

# **Joint Discourse-aware Concept Disambiguation and Clustering**

Dissertation

zur

Erlangung der Doktorwürde  
der Neuphilologischen Fakultät  
der Ruprecht-Karls-Universität Heidelberg

vorgelegt

von

**Angela Petra Fahrni**

Referent: Prof. Dr. Michael Strube  
Korreferent: Prof. Dr. Anette Frank  
Einreichung: 31.10.2014  
Disputation: 21.12.2015

# Abstract

This thesis addresses the tasks of concept disambiguation and clustering. *Concept disambiguation* is the task of linking common nouns and proper names in a text – henceforth called mentions – to their corresponding concepts in a predefined inventory. *Concept clustering* is the task of clustering mentions, so that all mentions in one cluster denote the same concept. In this thesis, we investigate concept disambiguation and clustering from a discourse perspective and propose a discourse-aware approach for joint concept disambiguation and clustering in the framework of Markov logic. The contributions of this thesis are fourfold:

**Joint Concept Disambiguation and Clustering.** In previous approaches, concept disambiguation and concept clustering have been considered as two separate tasks (Schütze, 1998; Ji & Grishman, 2011). We analyze the relationship between concept disambiguation and concept clustering and argue that these two tasks can mutually support each other. We propose the – to our knowledge – first joint approach for concept disambiguation and clustering.

**Discourse-Aware Concept Disambiguation.** One of the determining factors for concept disambiguation and clustering is the context definition. Most previous approaches use the same context definition for all mentions (Milne & Witten, 2008b; Kulkarni et al., 2009; Ratnov et al., 2011, inter alia). We approach the question which context is relevant to disambiguate a mention from a discourse perspective and state that different mentions require different notions of contexts. We state that the context that is relevant to disambiguate a mention depends on its embedding into discourse. However, how a mention is embedded into discourse depends on its denoted concept. Hence, the identification of the denoted concept and the relevant concept mutually depend on each other. We propose a binwise approach with three different context definitions and model the selection of the context definition and the disambiguation jointly.

**Modeling Interdependencies with Markov Logic.** To model the interdependencies between concept disambiguation and concept clustering as well as the

interdependencies between the context definition and the disambiguation, we use Markov logic (Domingos & Lowd, 2009). Markov logic combines first order logic with probabilities and allows us to concisely formalize these interdependencies. We investigate how we can balance between linguistic appropriateness and time efficiency and propose a hybrid approach that combines joint inference with aggregation techniques.

**Concept Disambiguation and Clustering beyond English: Multi- and Cross-linguality.** Given the vast amount of texts written in different languages, the capability to extend an approach to cope with other languages than English is essential. We thus analyze how our approach copes with other languages than English and show that our approach largely scales across languages, even without retraining.

Our approach is evaluated on multiple data sets originating from different sources (e.g. news, web) and across multiple languages. As an inventory, we use Wikipedia. We compare our approach to other approaches and show that it achieves state-of-the-art results. Furthermore, we show that joint concept disambiguating and clustering as well as joint context selection and disambiguation leads to significant improvements *ceteris paribus*.

# Zusammenfassung

Diese Dissertation beschäftigt sich mit Konzeptdisambiguierung und Konzeptclustering. Unter *Konzeptdisambiguierung* verstehen wir die Aufgabe, Gattungs- und Eigennamen in Texten – im Folgenden Erwähnungen genannt – zu ihren entsprechenden Konzepten in einem vorab definierten Inventar zu verlinken. *Konzeptclustering* ist die Aufgabe, Erwähnungen so zu gruppieren, dass alle Erwähnungen in einem Cluster das gleiche Konzept denotieren. In dieser Dissertation untersuchen wir Konzeptdisambiguierung und -clustering von einer Diskursperspektive und schlagen einen diskursbezogenen Ansatz für ein vereintes Disambiguieren und Clustern von Konzepten in Markov Logik vor. Die Forschungsbeiträge dieser Dissertation umfassen vier Bereiche.

**Vereintes Disambiguieren und Clustern von Konzepten.** Vorherige Ansätze modellieren Konzeptdisambiguierung und Konzeptclustering als zwei separate Aufgaben (Schütze, 1998; Ji & Grishman, 2011). Wir analysieren die Beziehung zwischen Konzeptdisambiguierung und Konzeptclustering und argumentieren, dass diese zwei Aufgaben sich wechselseitig unterstützen können. Wir schlagen den – unseres Wissens – ersten Ansatz für vereintes Disambiguieren und Clustern von Konzepten vor.

**Diskursbezogene Konzeptdisambiguierung.** Ein bestimmender Faktor für das Disambiguieren und Clustern von Konzepten ist die Kontextdefinition. Die meisten vorherigen Ansätze verwenden die gleiche Kontextdefinition für alle Erwähnungen (Milne & Witten, 2008b; Kulkarni et al., 2009; Ratinov et al., 2011, inter alia). Wir nähern uns der Frage, welcher Kontext relevant für die Disambiguierung von Erwähnungen ist, von einer Diskursperspektive und argumentieren, dass verschiedene Erwähnungen unterschiedliche Kontextdefinitionen erfordern. Wir legen dar, dass der für die Disambiguierung relevante Kontext davon abhängt, wie diese Erwähnung in den Diskurs eingebettet ist. Die Einbettung einer Erwähnung in den Diskurs hängt jedoch vom Konzept ab, das die Erwähnung denotiert. Dies führt dazu, dass die Identifikation des denotierten Konzeptes und die Bestimmung des relevanten Kontextes voneinander abhängen. In dieser Dissertation schlagen

wir einen Ansatz mit drei Kontextdefinitionen vor und modellieren die Identifikation des Kontextes für eine Erwähnung und deren Disambiguierung wechselseitig.

**Modellieren von Interdependenzen mit Markov Logik.** Um die Interdependenzen zwischen Konzeptdisambiguierung und Konzeptclustering sowie zwischen Kontextdefinition und Disambiguierung zu modellieren, verwenden wir Markov Logik (Domingos & Lowd, 2009). Markov Logik vereinigt Prädikatenlogik mit Wahrscheinlichkeiten und ermöglicht es, Interdependenzen präzise und prägnant zu formalisieren. Wir untersuchen, wie wir Konzeptdisambiguierung und Konzeptclustering einerseits linguistisch motiviert, andererseits zeiteffizient implementieren können, und schlagen einen hybriden Ansatz vor, der vereinte und aggregative Techniken kombiniert.

**Multi- und crosslinguales Disambiguieren und Clustern von Konzepten.**

Viele Texte sind nicht in Englisch verfügbar. Es ist daher zentral, dass ein Ansatz nicht nur für das Englische verwendbar ist, sondern auch andere Sprachen abdeckt. Wir analysieren, wie unser Ansatz auf andere Sprachen anwendbar ist, und zeigen, dass unser System erfolgreich andere Sprachen verarbeiten kann, selbst ohne sprachspezifisches Abstimmen der gelernten Parameter.

Wir evaluieren unseren Ansatz anhand von verschiedenen Datensätzen und berücksichtigen nicht nur unterschiedliche Textquellen (beispielsweise Zeitungen, Web), sondern auch verschiedene Sprachen. Als Inventar verwenden wir Wikipedia. Wir vergleichen unseren Ansatz mit verschiedenen anderen Ansätze und zeigen, dass die Ergebnisse unseres Ansatzes dem aktuellen Stand der Forschung entsprechen. Zudem zeigen wir, dass unser vereinter Konzeptdisambiguierungs- und -clusteringansatz sowie unsere vereinte Kontextmodellierung und Disambiguierung zu signifikant besseren Resultaten führen *ceteris paribus*.



# Acknowledgments

I am sitting in front of my thesis surrounded by a few boxes. Tomorrow, I will hand in my thesis and move out of my flat. So it is about time to add another two – I promise last – pages to my not too short thesis.

First of all, I would like to thank my supervisor Prof. Dr. Michael Strube. He managed to give me enough freedom to develop my own ideas, while being supportive at the same time. He always took a lot of time for discussions and I could count on his honest feedback.

I am very glad that Prof. Dr. Anette Frank is my co-referent. She was always encouraging and helped me to move on by asking relevant questions during my colloquium talks and providing valuable comments afterwards.

While at the beginning of writing this thesis, I was glad about each new page, at some point of time, I was glad about each sentence I could cut. Fortunately, Sebastian Martschat did a fantastic job in carefully reading my thesis and helping me to streamline it, similar to Nafise Moosavi, Jie Cai, Yufang Hou, Alex Judea, Mohsen Mesgar, Daraksha Parveen and Michael Roth who all read parts of it with great care.

I did not only obtain great support from all my colleagues while writing my thesis, but during the whole time as a PhD student. Dr. Vivi Nastase helped me a lot with her broad knowledge in my first years at HITS, introduced me in the project world and was always there for me as a friend. I really appreciate all the discussions I had with Jie Cai, Sebastian Martschat and Yufang Hou, which were very useful when I did not know how to solve a problem. In particular, I will not forget all the interesting conversations I had with Sebastian Martschat about evaluation metrics.

In addition to all the support I received in Heidelberg, I also learned a lot from people from other research institutes. Dr. Mathias Niepert helped me to better understand how to model a problem with Markov Logic and to "domesticate" *TheBeast*. Being part of the EU project CoSyne, I also had the chance to learn a lot from our project partners, in particular from Prof. Dr. Christof Monz, Dr. Matteo Negri and Amit Bronner.

Finally, I would like to thank Vivi Nastase, Michael Roth, Jie Cai, Viola Ganter, Federico Marinacci and all the other people from HITS that made me feel at home in Heidelberg very quickly. At the same time, I always had great support from my family and friends in Switzerland and Dr. Manfred Klenner who showed me how exciting our field is before I became a PhD student.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Concept Disambiguation and Clustering . . . . .	2
1.1.1	Concept Disambiguation . . . . .	4
1.1.2	Concept Clustering . . . . .	7
1.1.3	Multi- and Cross-linguality . . . . .	8
1.1.4	Applications . . . . .	9
1.2	Research Questions . . . . .	11
1.2.1	Interrelations between Concept Disambiguation and Clustering . . . . .	11
1.2.2	Identifying the Relevant Context . . . . .	11
1.2.3	Machine Learning Meets Linguistics: Modeling Interrelations . . . . .	12
1.2.4	Concept Disambiguation and Clustering beyond English: Scalability across Languages . . . . .	12
1.3	Research Contributions . . . . .	12
1.3.1	Joint Concept Disambiguation and Clustering . . . . .	13
1.3.2	Discourse-aware Concept Disambiguation . . . . .	13
1.3.3	Modeling Interdependencies with Markov Logic . . . . .	13
1.3.4	High Scalability across Languages . . . . .	13
1.4	Structure of the Thesis and Thesis Overview . . . . .	14
1.5	Published Work . . . . .	15
<b>2</b>	<b>Linguistic Background and Motivation</b>	<b>17</b>
2.1	Semiotic Perspective . . . . .	18
2.1.1	Mentions, Concepts and Entities . . . . .	18
2.1.2	The Role of Proper Names . . . . .	21
2.1.3	Summary . . . . .	23
2.2	Discourse Perspective . . . . .	25
2.2.1	Semiotics Meets Discourse . . . . .	25
2.2.2	Cohesive Ties between Concepts . . . . .	28
2.2.3	Cohesive Ties between Entities . . . . .	36

2.2.4	Cohesive Ties on the String Level . . . . .	38
2.2.5	Discussion . . . . .	39
2.3	Identification of the Disambiguation Context . . . . .	40
2.3.1	Uniform, Individual and Binwise Context Definitions . . . . .	41
2.3.2	Context Definition Based on Cohesive Scopes . . . . .	42
2.4	Implications . . . . .	46
2.4.1	Implications for Selecting an Inventory . . . . .	47
2.4.2	Implications for Modeling . . . . .	47
<b>3</b>	<b>Wikipedia as an Inventory</b>	<b>51</b>
3.1	Selecting an Inventory . . . . .	51
3.1.1	Wikipedia . . . . .	52
3.1.2	Comparison with Other Resources . . . . .	57
3.2	From Wikipedia to an Inventory . . . . .	67
3.2.1	Deriving a Concept Inventory from Wikipedia . . . . .	68
3.2.2	Building a Multilingual Concept Inventory . . . . .	69
3.3	Statistics . . . . .	70
3.4	Summary . . . . .	73
<b>4</b>	<b>Method</b>	<b>75</b>
4.1	Motivation and Prerequisites . . . . .	75
4.1.1	Aggregative Approaches . . . . .	76
4.1.2	Iterative Approaches . . . . .	78
4.1.3	Joint Inference . . . . .	79
4.1.4	Discussion . . . . .	80
4.2	Markov Logic . . . . .	82
4.2.1	Markov Logic as a Probabilistic Extension of First-order Logic . . . . .	82
4.2.2	Markov Logic as a Compact Representation of Markov Networks . . . . .	85
4.2.3	Inference . . . . .	88
4.2.4	Learning . . . . .	90
4.2.5	Implementation . . . . .	91
4.3	Joint Concept Disambiguation and Clustering . . . . .	92
4.3.1	Modeling Concept Disambiguation . . . . .	92
4.3.2	Modeling Concept Clustering in Markov Logic . . . . .	97
4.3.3	Modeling Joint Disambiguation and Clustering in Markov Logic . . . . .	100
4.3.4	Discussion . . . . .	101
4.4	Integrating Discourse Information . . . . .	103
4.4.1	A Scope-aware Approach . . . . .	104

---

4.4.2	Learning with Latent Variables . . . . .	108
4.4.3	Discussion . . . . .	110
<b>5</b>	<b>Features</b>	<b>111</b>
5.1	Features for Concept Disambiguation . . . . .	113
5.1.1	Prominence of a Concept . . . . .	115
5.1.2	Co-occurrence Information . . . . .	117
5.1.3	Concept Type Information . . . . .	120
5.2	Features for Concept Clustering . . . . .	121
5.2.1	Features for Within-document Concept Clustering . . . . .	121
5.2.2	Features for Cross-document Concept Clustering . . . . .	123
5.3	Features for Scope Assignment . . . . .	123
5.3.1	Mention-based Features . . . . .	124
5.3.2	Features Based on Modification . . . . .	124
5.3.3	Features Based on the Sentence and Text Structure . . . . .	125
<b>6</b>	<b>Monolingual English Architecture</b>	<b>127</b>
6.1	Preprocessing . . . . .	128
6.2	Mention Recognition and Candidate Concept Identification . . . . .	130
6.3	Feature Extraction and Inference . . . . .	131
6.4	Discussion . . . . .	131
<b>7</b>	<b>Multi- and Cross-linguality</b>	<b>133</b>
7.1	Strategies for Multi- and Cross-lingual Concept Disambiguation and Clustering	134
7.1.1	Multilingual Concept Disambiguation and Clustering . . . . .	134
7.1.2	Cross-lingual Concept Disambiguation and Clustering . . . . .	137
7.1.3	Discussion . . . . .	138
7.2	Scalability . . . . .	138
7.2.1	Inventory . . . . .	140
7.2.2	Preprocessing . . . . .	140
7.2.3	Mention Recognition and Candidate Concepts Identification . . . . .	140
7.2.4	Backbone of Our Approach . . . . .	141
7.2.5	Features . . . . .	141
7.2.6	Model Parameters . . . . .	142
7.3	Multi- and Cross-lingual Concept Disambiguation and Clustering . . . . .	143
7.3.1	Multilingual Adaptation . . . . .	143
7.3.2	Cross-lingual Adaptation . . . . .	144

7.4	Selection of Languages for the Evaluation . . . . .	144
<b>8</b>	<b>Data</b>	<b>147</b>
8.1	Blazing a Trail Through the Data Jungle . . . . .	147
8.1.1	Selection Criteria . . . . .	148
8.1.2	Data Set Selection . . . . .	149
8.2	Training Data . . . . .	149
8.3	Development Data . . . . .	151
8.4	Testing Data . . . . .	151
8.4.1	Monolingual Data Sets . . . . .	152
8.4.2	Multi- and Cross-lingual Data Sets . . . . .	154
8.5	Summary . . . . .	157
<b>9</b>	<b>Experiments</b>	<b>165</b>
9.1	Settings . . . . .	166
9.1.1	Baselines . . . . .	166
9.1.2	Upper Bounds . . . . .	167
9.1.3	System Variants . . . . .	167
9.1.4	State-of-the-Art Approaches . . . . .	170
9.2	Evaluation Metrics . . . . .	174
9.2.1	Metrics for Concept Disambiguation . . . . .	174
9.2.2	Metrics for Concept Clustering . . . . .	176
9.3	Results . . . . .	176
9.3.1	Monolingual English Results . . . . .	176
9.3.2	Multi- and Cross-lingual Results . . . . .	189
9.4	Analysis . . . . .	197
9.4.1	Joint Concept Disambiguation and Clustering . . . . .	197
9.4.2	Scope-Aware Approach . . . . .	203
9.4.3	Multi- and Cross-linguality . . . . .	208
9.4.4	Remaining Errors . . . . .	211
9.5	Summary . . . . .	214
<b>10</b>	<b>Related Work</b>	<b>215</b>
10.1	Task Definitions . . . . .	216
10.1.1	Lexical Semantics . . . . .	216
10.1.2	Information Extraction . . . . .	220
10.1.3	Information Retrieval . . . . .	222
10.1.4	Discussion . . . . .	223

---

10.2	Methods . . . . .	223
10.2.1	Methods for Concept Disambiguation . . . . .	224
10.2.2	Methods for NIL Recognition . . . . .	235
10.2.3	Methods for Concept Clustering . . . . .	238
10.3	Context Definitions . . . . .	239
10.3.1	Concept-level Context Modeling . . . . .	242
10.3.2	String-level Context Modeling . . . . .	248
10.4	Modeled Information . . . . .	250
10.4.1	Modeled Information for Concept Disambiguation and Recognition of NILs . . . . .	250
10.4.2	Modeled Information for Concept Clustering . . . . .	254
10.5	Multi- and Cross-linguality . . . . .	254
10.5.1	Multi- and Cross-lingual Tasks . . . . .	255
10.5.2	Multi- and Cross-lingual Strategies . . . . .	256
10.5.3	Multi- and Cross-lingual Features . . . . .	259
10.5.4	Discussion . . . . .	260
10.6	Summary . . . . .	260
<b>11</b>	<b>Conclusions and Future Work</b>	<b>261</b>
11.1	Contributions . . . . .	261
11.1.1	Joint Concept Disambiguation and Clustering . . . . .	261
11.1.2	Discourse-aware Concept Disambiguation and Clustering . . . . .	262
11.1.3	Modeling Interdependencies with Markov Logic . . . . .	263
11.1.4	High Scalability across Languages . . . . .	264
11.2	Limitations . . . . .	265
11.2.1	Modeling of NILs . . . . .	265
11.2.2	Restriction to Common Nouns and Proper Names . . . . .	265
11.2.3	Speed and Scalability . . . . .	265
11.3	Future Research Directions . . . . .	266
11.3.1	Evaluation Settings and Metrics . . . . .	266
11.3.2	A Multi-layer Model for Concept Disambiguation and Clustering . . . . .	266
11.3.3	Beyond the Textual Context . . . . .	267
	<b>Bibliography</b>	<b>269</b>



# List of Figures

1.1	Example text with mentions . . . . .	3
1.2	Example for concept disambiguation . . . . .	5
1.3	Different variations of the concept disambiguation task . . . . .	6
1.4	Example for concept clustering . . . . .	7
1.5	Different variations of the clustering task . . . . .	8
1.6	Different cross- and multilingual variations of the concept disambiguation and clustering task . . . . .	9
2.1	Semiotic triangle after Ogden and Richards (1923), p.11. . . . .	18
2.2	Mentions, concepts, entities . . . . .	21
2.3	Common nouns vs. proper names . . . . .	24
2.4	Cohesive ties on the string, concept and entity level . . . . .	27
2.5	Relationship between concept disambiguation and concept clustering . . . . .	28
2.6	Joint concept disambiguation and clustering . . . . .	29
2.7	Concept-level cohesive ties for two candidate concepts . . . . .	31
2.8	Relatedness relations for the mention <i>laws</i> . . . . .	33
2.9	Exploiting clustering . . . . .	40
2.10	Mentions with local cohesive ties . . . . .	43
2.11	Mentions with intermediate cohesive ties . . . . .	44
2.12	Mentions with global cohesive ties . . . . .	45
3.1	Sentence from Wikipedia with an internal hyperlinks . . . . .	55
3.2	Example for a disambiguation page in Wikipedia . . . . .	56
3.3	Examples for noun synsets in WordNet . . . . .	61
4.1	Graph-based representation of an aggregative approach . . . . .	77
6.1	Cascaded vs. joint approach . . . . .	128
6.2	Workflow . . . . .	129
7.1	Strategies for multilingual concept disambiguation and clustering . . . . .	135

---

7.2	Strategies for cross-lingual concept disambiguation and clustering. . . . .	139
9.1	Accuracy on the English TAC data sets from 2011 to 2013 . . . . .	182
9.2	$B^{3+}$ scores on the English TAC data sets from 2011 to 2013 . . . . .	183
9.3	Distribution across induced scopes . . . . .	205
9.4	Distribution of errors across different types . . . . .	212
10.1	Modeling the context . . . . .	241
11.1	Towards a multi-layer model for concept disambiguation and clustering . . .	267

# List of Tables

1.1	Relation between our previously published work and this thesis . . . . .	16
2.1	Examples for cohesive ties of type identity and relatedness . . . . .	27
3.1	Comparison of inventories, part I . . . . .	64
3.2	Comparison of inventories, part II . . . . .	66
3.3	Number of concepts and lexicalizations for different languages . . . . .	70
3.4	Number of concepts that are mapped to English . . . . .	71
3.5	Number of concepts with at least one inlink . . . . .	72
3.6	Number of inlinks for different languages . . . . .	73
4.1	Example for a knowledge base in first-order logic . . . . .	83
4.2	Example for a ground formula . . . . .	84
4.3	Example for a knowledge base in Markov logic . . . . .	86
4.4	Backbone for concept disambiguation in Markov logic . . . . .	96
4.5	Backbone for concept clustering in Markov logic . . . . .	99
4.6	Backbone for joint concept disambiguation and clustering in Markov logic . .	102
4.7	Backbone for scope-aware concept disambiguation in Markov logic . . . . .	107
5.1	Features for concept disambiguation . . . . .	112
5.2	Features for concept clustering . . . . .	113
5.3	Features for the scope assignment task (part I) . . . . .	114
5.4	Features for the scope assignment task (part II) . . . . .	115
7.1	Language-specific adaptations for Spanish and Chinese. . . . .	145
8.1	Training and development data . . . . .	151
8.2	Spanish cross-lingual data sets . . . . .	154
8.3	Chinese cross-lingual data sets . . . . .	156
8.4	Comparison of different data sets that use Wikipedia as an inventory . . . . .	164
9.1	System variants . . . . .	169

9.2	Summary of related work . . . . .	173
9.3	Results on ACE 2005 . . . . .	177
9.4	Results on ACE 2004 . . . . .	180
9.5	Results on the English TAC 2011 data set . . . . .	186
9.6	Results on the English TAC 2012 data set . . . . .	187
9.7	Results on the English TAC 2013 data set . . . . .	188
9.8	Results on the Spanish cross-lingual TAC 2012 and 2013 data sets . . . . .	191
9.9	Results on the Chinese cross-lingual TAC 2011 and 2012 data sets . . . . .	194
9.10	Results on the Chinese cross-lingual TAC 2013 data set . . . . .	195
9.11	Results on the cross-lingual link discovery task at NTCIR 9 . . . . .	197
9.12	Cascaded vs. joint concept disambiguation and clustering: results for disambiguation . . . . .	198
9.13	Cascaded vs. joint concept disambiguation and clustering: results for clustering	199
9.14	Within-document vs. cross-document clustering . . . . .	200
9.15	Influence of the clustering on mentions that do not contain the denotated concept among their candidate concepts . . . . .	201
9.16	Influence of the clustering on mentions that contain the denotated concept among their candidate concepts . . . . .	201
9.17	Scope-unaware vs. cascaded scope-aware vs. joint scope-aware approach . . . . .	204
9.18	Results for local, intermediate and global scope . . . . .	206
9.19	Results on the CoNLL 2003 data set . . . . .	207
9.20	Results for scope-informed keyword selection . . . . .	208
9.21	Monolingual vs. cross-lingual concept-level co-occurrences . . . . .	210
9.22	Language-independent vs. language-specific features . . . . .	211
9.23	Accuracy for different text sources . . . . .	214
10.1	Classification of global approaches based on how they model interdependencies between concepts . . . . .	230
10.2	Global approaches for disambiguation: overview . . . . .	244
10.3	Commonly modeled information in concept disambiguation and NIL recognition. . . . .	251
10.4	Different multi- and cross-lingual tasks . . . . .	257
10.5	Comparison of different multi- and cross-lingual disambiguation approaches . . . . .	258

# Chapter 1

## Introduction

If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. “Fast” may mean “rapid”; or it may mean “motionless”; and there is no way of telling which. But if one lengthens the slit in the opaque mask, until one can see not only the central word in question, but also say  $N$  words on either side, then if  $N$  is large enough one can unambiguously decide the meaning of the central word. [...] The practical question is, what minimum value of  $N$  will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word? (Weaver, 1955, p. 8, written in 1949).

Resolving meaning ambiguities of single and multiword tokens in a text is one of the oldest research problems in computational linguistics. From the beginning, the context definition has been considered a determining factor to solve the meaning ambiguity problem (Weaver, 1955; Bar-Hillel, 1960). Since Weaver (1955), many different context definitions have been explored, ranging from a few surrounding words (Ng & Lee, 1996) up to the whole document-level context (Navigli & Lapata, 2010). However, the factors that influence the context relevant to resolve meaning ambiguities are still hardly understood. After many years of research and many shared tasks organized at *Senseval* (Kilgarriff & Rosenzweig, 2000; Palmer et al., 2001, inter alia) and *SemEval* (Pradhan et al., 2007, inter alia), the performance of state-of-the-art approaches has reached a plateau (Agirre & Edmonds, 2006, p. 7). The idea of using Wikipedia as a meaning inventory (Bunescu & Paşca, 2006; Cucerzan, 2007; Csomai & Mihalcea, 2008) has given the field a new spin. In the last nine years, new approaches have been proposed – mainly in the information extraction and information retrieval communities – focusing on resolving ambiguities of nouns, in particular of proper names. However, despite of the huge amount of research, resolving meaning ambiguities is still an open problem.

This thesis is about resolving meaning ambiguities of common nouns and proper names.

The contributions of this thesis are fourfold. We analyze the problem from a semiotic and discourse perspective and develop a deeper understanding of the task and of one of its determining factors, the context. Based on these insights, we propose a joint concept disambiguation and clustering approach (first contribution) and a binwise context model (second contribution). Our approach makes use of state-of-the-art machine learning techniques and is empirically evaluated and compared to related research (third contribution). The proposed approach is developed for English. However, given the vast amount of texts in other languages, scalability across languages is essential. We thus analyze how our approach can be ported to other languages (fourth contribution). This also allows us to better understand the properties and the underlying assumptions of our proposed approach.

In the remainder of this chapter, we first describe the task and some applications (Section 1.1). We then discuss the main questions addressed in this thesis (Section 1.2) and summarize its contributions (Section 1.3). To help the reader to navigate through the thesis, we outline its structure (Section 1.4) and point to our publications related to this thesis (Section 1.5).

## 1.1 Concept Disambiguation and Clustering

The task addressed in this thesis is ambiguity resolution of common nouns and proper names. Hence, the aim is to automatically identify relations between two linguistic levels: the token or string level and the meaning level. In the following, we specify the task by defining these two linguistic levels and the relation between them.

Words of almost all parts of speech can be ambiguous, including nouns, verbs, adjectives or prepositions. In this thesis, we focus on common nouns and proper names. Common nouns and proper names are content-bearing linguistic units. They realize the discourse entities, which are essential for e.g. analyzing the discourse structure (e.g. Barzilay & Lapata (2008)) and extracting relations from texts (e.g. Banko et al. (2007), Ji & Grishman (2011)). Resolving ambiguities of common nouns and proper names is thus essential for automatic text understanding. Following previous work (e.g. Ratinov et al. (2011)), we call the token sequences which we aim to resolve *mentions*.

**Definition 1.1** A *mention* is an occurrence of a common noun or proper name in a text and spans one single or multiple tokens. A multi-token mention can include a preposition phrase or an adjective.

In the whole thesis, mentions are written in italics. Figure 1.1 shows a news article, which we use in the following for illustration purposes. The mentions are surrounded by boxes. As for instance the sequence *online-dating app Tinder* illustrates, it is not always straightforward to determine the boundaries of mentions. Data sets slightly vary in their mention definitions.

The *state of New York* has effectively banned the popular *trend* of taking *selfie photos* with *tigers* or *big cats* by saying *people* are no longer allowed near dangerous *animals* at *zoos*, *circuses* and *carnivals*.

The new *law* comes after the *online-dating app* *Tinder* saw a *surge* of *photos* – mainly of *men* – posting *profile pictures* of themselves next to *tigers* and other *big cats*. The *phenomenon* has come to be known as the “*tiger selfie*” – local *website* *Politics on the Hudson* reports. A *report* earlier this *year* from the *Wall Street Journal* estimates that one in 10 *photos* on *Tinder* has a *tiger* in it – perhaps because the *men* want to appear adventurous to potential *partners*.

But *assembly member* *Linda Rosenthal*, who sponsored the *bill*, tells the *Daily News* *website* the *measure* is there to stop *animals* from being exploited. *Wildlife activists* say *tiger selfies* encourage *people* to take *cubs* from *big cats* who are later neglected, mistreated and abandoned when they grow up. Similar *laws* are already in *force* in *states* such as *Mississippi*, *Arizona* and *Kansas*.

Figure 1.1: English news article from BBC News.<sup>1</sup><sup>1</sup><http://www.bbc.com/news/blogs-news-from-elsewhere-28778947>, 13.8.2014.

Different linguistic theories define the *meaning* of a word differently (Section 2.1). We assume that the meaning of a mention is the *concept* it denotes.

**Definition 1.2** *Based on Eco (2002), we define a **concept** as a cultural unit that is shared among different people.*

In other work, concepts are often called *senses* (Agirre & Edmonds, 2006; Navigli, 2009) or *topics* (Zhou et al., 2010a). In proper name disambiguation, the term *entities* is commonly used (Ji & Grishman, 2011). We stick to the term *concept*, as senses are often equated with WordNet senses (Miller, 1995), and topics and entities can be confused with text topics and real-world entities respectively. We indicate concepts by small capitals. In the text in Figure 1.1, the mention *trend* denotes the concept FAD<sup>2</sup>, while *state of New York* denotes NEW YORK.

Ambiguities of mentions can either be resolved by identifying the corresponding concept in a predefined inventory (*concept disambiguation*) or by clustering mentions so that all occurrences that denote the same concept are in the same cluster (*concept clustering*). In the case of concept disambiguation, concepts are explicitly represented and relationships between mentions and concepts are established. In contrast, concept clustering approaches lack such an explicit concept representation and model relationships between mentions.

In both concept disambiguation and clustering, two main challenges need to be addressed: *ambiguity* and *variability* (e.g. Dredze et al. (2010)).

**Definition 1.3** *A mention string is **ambiguous** if it can denote different concepts given that no context is considered.*

**Definition 1.4** *Different mention strings can denote the same concept. This phenomenon is called **variability**.*

For instance, *law* is ambiguous and can e.g. denote a man-made law, a physical law or a scientific law. The concept FAD can be realized by e.g. the two variants *fad* or *trend* and thus exhibits variability. In the following, we discuss the two tasks concept disambiguation and clustering in more detail.

### 1.1.1 Concept Disambiguation

In this thesis, we define concept disambiguation as follows.

**Definition 1.5** ***Concept disambiguation** is the task of linking mentions in a text to the concepts they denote. The concepts are predefined and part of a larger concept inventory.*

<sup>2</sup>The concept names correspond to Wikipedia article names.

Thus, concept disambiguation requires that concepts are defined a priori. A collection of such concept definitions is called an inventory.

**Definition 1.6** An *inventory* consists of concept definitions and contains different linguistic realizations for each concept. Concepts may be interlinked.

Examples of such inventories for English are *WordNet* (Miller, 1995) and *Roget's Thesaurus* (Roget, 1964). We use Wikipedia as an inventory as it best fits our concept definition (Chapter 3). In Figure 1.2, the inventory is illustrated on the right-hand side. Bold lines indicate which mention denotes which concept. For instance, *men* denotes MAN (MALE) and not HUMAN. Due to space limitations, the inventory only shows a few candidate concepts. In fact, these mentions have many more candidates. For instance, *men* has more than 25 candidate concepts according to our lexicon, including Tolkien's concept MAN (MIDDLE-EARTH), different places and companies with this name. Concept disambiguation can be considered as a labeling task where the candidate labels depend on the respective mention string (Navigli, 2009). Hence, the task consists of two main steps: during the *candidate concept identification*, candidate concepts are identified for each mention; in the effective *disambiguation* step, one of the respective candidate concepts is selected for each mention.

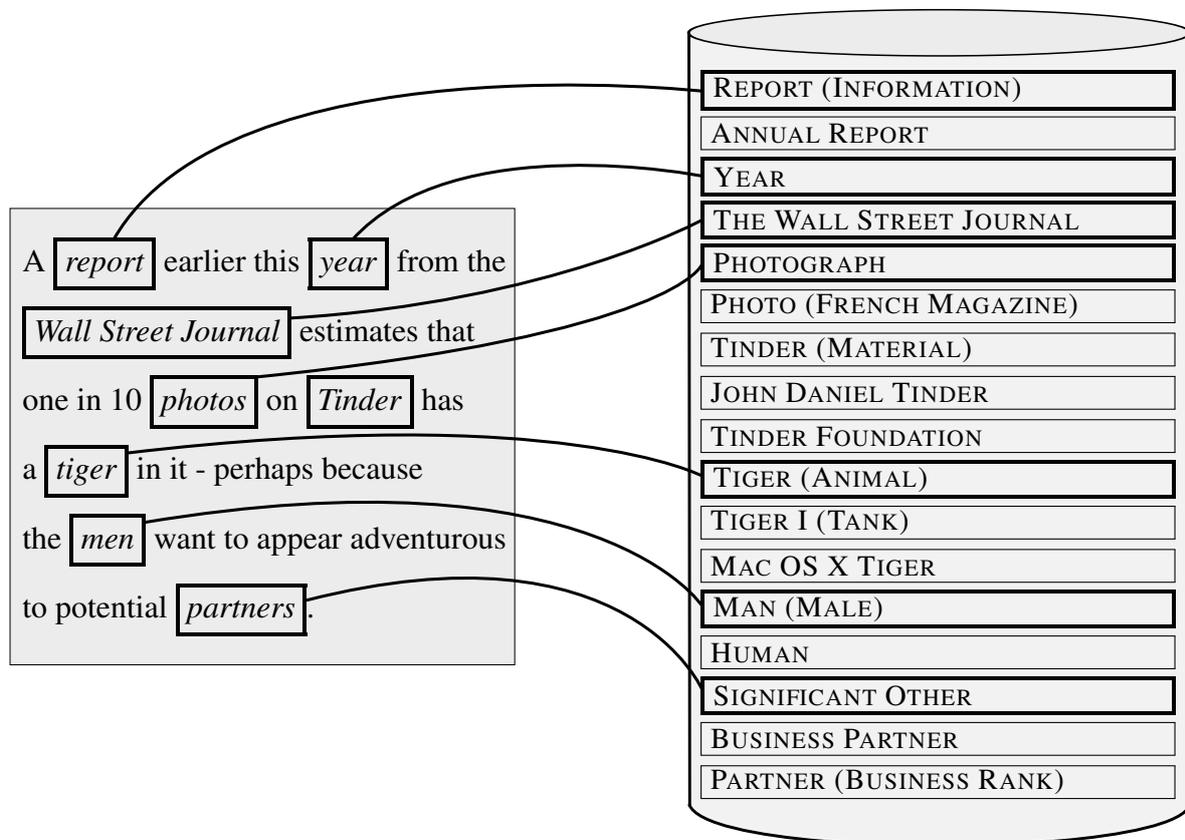


Figure 1.2: Concept disambiguation: on the right-hand side, parts of the inventory are shown. Sentence from the same English news article from BBC News as in Figure 1.1.

	Proper names	Common nouns	Other parts of speech
<b>All occurrences</b>	Entity disambiguation, entity linking	Word sense disambiguation	
	Concept disambiguation		
<b>Keywords</b>	Wikification		

Figure 1.3: Different variations of the concept disambiguation task. Our definition encloses the gray area.

Many variants of the concept disambiguation tasks exist. In work that uses *WordNet* as an inventory, the task is commonly called *word sense disambiguation* (Navigli, 2009). In word sense disambiguation, proper names are not disambiguated. If the emphasis lies on the disambiguation of keywords with Wikipedia as an inventory, the task is known as *wikification* (Csomai & Mihalcea, 2008). In proper name disambiguation, the task is usually called *entity disambiguation* or *entity linking* (Ji & Grishman, 2011). The differences between the tasks are summarized in Figure 1.3. They all have in common that the aim is to link occurrences to entries in a predefined inventory.

While the inventory is often assumed to be complete (e.g. Csomai & Mihalcea (2008), Milne & Witten (2008b)), more recent work considers that the inventory may lack some concepts (e.g. Bunescu & Paşca (2006), Ji & Grishman (2011)). Hence, we can distinguish between mentions that lack a corresponding concept in the inventory and mentions that have a corresponding concept in the inventory.

**Definition 1.7** *If a mention denotes a concept that is not part of the inventory, we call it a NIL. If a mention denotes a concept that is part of the inventory, we call it a Non-NIL.*

Common terms for the concepts denoted by *NILs* are *emerging entities* or *out-of-knowledge-base entities* (Hoffart et al., 2014). In Figure 1.2, the mention *Tinder* is a *NIL*, as it lacks a corresponding concept in the inventory. Although we can identify different candidates for it – e.g. the person JOHN DANIEL TINDER, a material called tinder and the organization TINDER FOUNDATION –, the actually denoted one, TINDER (APPLICATION), an online dating app that has become popular recently, is missing in our Wikipedia version from 2012. If *NILs* were ignored, *Tinder* would be wrongly linked to one of the existing candidate concepts. Another example is the mention *tiger selfie* (Figure 1.1). In 2012, neither a concept SELFIE nor a concept TIGER SELFIE exists in Wikipedia. SELFIE is a novel concept that has been established since 2012.<sup>3</sup> We define the *NIL* recognition task as follows:

**Definition 1.8** *NIL recognition is the task of recognizing mentions that denote a concept which is not part of the inventory.*

<sup>3</sup>It is now part of the current Wikipedia version: <http://en.wikipedia.org/wiki/Selfie>, 14.8.2014.

In this thesis, we assume that the inventory is incomplete. We therefore address the NIL recognition task.

### 1.1.2 Concept Clustering

In this thesis, we use the following definition of concept clustering:

**Definition 1.9** *Concept clustering is the task of clustering mentions, so that all mentions in one cluster denote the same concept.*

In contrast to concept disambiguation, concept clustering does not presuppose a predefined inventory. Concept clustering can either be done within one single document or across docu-

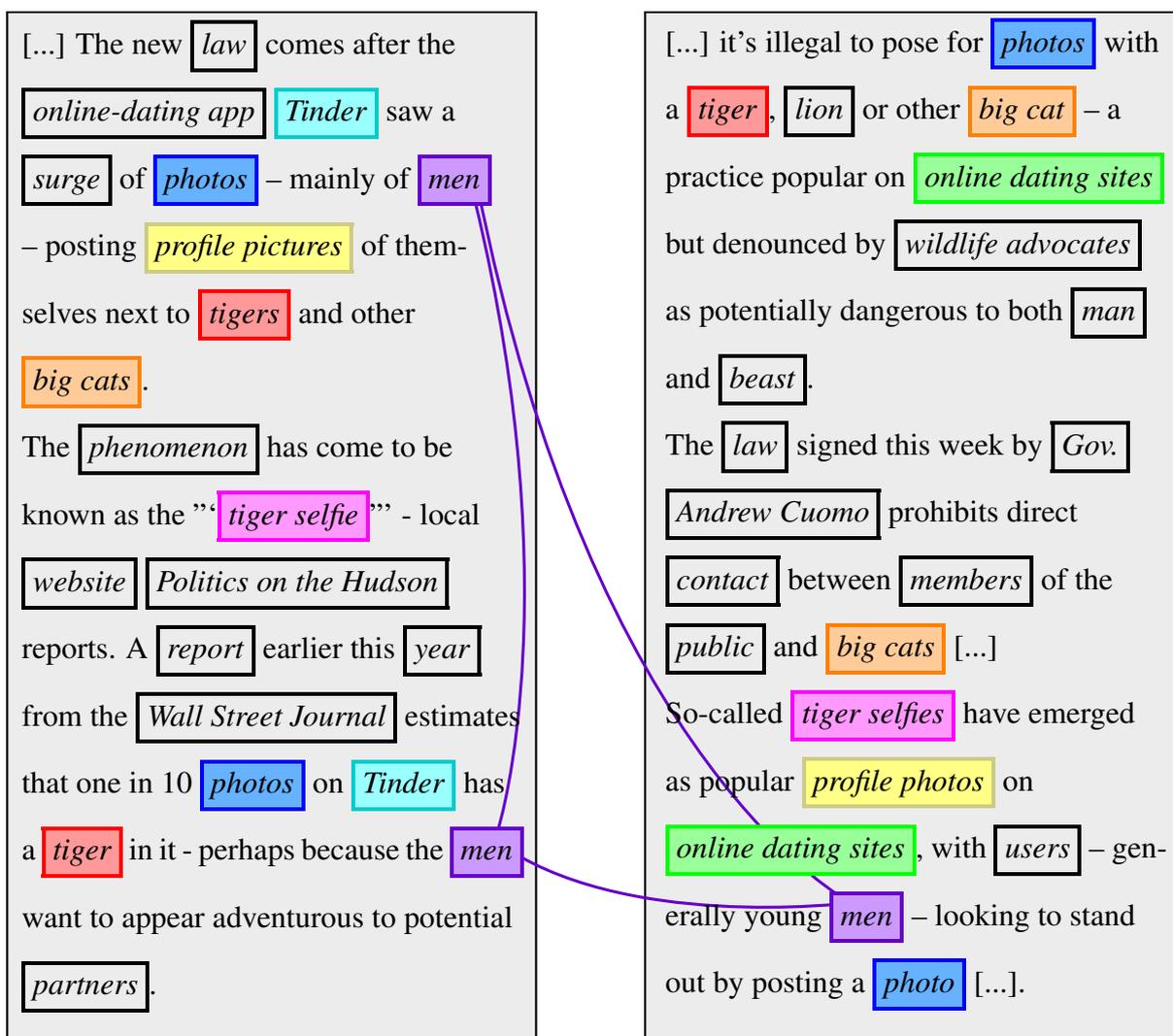


Figure 1.4: Concept clustering: the mentions in the same cluster share the same color. For one example, the mentions are connected by a line.

	Proper names	Common nouns	Other parts of speech
Token-based	Concept clustering		
	Entity clustering	Sense discrimination	
Type-based			

Figure 1.5: Different variations of the clustering task. Our definition encloses the gray area.

ments. In Figure 1.4, the clustering relations between mentions are indicated by colors. For the mentions *men*, we additionally indicate the clustering relations by bold lines. The three occurrences of *men* in Figure 1.4 share the same cluster – they all denote MAN (MALE). The mention *man* is part of another cluster, as it denotes another concept, i.e. HUMAN.

Similar as in disambiguation, different variants of the clustering task exist. According to our task definition, occurrences are clustered. It is thus a *token-based* task (Pedersen, 2006). In *type-based* tasks, all occurrences of a word type are considered and compared to the ones of other word types (Pedersen, 2006). Figure 1.5 illustrates the differences between task definitions.

### 1.1.3 Multi- and Cross-linguality

Most work on disambiguation and clustering focuses on one single language, usually English. However, ambiguity and variability is present in all languages. It is therefore desirable to also support other languages than English. We distinguish between multi- and cross-lingual approaches.

**Definition 1.10** *In multilingual concept disambiguation and clustering, the tasks are not only performed in one language, but in multiple languages. Each language is processed in isolation. In multilingual concept disambiguation, the concept definitions in the inventory and the texts to disambiguate are in the same language.*

**Definition 1.11** *In cross-lingual concept disambiguation, the concept description in the inventory and the texts to disambiguate are in different languages.<sup>4</sup> In cross-lingual concept clustering, mentions from texts in different languages are clustered across languages.<sup>5</sup>*

Figure 1.6 illustrates different multi- and cross-lingual variations of the concept disambiguation and clustering tasks and compares them to the monolingual setting. Depending on the downstream task or on the resources available for a language, either a cross-lingual or a multilingual task definition might be more suitable. For instance, if a downstream task requires a common representation of texts in different languages, a cross-lingual approach better

<sup>4</sup>This definition is derived from McNamee et al. (2011) for cross-lingual entity linking.

<sup>5</sup>This definition is analogous to the definition in Green et al. (2012) to cross-lingual entity clustering.

fits the requirements. In this thesis, we address the monolingual task (Figure 1.6, 1) and the two cross-lingual tasks (Figure 1.6, 2 and 4).

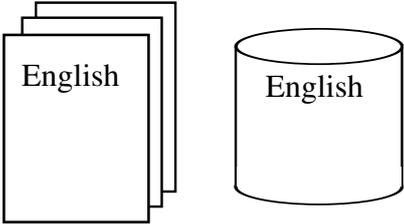
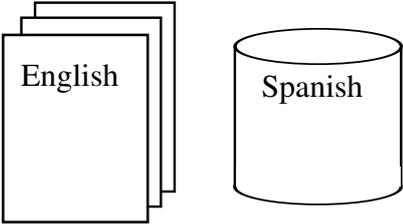
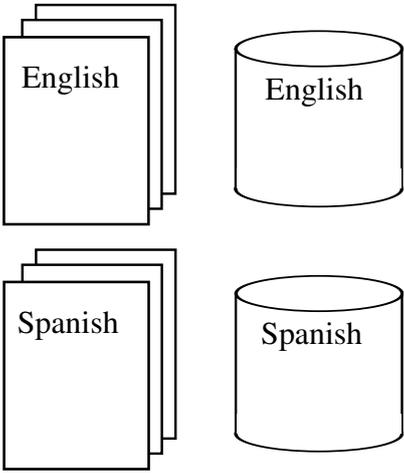
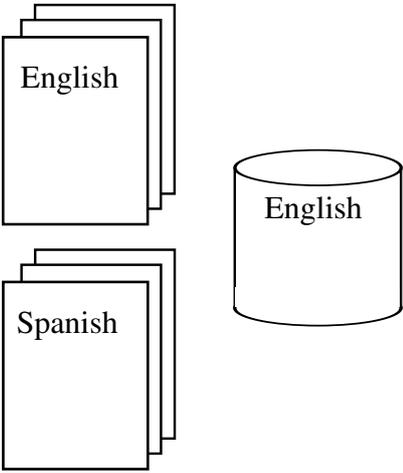
	<b>not cross-lingual</b>	<b>cross-lingual</b>
<b>not multilingual</b>	<p><b>(1):</b> monolingual concept disambiguation and clustering (one single language, e.g. English)</p> 	<p><b>(2):</b> cross-lingual concept disambiguation from one language to another one (e.g. texts are in English, inventory is in Spanish); cross-lingual clustering in this setting is not possible</p> 
<b>multilingual</b>	<p><b>(3):</b> multi-lingual concept disambiguation and clustering (multiple languages, e.g. English, Spanish)</p> 	<p><b>(4):</b> cross-lingual concept disambiguation and clustering across multiple languages (e.g. inventory is in English, texts are in English, German, Spanish)</p> 

Figure 1.6: Different cross- and multilingual variations of the concept disambiguation and clustering task. In this thesis, we focus on task 1, task 2 and task 4.

### 1.1.4 Applications

Concept disambiguation and clustering allow to study the meaning of nouns and to identify the relevant factors that determine the meaning of a noun in a certain context. Apart from addressing such linguistically oriented research questions, concept disambiguation and cluster-

ing have many applications in the field of computational linguistics and information retrieval. The massive impact concept disambiguation and clustering have or could have on applications is reflected in the governmental funding available for this research field (e.g. by DARPA) and the recent interest companies have shown for this area. In the following, we discuss some of these applications.

**Indexing, Text Similarity and Text Classification.** Every task that can be solved using a bag-of-words representation can be addressed with a concept-based text representation instead (Milne & Witten, 2008b). For instance, indexing could be done based on concepts instead of words (Kulkarni et al., 2009). While the usefulness of disambiguation for information retrieval has been controversial for a long time (Krovetz & Croft, 1989; Sanderson, 1994; Voorhees, 1999), recent studies indicate positive results (Agirre et al., 2010a; Zhong & Ng, 2012). Some researchers even assume that the disambiguation of proper names could trigger a paradigm shift in web search (Jin et al., 2014). A concept-based text representation not only accounts for ambiguity and variability, but – if a cross-lingual approach is chosen – can also bridge between different languages. We have shown in the context of an EU project on multilingual content synchronization (CoSyne) that a cross-lingual concept-based representation is useful to calculate cross-lingual text similarity (Nastase et al., 2011). Another application for the concept-based representation is text classification (Gabrilovich & Markovitch, 2007b). In particular, in the context of Twitter, concept-based approaches are of interest (Michelson & Macskassy, 2010).

For the applications in this field, both concept disambiguation and clustering approaches are applicable, as the main goal is to address ambiguity and variability issues. However, by linking mentions in a text to concepts in an inventory, not only ambiguity and variability are resolved, but the text can be enriched with additional information retrieved from the concept descriptions in the inventory or from the relations between the concepts if available. On the downside, it is controversial if predefined concepts meet the granularity required by an application (e.g. McCarthy (2006a)).

**Hyperlink Identification.** Hyperlink identification is a direct application of concept disambiguation. Using a resource such as Wikipedia as an inventory, keywords in texts can be linked to their corresponding articles in Wikipedia providing users with easy-to-access background information. Csomai & Mihalcea (2008) and Milne & Witten (2008b) show that state-of-the-art disambiguation algorithms lead to good results on this task. At the cross-lingual link discovery task at NTCIR (Tang et al., 2011; 2013), this task has been extended to a cross-lingual task. In this cross-lingual task, systems are required to insert hyperlinks to articles that are written in another language than the input texts.

**Information Extraction.** Concept disambiguation and clustering are essential in information extraction (Ji & Grishman, 2011). To extract information from a sentence such as “*Mueller* recently became *father*”, the mentions need to be disambiguated or clustered. Disambiguation or clustering prevents that all information extracted for *Mueller* is pooled.

**Coreference Resolution.** Coreference resolution benefits from semantic and encyclopedic relations present in concept inventories (Ponzetto & Strube, 2011; Ratinov & Roth, 2012). Recently, coreference resolution and proper name disambiguation have even been addressed jointly (Hajishirzi et al., 2013). In the case of coreference resolution, concept disambiguation is more useful than clustering. By linking mentions to their corresponding concepts in the inventory, more information can be obtained from the inventory, which is useful for e.g. similarity calculations between mentions.

## 1.2 Research Questions

Concept disambiguation and clustering involve many different facets and can be studied from many perspectives. In this thesis, we examine the problem from a semiotic and discourse perspective. We mainly investigate (1) interrelations between disambiguation and clustering, (2) the identification of the context that is relevant to disambiguate a mention, (3) the modeling of interrelations using state-of-the-art machine learning approaches, and (4) scalability across languages. In the following we discuss these research questions in more detail.

### 1.2.1 Interrelations between Concept Disambiguation and Clustering

Concept disambiguation and clustering solve the ambiguity and variability problem from different perspectives. While for instance Schütze (1998) proposes to first perform clustering to obtain senses and then use them for disambiguation, the entity linking task at TAC requires to link mentions to an inventory and to cluster the NILs (Ji & Grishman, 2011). In both cases, disambiguation and clustering are treated as two cascaded tasks. In this thesis, we study the interrelations between the two tasks and analyze if they could mutually support each other. Our first research question is:

**Question 1.** What is the relation between concept disambiguation and clustering? Can concept clustering be used to improve concept disambiguation and vice versa?

### 1.2.2 Identifying the Relevant Context

The definition of the context is one of the determining factors in concept disambiguation and clustering. Most previous approaches use one single context definition and apply it to all

mentions (inter alia Kulkarni et al. (2009), Navigli & Lapata (2010), Ratinov et al. (2011), Ferragina & Scaiella (2012)). The common assumption is that the better a candidate concept fits its context, the more likely it is. In this thesis, we question this assumption and analyze the factors that influence the context that is relevant to disambiguate or cluster a mention. We thus pick up a question already posted by Weaver (1955) (see above).

**Question 2.** What is the context relevant to disambiguate a mention? Which factors determine which context is relevant to disambiguate a mention?

### 1.2.3 Machine Learning Meets Linguistics: Modeling Interrelations

Our aim is to model the findings from our linguistic analysis (first two questions). From a machine learning perspective, both questions address label interdependencies, i.e. relations between concepts. We analyze how we can model these interdependencies using state-of-the-art machine learning techniques. Our third research question is:

**Question 3.** How can we model concept disambiguation and clustering in accordance to our linguistic findings using state-of-the-art machine learning techniques?

### 1.2.4 Concept Disambiguation and Clustering beyond English: Scalability across Languages

As the support of languages other than English – including languages for which only a few resources and tools are available – is important, we analyze how our system scales across languages. Porting our system to other languages also allows us to reveal its underlying assumption (Bender, 2011). We study different strategies to port our system from one language to another. In particular, we investigate if information obtained for a language with many resources is beneficial for a language with only few available resources. Thus, our fourth research question is:

**Question 4.** How can we port our system to languages other than English? Can we share information across languages to boost the performance?

## 1.3 Research Contributions

The research contribution of this thesis is a joint discourse-aware concept disambiguation and clustering approach. We not only question common assumptions made in concept disambiguation and clustering, but also provide model ideas that are transferable to other natural language processing problems. For instance, the model idea for joint concept disambiguation and clustering has been successfully adapted for bridging anaphora resolution (Hou et al., 2013).

### 1.3.1 Joint Concept Disambiguation and Clustering

We show that concept disambiguation and clustering are two complementary tasks and propose an approach for joint concept disambiguation and clustering using Markov logic. The results indicate that the two tasks mutually support each other. As our clustering is trained based on concept-annotated data, the granularity of the produced clusters matches the granularity of the concepts in our inventory – a prerequisite for a mutual enforcement. This leads to our first research contribution:

**Contribution 1.** Concept disambiguation and clustering mutually support each other given that the granularity of the clusters matches the one of the concepts in the inventory.

### 1.3.2 Discourse-aware Concept Disambiguation

Taking a discourse perspective, we conclude that the relevant context to disambiguate a mention consists of the lexical units with which it shares concept-level cohesive ties. As the concept-level cohesive ties of a mention depend on the concept it denotes, the context relevant to disambiguate a mention depends on its concept and vice versa. These observations lead to the following conclusion, which is our second research contribution:

**Contribution 2.** Different mentions require different notions of contexts. Which context is relevant to disambiguate or cluster a mention depends on its embedding into discourse and its denoted concept.

### 1.3.3 Modeling Interdependencies with Markov Logic

The interdependencies between concept disambiguation and clustering and between the context definition and the denoted concepts can be formalized as label interdependencies. Markov logic allows to express such interdependencies between labels. We thus implement our approach in Markov logic and combine explicit modeling of interrelations with aggregative techniques to keep it tractable. In the same vein, we distinguish between three different context definitions – instead of using an individual context definition for each mention – and model them as latent variables.

**Contribution 3.** We propose an approach in Markov logic that combines explicit modeling of interrelations with aggregative techniques and uses latent variables to model different context definitions.

### 1.3.4 High Scalability across Languages

We show that our approach largely scales across languages and only requires a few adaptations. Our approach even achieves state-of-the-art results without retraining the model

for a new language. This indicates that the features and the weights are mainly language-independent, at least for the languages we consider.

**Contribution 4.** Our proposed joint concept disambiguation and clustering approach largely scales across languages and achieves state-of-the-art results even without retraining.

## 1.4 Structure of the Thesis and Thesis Overview

This thesis consists of eleven chapters. We first analyze the task from a semiotic and discourse perspective and lay the ground for the proposed approach (Chapter 2). Then, we describe our approach including the inventory (Chapter 3), the method (Chapter 4), the features (Chapter 5) and the architecture (Chapter 6). While until this point we focus on English, we address the multi- and cross-lingual extensibility in Chapter 7. In Chapter 8, we describe the data sets we use for the evaluation of the proposed approach (Chapter 9). Chapter 10 is devoted to related work and embeds the proposed approach in the broader research context. Finally, we draw some conclusions and indicate future research directions (Chapter 11). In the following, we discuss the content of each chapter in more detail.

**1 Introduction.** In this chapter, we introduce the task of concept disambiguation and clustering and describe its significance for downstream applications. We provide the research questions that are addressed in this thesis and summarize the main contributions. This chapter also contains this thesis overview and points to our previously published work.

**2 Linguistic Background and Motivation.** In this chapter, we provide a linguistic analysis of the tasks and one of its determining factors, i.e. the context definition. From a semiotic and discourse perspective, we lay the ground for the proposed approach (Question 1 and Question 2).

**3 Wikipedia as an Inventory.** Concept disambiguation requires an inventory. In this chapter, we compare different inventories based on the insights gained in the previous chapter and justify why we use Wikipedia. The selection of the inventory influences the method – e.g. with respect to available training data –, the features (Question 3) and the scalability across languages (Question 4).

**4 Method.** In this chapter, we describe how the linguistic findings (Chapter 2) are implemented using Markov logic. We discuss different strategies to model interdependencies between labels and provide an overview over Markov logic. We then introduce the backbone

of our approach. As we model the context definitions as latent variables, the parameter estimation needs to be adapted accordingly. In this respect, we follow the approach of Poon & Domingos (2008) proposed in the context of coreference resolution. This chapter mainly addresses Question 3.

**5 Features.** This chapter contains a description of our features for concept disambiguation, concept clustering and the context selection. The features influence both the method (Question 3) and the scalability across languages (Question 4).

**6 Architecture.** While all previous chapters exclusively focus on the effective disambiguation and clustering step, this chapter outlines our whole end-to-end system including the pre-processing, the mention recognition and the candidate concept identification (Question 3).

**7 Multi- and Cross-lingual Concept Disambiguation and Clustering.** In this chapter, we discuss how the proposed approach developed for English scales across languages. We discuss different strategies to port a system along the multi- and cross-lingual dimensions and reveal assumptions behind the proposed approach (Question 4).

**8 Data.** For concept disambiguation and clustering, many different data sets are annotated. We discuss some criteria to select some of them and describe the data sets we use for training, development and testing in more detail.

**9 Experiments.** In this chapter, we compare different versions of our proposed approach to some baselines and other state-of-the-art approaches. We provide some error analysis and test the effectiveness of our approach (Questions 1–4).

**10 Related Work.** Disambiguation and clustering is a prominent topic in different fields, including lexical semantics, information extraction and information retrieval. In this chapter, we embed the proposed approach in the broader research context. We mainly focus on approaches that also use Wikipedia as an inventory.

**11 Conclusions and Future Work.** In this chapter, we summarize the results and insights drawn from this thesis and indicate directions for future work.

## 1.5 Published Work

Most of the ideas and contributions described in this thesis has been published earlier. The Markov logic based approach for joint concept and entity disambiguation and clustering has

Topic	Research Questions	Chapters	Previously Published Work
Joint approach for concept disambiguation and clustering using Markov logic	Question 1 Question 3	Chapter 2, 4, 5, 6	Fahrni & Strube (2012) Fahrni et al. (2013) Fahrni et al. (2014)
Discourse-aware approach with latent variables	Question 2 Question 3	Chapter 2, 4, 5, 6	Fahrni & Strube (2014)
Mutli- and cross-lingual extension	Question 4	Chapter 4, 5, 6, 7	Fahrni et al. (2011b) Fahrni et al. (2012) Fahrni et al. (2013) Fahrni et al. (2014)
Previously proposed graph-based approach	Question 2	Chapter 4, 5, 7	Fahrni et al. (2011b) Fahrni et al. (2012) Fahrni et al. (2011a) Monz et al. (2011) Nastase et al. (2011) Bronner et al. (2012)

Table 1.1: Relation between our previously published work and this thesis

been proposed in Fahrni & Strube (2012). We participated with this system in two shared tasks on monolingual English and cross-lingual Chinese and Spanish entity linking. These results are reported in Fahrni et al. (2013) and Fahrni et al. (2014). The latent variable model and the discourse-aware extension are presented in Fahrni & Strube (2014). Multilingual and cross-lingual aspects of the thesis are discussed in Fahrni et al. (2011b), Fahrni et al. (2012), Fahrni et al. (2013) and Fahrni et al. (2014). An efficient version of the proposed disambiguation approach has been integrated in the prototype of the EU Project *CoSyne* on multilingual content synchronization with Wikis. This project has built the starting point and frame for this thesis (Monz et al., 2011; Bronner et al., 2012). A description of a precedent graph-based approach has been described in Fahrni et al. (2011b), Fahrni et al. (2012) and in some project deliverables (Fahrni et al., 2011a; Nastase et al., 2011). This version has been extrinsically evaluated for cross-lingual alignment of comparable corpora (Nastase et al., 2011). Table 1.1 summarizes the relation between our previously published work and this thesis.

## Chapter 2

# Linguistic Background and Motivation

An aim of this thesis is to propose a linguistically motivated approach for concept disambiguation and clustering. We argue that analyzing the task from a linguistic perspective should precede the selection of the method. Even more, the linguistic requirements should drive the selection of the method. Although each method implies some constraints, e.g. in terms of complexity, which may disallow to model a task as intended, the method should at least approximate the linguistic requirements.

Following this credo, we analyze concept disambiguation and clustering from a linguistic perspective and lay the linguistic foundation for the modeling. One of the key factors for concept disambiguation and clustering is the identification of the context that is relevant for the mentions (Chapter 1). In order to achieve a better understanding of the context, a deeper understanding of the tasks, i.e. concept disambiguation and clustering, is required. Based on the well-known semiotic model from Ogden & Richards (1923), we thus further specify concept disambiguation and clustering (Section 2.1). Based on this model and Halliday & Hasan (1976), we investigate the tasks from a discourse perspective and discuss their relationship to cohesion. We then derive consequences for the modeling of the context (Section 2.2) and propose a discourse-aware concept disambiguation approach (Section 2.3). In Section 2.4, we summarize the linguistic insights and discuss the requirements with respect to the inventory and the method. How to implement or at least approximate these requirements using currently available resources and state-of-the-art machine learning techniques is then discussed in Chapter 3 and Chapter 4 respectively.

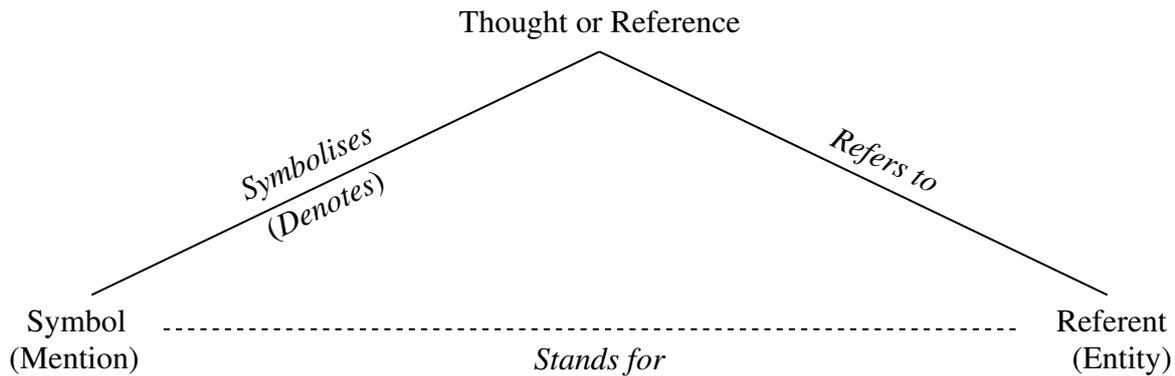


Figure 2.1: Semiotic triangle after Ogden and Richards (1923), p.11.

## 2.1 Approaching Concept Disambiguation and Clustering from a Semiotic Perspective

Following Eco (2002), we have defined a concept as a *cultural unit* that is denoted by a mention (Section 1.1.1). This definition emphasizes the social aspects of concepts, but is still vague and needs further specifications. In this section, we elaborate on the implications of defining a concept as a cultural unit and discuss how concepts relate to mentions, discourse entities and real-world entities (Section 2.1.1). To specify these relations, we take a semiotic perspective considering each mention being part of a linguistic sign. We then broach the controversial discussion if common nouns and proper names can be explained by one single sign type and discuss how we treat them in this thesis (Section 2.1.2). The implications for concept disambiguation and clustering are summarized in Section 2.1.3.

### 2.1.1 Mentions, Concepts and Entities

From a semiotic perspective, mentions can be considered as parts of linguistic signs. Different binary and triadic models have been proposed to analyze signs (Bußmann, 2002, p. 761). In this thesis, we build upon the triadic model of Ogden & Richards (1923), also known as the *semiotic triangle*. The semiotic triangle is helpful, as it distinguishes between different linguistic levels that are relevant for this thesis. Figure 2.1 shows the semiotic triangle as it has been proposed by Ogden & Richards (1923) in the early twenties of the last century, supplemented with the terminology used in this thesis in parentheses. The model comprises three components: the *symbol* equals in our case the mention; the *thought* or *reference* corresponds to our notion of concept and the *referent* is what we call the entity.

**Mentions.** The mentions are the surface forms in a text we aim to disambiguate or cluster. They are the only observed parts. In this discussion, we focus on common nouns and proper

names. However, also other lexical units in a text such as e.g. verbs can be part of linguistic signs.

**Concepts.** What a concept is or – to formulate it more generally – what the *meaning* of a word is, has been studied since and even before Aristotle. Peters (1999) gives a concise overview over different ideas of communication and reveals at the same time different conceptions of meanings that have been developed since the antiquity. From Augustine’s introversive perspective to John Dewey’s conception of meaning as partaking to Wittgenstein’s famous formulation “Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache” (Wittgenstein, 2001, p. 43) – all these different views emphasize different aspects of meaning and are influenced by the respective prevailing philosophic streams. In this thesis, we refrain from renegotiating this tremendous question of what the meaning of meaning is, but focus on a definition that is operational and suitable for automatic text processing.

By considering a concept as a cultural unit, we define it as a unit that is *shared* by a social community. We are only interested in the parts of meaning that are not idiosyncratic to one single person, but that are generally accepted by a community, i.e. that are ‘institutionalized’ (Eco, 2002, p. 66). Schneider (1968, p. 2) defines a unit as “simply anything that is culturally defined and distinguished as an entity. It may be a person, place, thing, feeling, state of affairs, sense of foreboding, fantasy, hallucination, hope or idea. In American culture such units as uncle, town, blue (depressed), a mess, a hunch, the idea of progress, hope, and art are cultural units.” To refine what *culturally defined and distinguished* implies we introduce the common linguistic distinction between denotation and connotation. The *denotation* is the fraction of meaning that forms the literal or core meaning and that is usually summarized in dictionary entries (Bußmann, 2002, p. 152). The denotation of CURTAIN could be paraphrased by “a piece of cloth intended to block or obscure light”<sup>1</sup> and is distinguishable from similar cultural units such as SHUTTER, which is a “solid and stable window covering usually consisting of a frame of vertical stiles and horizontal rails”<sup>2</sup>. The denotation has the potential to evoke itself other cultural units (Eco, 2002). Following Eco (2002, p. 108), we consider all these cultural units that the denotation can evoke in the recipient as the *connotation*. The connotation encompasses properties of the denoted cultural unit, hyponyms, hypernyms, antonyms and emotive associations. Connotations for CURTAIN are for example LACE or POLYESTER – typical materials curtains are made of – or SLEEPING, SEPARATION and DISRUPTION. Such connotations can originate metaphors: in the case of CURTAIN for example IRON CURTAIN. Both denotations and connotations are cultural, i.e. shared in a community and not idiosyncratic to the experience of a single person.

In this thesis, we consider the denotation as primary. For instance, we say that *dough* and

<sup>1</sup>English Wikipedia, <http://en.wikipedia.org/wiki/Curtain>, 18.2.2014.

<sup>2</sup>English Wikipedia, [http://en.wikipedia.org/wiki/Window\\_shutter](http://en.wikipedia.org/wiki/Window_shutter), 18.2.2014.

*money* denote the same concept, although their connotations only partially overlap. However, a concept description ideally also accounts for connotative aspects, as connotations are relevant for text analysis. They often strengthen textual cohesion, i.e. establish ties between different text parts, and contribute to the subtext. But in contrast to the denotation, connotations are optional (Eco, 2002, p. 67) and only selectively relevant in certain contexts: connotations are more like a cloud of shared – i.e. culturally available (Eco, 2002, p. 108) – associations that are potentially – but not necessarily – activated in context. We therefore prefer the term *connotative potential* in the following.

**Entities.** Entities can either be considered as *real-world entities*, i.e. real-world objects and persons with an extralinguistic existence, or as *discourse entities*, i.e. “components of the discourse context rather than elements of the (concrete) world” (Cumming, 2008, p. 1).

**Relations between Mentions, Concepts and Entities.** The main claim of the semiotic triangle is that symbols (mentions) and things (entities) are not directly connected, but intermediated by thoughts (concepts). Hence, we say:

**Definition 2.1** *A mention denotes a concept, while the evoked concept can refer to a specific (discourse) entity.*

These two relations are indicated in Figure 2.1 by lines. In this thesis, we only consider mentions, concepts and discourse entities. How discourse entities or concepts are societally (De Vault et al., 2006) or perceptually grounded in real-world entities and how language can be used to point to real-world entities is a topic on its own and not explored in this thesis (cf. Putnam (1975) or Devitt & Sterelny (1999)). We assume that what Karttunen (1976) said about coreference resolution also applies to concept disambiguation and clustering, i.e. that it is “a linguistic problem and can be studied independently of any general theory of extralinguistic reference.” (Karttunen, 1976, p. 366).

Figure 2.2 shows an excerpt from our running example (Figure 1.1). We marked all mentions (including pronouns) that share the same string, the same concept or the same discourse entity with another mention in the text snippet. The table on the right side of the figure shows all three levels. We added determiners in parentheses to the mentions as they are relevant with respect to discourse entities. It illustrates that not all mentions that are denoting a concept are referring to an entity. Generics such as *photos* or *men* for example are not considered as referring here. Pronouns on the other hand lack a concept. At the same time, the table indicates that a coincidence on one level does not imply a coincidence on another level. For instance, (*the*) *online-dating app* and *Tinder* share the same discourse entity, but not the same concept. In contrast, *Tinder* and *Tinder* correlate on all levels. We will discuss such correlations in more detail in Section 2.2.

### 2.1.2 The Role of Proper Names

The semiotic triangle is not only useful to explain the relationship between mentions, concepts and discourse entities, but also to describe the difference between common nouns and proper names. This distinction highly correlates with the distinction between *classes* (a-box) and *instances* (t-box) commonly made in research related to ontologies. The relevant question in this thesis with respect to proper names is “Do proper names have a sense” (Searle, 1958, 166), i.e. do proper names denote a concept or do they directly refer to an entity. This question has been extensively discussed by philosophers such as Frege, Russel and Searle and is still unsolved (for an overview see Van Langendonck (2008)). While for example Frege argues that proper names denote a sense which determines the entity (in the terminology of Frege the *meaning*, Frege (1892)), other theories postulate proper names as directly referring expressions (e.g. Kripke (1980)). Such theories deny proper names any denotation or connotation (as introduced above) and argue for a direct connection between a proper name mention and a discourse or often real-world entity. For instance, Ullmann (1962, p. 77) states: “The essential difference between common nouns and proper names lies in their function : the former are meaningful units, the latter mere identification marks.” We beware of discussing the differ-

[...] The new law comes after the **online-dating app** **Tinder** saw a surge of **photos** – mainly of **men** – posting profile pictures of **themselves** next to **tigers** and other big cats.

[...] that one in 10 **photos** on **Tinder** has a **tiger** in it – perhaps because the **men** want to appear adventurous to potential partners.

Mention	Concept	Entity
<i>(the) online-dating app</i>	ONLINE DATING SERVICE	entity <sub>1</sub>
<i>Tinder</i>	NIL_1	entity <sub>1</sub>
<i>photos</i>	PHOTOGRAPH	
<i>men</i>	MAN (MALE)	entity <sub>2</sub>
<i>themselves</i>		entity <sub>2</sub>
<i>tigers</i>	TIGER (ANIMAL)	
<i>(10) photos</i>	PHOTOGRAPH	
<i>Tinder</i>	NIL_1	entity <sub>1</sub>
<i>(a) tiger</i>	TIGER (ANIMAL)	
<i>(the) men</i>	MAN (MALE)	

Figure 2.2: Comparison between mentions (including pronouns if relevant), concepts and discourse entities.

ence between common nouns and proper names in all its facets in this thesis, but in order to decide how to model proper names in the context of automatic disambiguation and clustering, we need to consider at least the two main options.

One option would be to consider proper names as directly referring expressions (Kripke, 1980) implying that a proper name does not denote a concept, but directly points to a discourse entity. In this case, proper names could only be analyzed as discourse entities and concept disambiguation and clustering would be omitted. In terms of modeling, a common representation of proper names and common nouns on the concept-level would be missing. The other option would be to assume that proper names denote cultural units, i.e. concepts, as proposed by for example Eco (2002). In this case, we would assume a common representation of common nouns and proper names both on the concept and the entity level.

In this thesis, we argue that with respect to modeling, it is preferable to assume that proper names denote concepts similar to common nouns (second option). For instance, many cohesive relations between proper names and common nouns are difficult to explain without a common representation at the concept level. Without this common concept-level representation, it is unclear how connotations associated with proper names are accessible for the analysis. However, these connotations are essential to explain why for example the following two sentences are coherent:

(2.1) *Federica: Food is so important for me.*

*Leander: You should move to Lyon.*

These two sentences are partially conceived as coherent because FOOD and LYON are related. How this relatedness can be explained on the level of discourse entities is unclear. As *food* is used generically in this sentence, it is disputable if FOOD refers to a discourse entity. If we assumed FOOD does not refer any entity and we further assumed that proper names are directly referring expressions, no common representation for the two noun phrases would be available at all. How to calculate relatedness between the two noun phrases without a common representation and thus establish a connection between them is unclear. However, if the proper name *Lyon* denotes a concept, we could partially explain why Lyon is mentioned: Lyon denotes a cultural unit with culturally shared connotations such as GASTRONOMY which is highly related to FOOD.

This example demonstrates that there is a need for a common representation for both common nouns and proper names to explain cohesive relations between them. In this thesis, we thus assume that proper names denote cultural units, i.e. concepts. We further assume that all the differences between proper names and common nouns relevant to this thesis can be explained by the fact that concepts denoted by proper names usually refer to one single entity across text boundaries, while this is not the case for common nouns. This difference is illustrated in Figure 2.3. Common nouns, especially sortal nouns, classify discourse entities

into different sorts (Löbner, 1985). For example PHOTOGRAPH can be used in one single discourse to refer to the photograph of a famous actor, to some photographs of a tiger and to a photograph of an actor with a tiger. In contrast, proper names denote concepts that refer to one single entity across text boundaries. For instance TINDER (APPLICATION) refers to the same discourse entity across documents. However, the transition between common nouns and proper names is fluent. As the following example illustrates, proper names may have the potential to refer to multiple discourse entities:

(2.2) *Many people daily use bar-bells. But only a few turn into an Arnold Schwarzenegger.*

It is important to notice that most concepts denoted by proper names are not shared by a large community such as in the case of e.g. JAQUES CHIRAC, but by a smaller community such as in the case of LORENZO SCHMID (ICE-HOCKEY PLAYER). However, from an analysis point of view the size of the community that shares a concept denoted by a proper name is irrelevant (see also Eco (2002, p. 105)).

### 2.1.3 Summary

Given these considerations, we can now refine the definition of the tasks we investigate.

**Definition 2.2** *Concept disambiguation* is the task of linking both common noun and proper name mentions to their denoted concepts represented in an inventory.

**Definition 2.3** *Concept clustering* is the task of grouping common noun and proper name mentions into clusters, so that all mentions in a cluster denote the same concept.

Both tasks connect the mention level with the concept level and are neither localized on the discourse entity nor on the real-world entity level. In the context of proper name disambiguation usually the term *entity* is used, e.g. in *entity disambiguation*, *entity linking* or *entity clustering*. Most contributions lack an exact definition of what an entity is. For instance, in the knowledge base population task at TAC the task of entity linking is vaguely described with “[...] automatically identify salient and novel entities, link them to corresponding Knowledge Base (KB) entries [...]” (Ji & Grishman, 2011, p. 1149). It is not clear whether an entity is a discourse entity, a real-world entity or a concept; although formulations such as the following one indicate that a real-world entity is meant: “This requires the ability to link individuals mentioned in a document, and information about these individuals, to entries in the data base” (Ji et al., 2010, p. 1). In order to avoid any misunderstandings we consequently use the term *entity* only in the sense of *discourse* or *real-world entity* in this thesis.

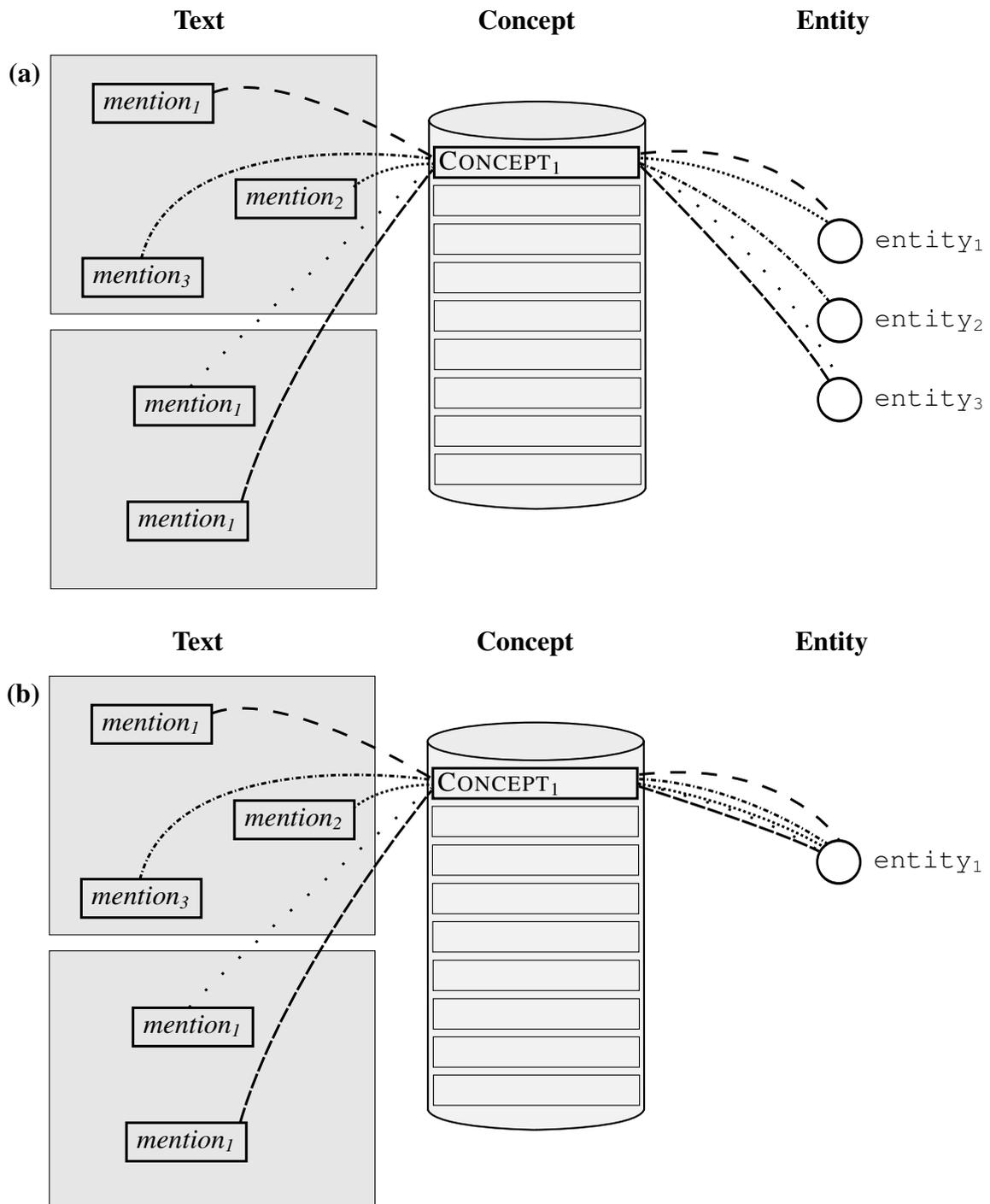


Figure 2.3: Difference between common nouns (a) and proper names (b): While concepts denoted by common nouns can be used to refer to different entities within and across documents (a), concepts denoted by proper names usually refer to one single entity (b).

## 2.2 Approaching Concept Disambiguation and Clustering from a Discourse Perspective

In the last section, we analyzed concept disambiguation and clustering from a semiotic perspective and mainly focused on mentions in isolation. However, mentions interact with other mentions and are embedded in a text. In this section, we thus approach concept disambiguation and clustering from a discourse perspective.

We first introduce a mention-centric model for cohesion based on Halliday & Hasan (1976) and the semiotic triangle (Section 2.2.1). This model allows us to systematically describe the relation between concept disambiguation and concept clustering and to obtain a better understanding of the context that is relevant to disambiguate mentions (Section 2.2.2). In addition, it helps us to analyze related tasks and to study how they can be exploited for concept disambiguation and clustering (Section 2.2.3 and Section 2.2.4). We conclude this section with a discussion (Section 2.2.5). This section thus lays the ground for both our proposed context model (Section 2.3) and the features (Chapter 5).

### 2.2.1 Semiotics Meets Discourse

Semiotics mainly focuses on single signs. While semiotic models (such as the semiotic triangle we discussed in the last section) are useful to analyze aspects of a single mention, they do not describe the relations between multiple mentions. However, such relations are essential for disambiguation and clustering, as they define the mention's context. This is where discourse comes into play. We build upon the notion of cohesion (Halliday & Hasan, 1976) and combine this discourse view with the semiotic model described in the last section.

Halliday & Hasan (1976, p. 4) define *cohesion* as “relations of meaning that exist within the text, and that define it as a text” (p. 4). For them, cohesion is a *semantic relation* and occurs where the interpretation of some element in the discourse is dependent on the interpretation of another element. Halliday & Hasan (1976) distinguish between grammatical cohesion (reference, substitution and ellipsis), lexical cohesion (reiteration and collocation) and cohesion through conjunction, which is between grammatical and lexical cohesion. While Halliday & Hasan (1976) aim to sub-classify cohesive relations based on the common linguistic distinction between grammar and lexicon, we approach cohesion from a semiotic perspective. We will show that such a semiotic perspective is in our case more valuable.

In contrast to Halliday & Hasan (1976), we have a broader notion of cohesion that is not restricted to semantic relations:

**Definition 2.4** *Cohesive relations are relations that exist within the text, and that define it as a text.*

**Definition 2.5** A (*cohesive*) *tie* is one instance of a cohesive relation between two items (Halliday & Hasan, 1976).

While cohesive ties occur on various linguistic levels, e.g. between words, subclauses, sentences or even whole text parts, we focus here on ties between mentions. To systematically study such mention-level cohesive ties, we start from the semiotic triangle introduced in Section 2.1. Each mention in a text corresponds to a sign and can be approximated using the semiotic triangle. From a purely mention-centered perspective, a text is a sequence of mentions, each consisting of the mention tokens and optionally a denoted concept and a referred discourse entity. Figure 2.4 schematically shows a text from a mention-centric perspective. Each mention in this text denotes a concept and refers a discourse entity (except *mention*<sub>5</sub> which lacks an entity). *CONCEPT*<sub>1</sub> indicates the concept denoted by *mention*<sub>1</sub> and *entity*<sub>1</sub> represents the entity referred to by *mention*<sub>1</sub>. We assume that on all three levels, i.e. on the mention, concept and entity level, cohesive ties can occur. They are illustrated in Figure 2.4 with red and blue solid lines. These ties can be classified into specific types, but we only distinguish here between two main relation types: identity and relatedness.

**Definition 2.6** A cohesive tie of type *identity* ties two items that are identical (same string or lemma, same concept or same entity).

The identity relations are represented by blue lines in Figure 2.4.

**Definition 2.7** A cohesive tie of type *relatedness* occurs between two items that are related, but not identical.

The relatedness relations are represented by red lines in Figure 2.4. We subsume similarity relations under this type. While identity relations are binary – they are either present or missing –, relatedness relations are gradual. Their strength is indicated by the thickness of the lines in Figure 2.4. An identity relation can also be seen as a special case of a relatedness relation, where the relatedness between two items is maximal and the two items are thus identical.

Table 2.1 gives for each level and type an example taken from our text on tiger selfies (or if not available a made up one). These examples are discussed in the following.

**String-level Cohesive Ties.** Cohesive ties on the mention level of type identity occur between mentions of the same string. Cohesive ties of type relatedness occur for instance if two mentions are derived from the same word stem (e.g. *photo* – *photography*). Another form of relatedness is co-occurrence. If the strings of two mentions often co-occur, we consider them as related. But also for instance alliterations or end rhyme contribute to cohesion and thus could be summarized under this type.

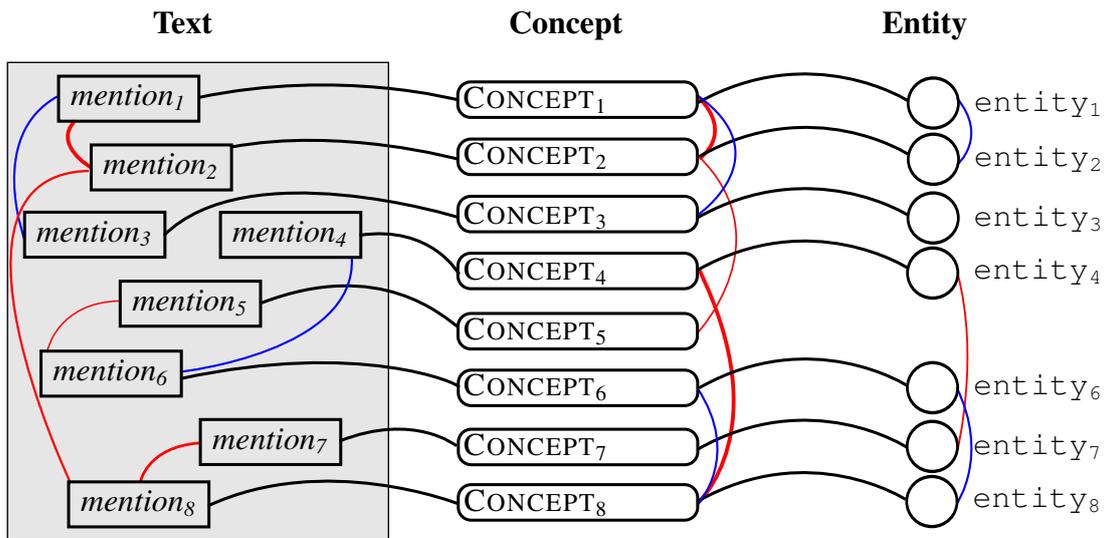


Figure 2.4: Cohesive ties on the string (left), concept (middle) and entity level (right). CONCEPT<sub>1</sub> is the concept denoted by mention<sub>1</sub>. entity<sub>1</sub> is the entity referred to by mention<sub>1</sub>. Cohesive ties of type identity are represented by blue lines. Cohesive ties of type relatedness are in red. Their thickness represents the degree of relatedness.

**Concept-level Cohesive Ties.** On the concept level, relatedness has many facets, e.g. hypernymy – in this case, one usually talks about similarity –, co-hyponymy, but also causation, activation, prohibition and concept co-occurrence relations belong to this type. Concept-level relatedness relations often depend on the connotative potential of the involved concepts and the relations are often difficult to further specify.

**Entity-level Cohesive Ties.** Discourse entities can exhibit relatedness, e.g. in terms of part-of or location-based relations. For instance in the text on selfies, the men are located next to tigers on the pictures.

Items	Relation Type	Example
Mention strings	Identical	<i>photos – photos</i>
	Related	<i>photo – photography</i>
Concepts	Identical	PHOTOGRAPH – PHOTOGRAPH
	Related	PHOTOGRAPH – TIGER SELFIE
Discourse Entities	Identical	online dating app – Tinder
	Related	Men – Tigers

Table 2.1: Examples for cohesive ties of type identity and relatedness on the string, concept and entity level.

Two mentions can be tied on all three levels, or only on a subset of them (e.g. only on concept level). If two mentions exhibit ties on more than one level, the relation types can be

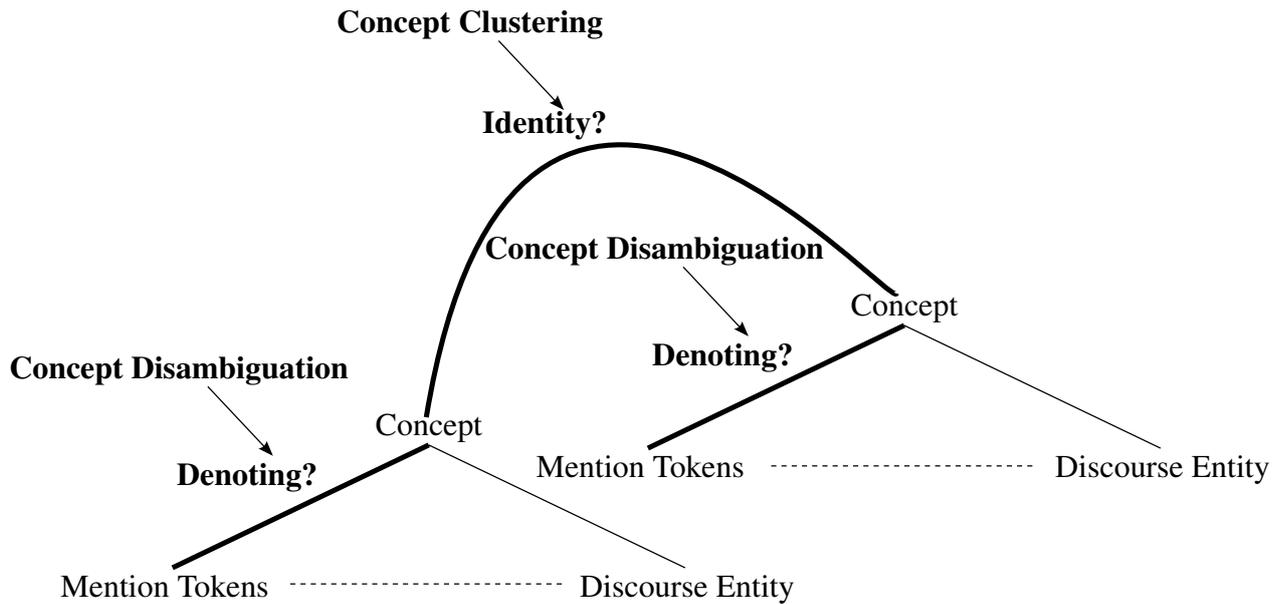


Figure 2.5: Concept disambiguation and concept clustering – two tightly connected tasks.

different across the different levels. For instance, an identity relation can exist between the corresponding discourse entities, while the respective concepts are connected by a relatedness relation.

In the following, we discuss concept-level cohesive ties in more detail and derive some consequences for the modeling of concept disambiguation and clustering. We then discuss correlations between concept-level cohesive ties and cohesive ties on the entity and the string level.

## 2.2.2 Cohesive Ties between Concepts

Cohesive ties on the concept level are essential for concept disambiguation and clustering. Given the mention-centric model, we can redefine concept clustering with respect to cohesive ties.

**Definition 2.8** *Concept clustering* is the task of identifying identity relations between common noun and proper name mentions on the concept level within and across documents.

Concept disambiguation lacks a direct correspondence to specific cohesive ties, as the aim is to identify a *denote*-relation between some mention and a concept. Nevertheless, cohesive ties are crucial to automatically approach concept disambiguation. We first discuss identity relations and then relatedness relations in more detail.

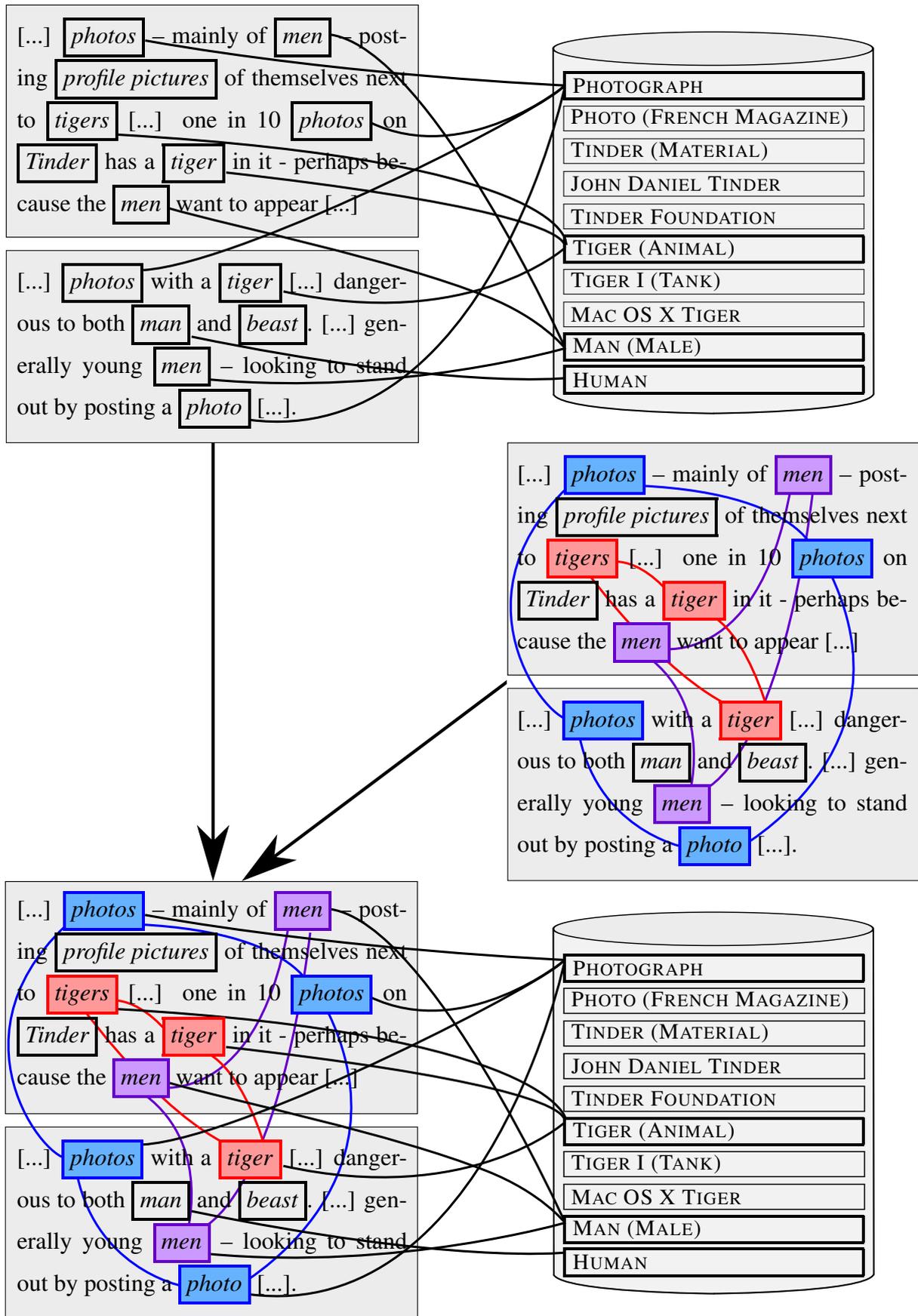


Figure 2.6: On the top: resolving ambiguity by linking mentions to concepts in an inventory; in the middle: resolving ambiguity by clustering mentions according to concept-level identity relations; on the bottom: combined perspective.

**Concept-level Cohesive Ties of Type Identity.** If two mentions are tied on the concept level by an identity relation, they must denote the same concept. At the same time, if two mentions denote the same concept, they must exhibit an identity relation on the concept level. This also implies that the negative formulation must be true: if two mentions lack an identity relation on the concept level, they cannot denote the same concept and if two mentions denote different concepts, they lack an identity relation on the concept level. Hence, concept disambiguation and concept clustering are tightly connected tasks: concept disambiguation decisions directly influence concept clustering decisions and vice versa. Both tasks approach the ambiguity problem, but from a different perspective as Figure 2.5 illustrates. In disambiguation, for each mention a *denote*-relation between the mention and a concept is identified. Thus, ambiguity and variability are resolved by linking. In clustering, ambiguity and variability are approached by identifying identity relations on the concept level. While these two tasks have been seen as two alternative tasks in previous research (Chapter 10), we argue that they are closely tied together and that they can support each other.

**Observation 2.1** *Concept disambiguation decisions and concept clustering decisions mutually influence each other. Hence, modeling them jointly can be beneficial for both of them.*

In Figure 2.6, we show how the two tasks interact using our running example. On the top, mentions are linked to their corresponding concepts (disambiguation). In the middle part, the mentions that denote the same concepts are clustered, indicated by lines (clustering). On the bottom, the two perspectives are combined. As the example illustrates, we assume that the granularity obtained via clustering corresponds to the granularity of the concepts in the inventory.

**Concept-level Cohesive Ties of Type Relatedness.** Not only cohesive ties of type identity are relevant for disambiguation, but also cohesive ties of type relatedness.

In work on disambiguation, it is usually (implicitly) assumed (1) that texts are cohesive, (2) that concept-level cohesion is a highly relevant form of cohesion and (3) that texts thus generally show concept-level cohesive ties (e.g. Cucerzan (2007) and Navigli & Lapata (2010)). Taking these three assumptions for granted, the crucial question is how concept-level cohesion can be exploited for disambiguation. In order to study the relationship between cohesion and disambiguation, let us assume that we can calculate a score that measures the relatedness of two concepts and thus approximates how strongly they are tied: the higher this score is, the stronger is the cohesive tie between the two concepts. Let us further assume that for all mentions in the text – except for one mention, which is our target mention – it is known which concepts they denote. Our aim is to disambiguate the target mention. Given this situation, one could argue that the candidate concept of the target mention that shows the strongest

concept-level cohesive ties with the concepts of the other mentions should be considered as the concept the target mention denotes. Selecting this concept makes the text more cohesive.

Figure 2.7 illustrates this situation using our running example. The mention *big cat* is our target mention. The concept-level cohesive ties are represented by (blue) edges. The thicker an edge is the higher is the relatedness score associated with it and the stronger is thus the cohesive tie between the respective mentions. To calculate relatedness, we used the relatedness measure based on the link structure from Milne & Witten (2008a). Among others, *big cat* has the candidate concepts BIG CAT and CURTIS HUGHES. Figure 2.7 shows on the left-hand side the cohesive ties for the candidate BIG CAT and on the right-hand side the ones for CURTIS HUGHES who has the nick name *big cat*. As this example illustrates, the actually denoted concept BIG CAT is not only tied to more other concepts, but the individual ties are also stronger.

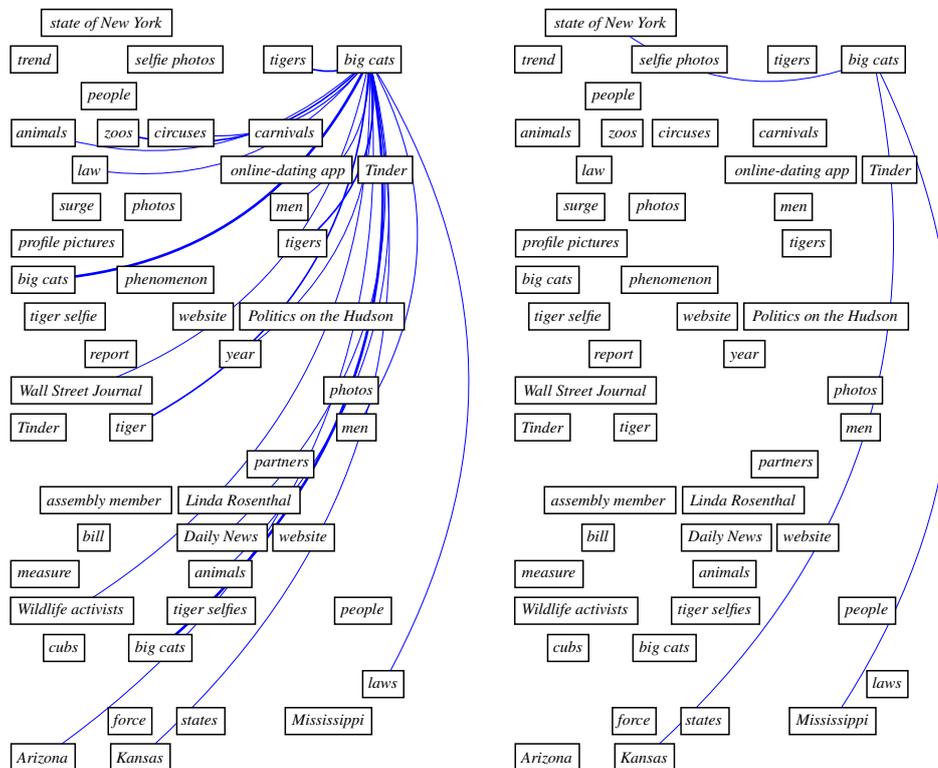


Figure 2.7: Concept-level cohesive ties for two concepts. It is assumed that the concepts are known for all mentions – except for *big cats*. On the left-hand side, the concept-level cohesive ties for the concept BIG CAT are depicted. We inserted an edge between *big cats* and another mention, if the relatedness score between the concept BIG CAT and the concept denoted by the other mention is higher than 0. On the right-hand side, an edge is established between *big cats* and another mention, if the relatedness score between the concept CURTIS HUGHES and the concept denoted by the other mention is higher than 0. We used the pairwise relatedness measure of Milne & Witten (2008a). Hence, the two figures approximate the concept-level cohesive ties for the two candidate concepts BIG CAT (left side) and CURTIS HUGHES (right side).

The idea of selecting the candidate concepts so that the concept-level cohesive ties between them are maximal underlies many disambiguation approaches (e.g. Cucerzan (2007), Navigli & Lapata (2010)). While some approaches exclusively rely on concept-level cohesive ties – approximated by relatedness measures –, others additionally include information such as string-level features (Ratinov et al., 2011). We applied such an approach that is inspired by Navigli & Lapata (2010) to our running example. We first calculated for each candidate concept of a mention its relatedness to each candidate concept of all other mentions and then summed up these scores. For each mention, the candidate concept with the highest summed relatedness score is then chosen. This method that exclusively relies on concept-level cohesive ties approximated by a relatedness score disambiguates 65.3% of the mentions in our running example correctly. The question is why these results are so low. The relatedness scores that are automatically calculated based on the link structure in Wikipedia may introduce some noise. However, this does not fully explain the low results. By revising the assumptions made by our simple approach, we identified two main flaws that affect many disambiguation approaches.

We implicitly assumed that high relatedness scores mainly occur between concepts that are actually denoted by the mentions. This assumption turns out to be wrong, as non-denoted concepts can be highly related with each other. For instance, QUANTUM STATE, a candidate of *state*, FORCE (PHYSICS), a candidate of *force*, or PHYSICAL LAW, a candidate of *law*, are all highly related to MEASURE (MATHEMATICS) or MEASUREMENT, two wrong candidate concepts of *measure*. PEOPLE (MAGAZINE), a candidate of *people*, and MEN (MAGAZINE), a candidate concept of *men*, are also highly related and the band THE ANIMALS shows high relatedness with other bands such e.g. KANSAS (BAND). Figure 2.8 illustrates that high relatedness scores between wrong candidate concepts are frequent. We calculated pairwise relatedness between the candidates of each context mention and the candidates of the mention *laws*. For each mention pair, we selected the highest relatedness score. If the candidate concepts associated with this score were the concepts that are actually denoted by the mentions, we added a blue line, otherwise a red line. To calculate relatedness, we used the same relatedness measure as above (Milne & Witten, 2008a). In the extreme case as in Figure 2.8, the highest relatedness score between the candidates of a mention pair involves almost always wrong candidate concepts. If we repeat our experiment from above and only consider the actually denoted concept of each context mention instead of all candidate concepts, 73.5% of the mentions are disambiguated correctly instead of 65.3%. These results are better, but still low given the fact that all context mentions are considered as correctly disambiguated. This leads us to the second flaw concerning our assumptions.

A common assumption in disambiguation (e.g. Cucerzan (2007), Navigli & Lapata (2010) and Ratinov et al. (2011)) including our simple approach is the *assumption of maximal cohesiveness*. More precisely, the assumption is that the concept-level cohesion in a text should be as high as possible. Thus, the concept which exhibits the highest relatedness with the other

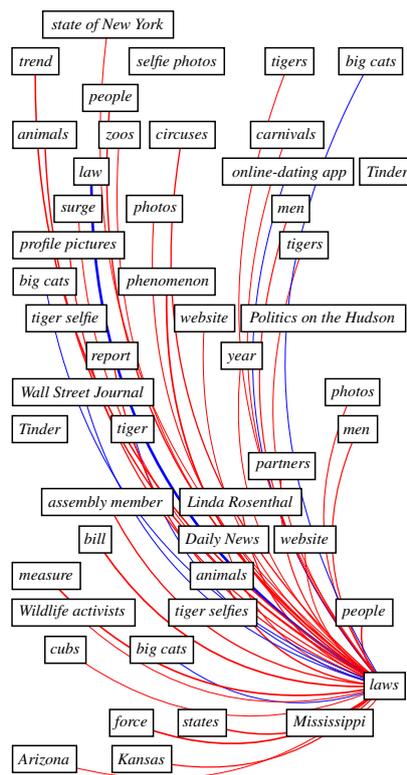


Figure 2.8: Relatedness relations for the candidate concepts of *laws* to other candidate concepts in the context: for each mention pair that involves *laws* we show the highest relatedness score between the respective candidate concepts. If this highest relatedness score involves the concepts that are actually denoted by the respective mentions it is in blue, otherwise it is in red.

concepts in the context must be the correct one. What is essential in such an approach is the definition of the context. Most approaches use one single context definition such as the whole document or a context window of a predefined fixed size. This context definition is then uniformly applied to all mentions. In our simple approach, we consider the concepts of all mentions in the document as context. By using such a uniform context definition, we implicitly assume that all mentions can be disambiguated using the same context definition. This is a common assumption underlying almost all disambiguation approaches. To revise this assumption, we first need to better understand how to define the context relevant to disambiguate a mention.

**Definition 2.9** *The context relevant to disambiguate a target mention consists of all concepts and text parts (lexical units, syntactic constructions etc.) that help to disambiguate it.*

To better understand the context, we need to specify what *help to disambiguate* means. To disambiguate a target mention, the concept denoted by another mention – henceforth called

the context concept – is helpful if the context concept and the concept that is actually denoted by the target mention are highly related. In this case, the two mentions disambiguate each other, as they are likely to denote the candidate concepts that are maximally related with each other. Given our former definition of cohesive ties, we thus can say:

**Definition 2.10** *The context relevant to disambiguate a mention (in terms of other mentions) comprises all other mentions with which it shares concept-level cohesive ties and the syntactic context to ensure local context compatibility (e.g. subcategorization restrictions, syntactic clues).*

Hence, if a mention exhibits a cohesive tie to another mention on the concept level, the two mentions are mutually relevant for their respective disambiguation in the sense that the selected concepts should be related or identical. However, to decide if mentions show cohesive ties on the concept level, they must be disambiguated, but to disambiguate mentions their concept-level cohesive ties are relevant. This leads to our second observation:

**Observation 2.2** *Identifying the relevant context (in terms of mentions) and resolving ambiguities of mentions are interconnected tasks and not separable from each other.*

For instance, if *bill* would denote BILL CLINTON in our running example, it would show different concept-level cohesive ties than if it denotes BILL (LAW). This mutual relationship between context definition and ambiguity resolution is one of the reasons why concept disambiguation is hard. This aspect has also been discussed by Asher & Lascarides (1995) in the context of the rhetorical structure theory.

Now that we have a better understanding of the context that is relevant to disambiguate a mention, we can investigate if all mentions should be disambiguated using the same context definition. As the context is defined by the concept-level cohesive ties a mention exhibits, this would imply that all mentions show the same amount of concept-level cohesive ties. However, this is wrong. Not all mentions equally contribute to the overall cohesiveness. Some mentions only glue the local text parts together, others the whole document. For instance, *force* in our running example shows completely different concept-level cohesive ties than for instance *tiger*. While the first is tied to the mentions in the surrounding context, the latter shows ties across the whole document. Hence, in the first case, the local context is relevant for its disambiguation, in the second case the relevant context spreads across the whole document. The following observation summarizes this insight:

**Observation 2.3** *Dependent on their exhibited concept-level cohesive ties, different mentions require different context definitions for their disambiguation.*

We thus conclude that, given a mention, the *assumption of maximal cohesiveness* is only true with respect to the context that is relevant to disambiguate the mention. If the same context definition is applied to all mentions, this assumption is wrong.

If the concept-level cohesive ties of a mention were known, disambiguation would be trivial. However, as we have discussed, they actually depend on the concepts denoted by the mentions. It is therefore essential to investigate if other factors can indicate if two mentions are tied on the concept level.

**Identification of Cohesive Ties.** Concept-level cohesive ties have been studied both in linguistics and computational linguistics. Halliday & Hasan (1976, p. 274) define *lexical cohesion* as cohesion “achieved by the selection of vocabulary”. Among other forms of cohesion they discuss, this is the closest one to our notion of concept-level cohesion. They classify lexical cohesion into two basic classes: *reiteration* includes repetition by using the same lexical item, a (near-) synonym, a hypernym or a general noun to refer to the same entity (p. 278); *collocation* is “cohesion that is achieved through the association of lexical items that regularly co-occur” (p. 284). While the definition for collocation is vague, we would classify reiterations as identity relations on the entity level.

Lexical cohesion has been discussed by different other authors (Martin, 1992; Tanskanen, 2006, *inter alia*). This work shows that it is difficult to further specify and sub-classify Halliday & Hasan (1976)’s collocation relations (an overview can be found in Tanskanen (2006)). Tanskanen (2006, p. 60–69) for example sub-classifies Halliday & Hasan (1976)’s collocation relation into three subcategories: ordered set relations (e.g. months or week days), activity-related collocations (e.g. *eat* and *meat*) and elaborative collocations for “all those pairs whose relation is impossible to define more specifically than stating that the items can somehow elaborate or expand on the same topic” (Tanskanen, 2006, p. 63). However, as Hoey (1991, p. 210) notices, “flexibility does not denote anarchy”; or to formulate it with Tanskanen (2006, p. 8): “A probabilistic stance is in effect the only feasible one, since the flexibility of lexical relations, like that of the patterns formed by lexical relations, entail potential novelty and therefore unpredictability.” Although these previous linguistic studies leave little hope that good indicators can be identified to predict which mentions tend to show concept-level cohesive ties and therefore disambiguate themselves, it might be possible to exploit the “collaborative nature of communication” (Tanskanen, 2006, p. 23). Tanskanen (2006, p. 66) states that “the association will be profitable to a text’s coherence only if it can be identified by the receiver”. This indicates that some textual clues might indicate cohesive relations.

**Observation 2.4** *Due to their collaborative nature, texts may contain some clues – e.g. with respect to the text organization – that allow to detect which mentions are related on the concept level.*

In computational linguistics, concept-level cohesive ties have been studied in work on lexical chains (Morris & Hirst, 1991; Remus & Biemann, 2013, *inter alia*).

**Definition 2.11** Lexical chains are “*sequences of related words*” (Morris & Hirst, 1991, p. 23).

Lexical chains have been used for e.g. summarization (Barzilay & Elhadad, 1997), text segmentation (Okumura & Honda, 1994) or malapropisms detection (Hirst & St-Onge, 1998). While Morris & Hirst (1991, p. 23) state that lexical chains “provide an easy-to-determine context to aid in the resolution of ambiguity and in the narrowing to a specific meaning of a word”, it turned out quickly that lexical chains are difficult to compute. In most approaches they are built based on a semantic resource such as a semantic network (e.g. WordNet) or a thesaurus. The assumption of all these approaches is that mentions in a text are only related with respect to the candidate concepts they actually denote. However, as we have seen above, mentions can be highly related with respect to their “wrong” candidate concepts. The wrong assumption that only “correct” concepts are highly related with each other is probably one of the main reasons why lexical chains fail to be successful for disambiguation (Nelken & Shieber, 2006).

As this discussion reveals, it is difficult to identify cohesive ties while only considering relatedness scores as for instance in work on lexical chains. A more promising direction to decide if two mentions are tied on the concept level might be to exploit correlations to cohesive ties on other levels. We therefore discuss in the following possible correlations between concept-level cohesive ties and cohesive ties on other levels.

### 2.2.3 Cohesive Ties between Entities

Cohesive ties on the entity level have been studied more extensively than concept-level cohesive ties. Correlations between these levels are therefore interesting, as they would help to identify concept-level cohesive ties. In the following, we discuss two tasks that address entity-level cohesive ties and investigate if these tasks might be beneficial to recognize concept-level cohesive ties.

**(Cross-document) Coreference Resolution.** Coreference resolution is the task of clustering mentions (e.g. common nouns, proper names and pronouns) in a document so that all mentions in a cluster refer to the same (discourse) entity (Cai & Strube, 2010). Cross-document coreference resolution is the task of clustering mentions across documents so that all mentions in one cluster refer to the same entity (Bagga & Baldwin, 1998b). We only consider coreference relations between common nouns and proper names. While concept clustering is concerned with

identity relations on the concept level, the aim in (cross-document) coreference resolution is to detect identity relations on the entity level.

We can identify the following correlations between concept-level and entity-level cohesive ties (see also e.g. Ponzetto & Strube (2006)):

**Observation 2.5** *If two common noun or proper name mentions show an identity relation on the entity level, they must be tied on the concept level, either by identity or similarity (which we summarized under relatedness above). The reverse is invalid: if two mentions show an identity or similarity relation on the concept level, they are not necessarily tied by an identity relation on discourse entity level.*

This correlation can be refined by considering the differences between proper name and common noun mentions. Given the observation that a concept denoted by a proper name mention usually refers to one single entity (Section 2.1), we can derive that for proper names the cohesive ties of type identity fully correlate on the concept and the entity level.

**Observation 2.6** *If two proper name mentions are coreferent, they must also show an identity relation on the concept level. At the same time, if two proper name mentions are tied by an identity relation on the concept level, they must be coreferent.*

This also implies that the negative formulation is true: if two proper name mentions are not coreferent, they cannot denote the same concept. In contrast to a concept denoted by a proper name mention, a concept denoted by a common noun mention can be used to refer to different discourse entities. This leads to an asymmetric relationship between concept-level and entity-level cohesive ties for common noun mentions.

**Observation 2.7** *If two common noun mentions are coreferent, they are tied on the concept level, but not necessary by an identity relation. In contrast, if two mentions exhibit an identity relation on the concept level, they are not necessarily coreferent.*

For common nouns, concept-level ties are a necessary, but not sufficient condition for coreference. The negative formulation is therefore invalid. If two common noun mentions are not coreferent, they still can be tied on the concept level by an identity or similarity relation. For instance, *law* and *laws* show an identity relation on the concept, but not on the entity level. Pairs consisting of a common noun and a proper name mention behave as pairs of two common nouns, except that they are necessarily tied by a similarity and not by an identity relation on the concept level, if they are coreferent.

**Relation Extraction.** Relation Extraction as the task of extracting relations between entities is concerned with cohesive ties of type relatedness. In relation extraction, usually tuples of the form (mention, relation-type, mention) are extracted. If mentions that are connected by such a relation also show a relatedness relation on the concept level depends on the extracted relation. If the extracted relation is a very discourse-specific relation (e.g. (*Vasella, read, newspaper*)), a cohesive tie on the concept level is less likely than for less discourse-specific relations such as (*Vasella, former-ceo-of, Novartis*).

By only considering certain relation types, relation extraction is a promising direction to identify mentions that are tied on the concept level and may disambiguate each other (Cheng & Roth, 2013).

### 2.2.4 Cohesive Ties on the String Level

Compared to cohesive ties between entities that can only be automatically determined by using advanced methods, string-level cohesive ties are straightforward to be detected automatically. Strong correlations between string- and concept-level cohesive ties are therefore even more appealing to investigate than correlations on the entity level. It is thus not surprising that string-level cohesive ties have been studied early in work on disambiguation. Gale et al. (1992c) and slightly later Yarowsky (1993) formulated two hypotheses concerning the correlation between string-level and concept-level cohesive ties.

**One Sense per Discourse Hypothesis.** The one sense per discourse hypothesis (Gale et al., 1992c) states that ambiguous words tend to denote the same sense within a discourse or a document, or in other words:

**Observation 2.8** *Mentions that are tied by an identity relation on the string level are likely to be tied by an identity relation on the concept level. The reverse is invalid.*

**One Sense Per Collocation Hypothesis.** The one sense per collocation hypothesis states that an ambiguous word denotes one sense per collocation (Yarowsky, 1993). A wide variety of collocation definitions have been explored in the disambiguation literature. Collocation-based features are probably the most widely used features in this area. However, to identify such collocations most approaches require concept-annotated data.

**Observation 2.9** *Mentions that exhibit a co-occurrence relation on the string level are likely to also exhibit a co-occurrence relation on the concept level.*

### 2.2.5 Discussion

In this section, we have analyzed concept disambiguation and clustering from a discourse perspective. This analysis has led to two main insights.

First, we showed that concept disambiguation and concept clustering are interrelated. If two mentions denote the same concept (disambiguation perspective), they share a concept-level cohesive tie of type identity (clustering perspective) and vice versa. This is only true given the assumption that the granularity of the clusters matches the granularity of the concepts in the inventory. After having discussed cohesive ties in more detail, we can now better understand why addressing the two tasks simultaneously might be beneficial. Figure 2.9 shows an example. We consider the two target mentions *law* and *laws*. On the left-hand side, a target mention is connected to a context mention if the relatedness score for the concepts denoted by these two mentions is higher than the relatedness scores between all other candidates of the two mentions. On the right-hand side, the mentions *law* and *laws* are connected to another mention if this condition is not met. As we see from this picture, *law* is easier to disambiguate as it has many more helpful connections than *laws*. If we know that the two mentions belong to the same cluster and thus should denote the same concept, it is more likely that we can also disambiguate *laws* correctly. The reason why the two mentions show these differences is due to their different candidate concepts. As they do not share the exact same string, different candidate concepts have been retrieved for them from the lexicon. If they shared exactly the same string, their candidate concepts would be the same and thus also their relatedness scores.

Second, we argue that the optimal context to disambiguate a mention (in terms of other mentions) includes all mentions with which it shares concept-level cohesive ties. However, the concept-level cohesive ties mentions depend on the concepts they denote, which we aim to determine, and vice versa. Because of this mutual relationship between concept-level cohesive ties and concepts, concept disambiguation is a challenging task. To obtain a better understanding of concept-level cohesive ties, we analyzed their correlations to entity- and string-level ties. This analysis has revealed some promising correlations. While these correlations are quite straightforward to be exploited for cohesive ties of type identity by using string or coreference relations, it is more challenging to use them to capture concept-level cohesive ties of type relatedness. Whereas some correlated phenomena are infrequent, e.g. some relations on the entity level that match a specific pattern might be rare, others require to resolve noun coreference beyond string match or to extract long-distance relations. These tasks are difficult by themselves and far from been solved yet.

In the next section, we therefore propose a more pragmatic approach to identify the context relevant to disambiguate a mention.

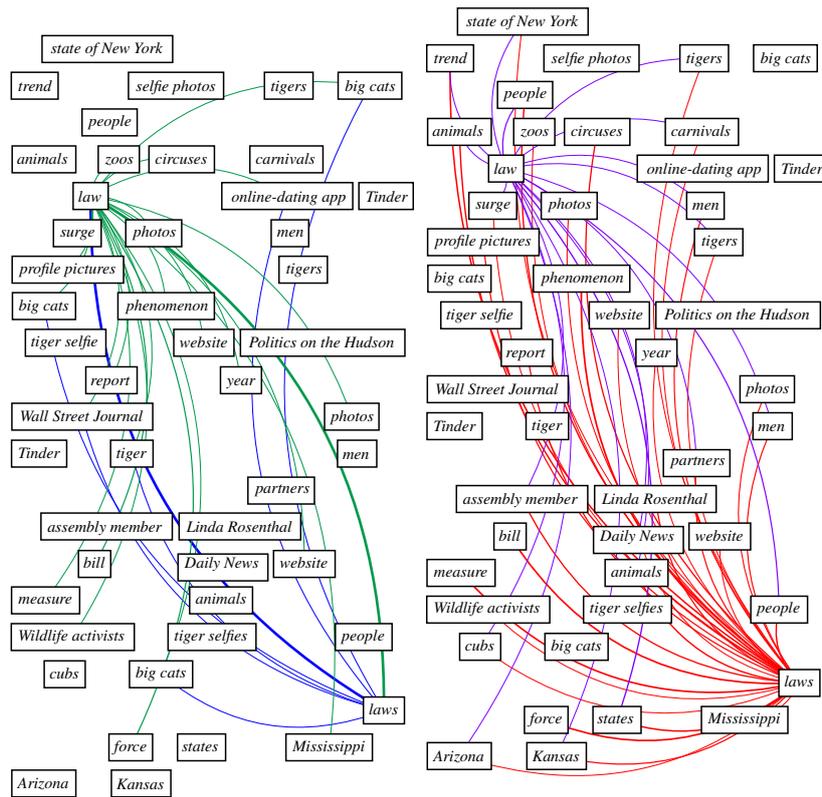


Figure 2.9: On the left-hand side, a target mention is connected to a context mention if the relatedness score for the concepts denoted by these two mentions is higher than the relatedness scores between all other candidates of the two mentions. The concept-level cohesive ties for *law* are in green, while the ties for *laws* are in blue. On the right-hand side, the mentions *law* and *laws* are connected to a context mention if the highest relatedness score involves concepts that are not denoted by the two mentions.

### 2.3 Identification of the Disambiguation Context

The key to concept disambiguation is to determine the context that is relevant to disambiguate a mention and to exploit this context effectively. In the last section, we have discussed the relationship between concept-level cohesive ties and the context that is relevant to disambiguate a mention. In fact, each mention would require an individual context definition that mirrors its concept-level cohesive ties. However, the concept-level cohesive ties of mentions depend on the concepts they denote and are thus difficult to identify.

In this section, we therefore propose a more pragmatic approach to identify the context that is relevant to disambiguate a mention. Instead of identifying for each mention its individual context, the idea of the proposed approach is to group mentions into bins according to their context behavior and to treat all mentions in the same bin alike. In the following, we first describe different context definition strategies (Section 2.3.1), before we discuss our proposed

binwise scope-aware approach (Section 2.3.2).

### 2.3.1 Uniform, Individual and Binwise Context Definitions

Previous work on concept disambiguation often uses a uniform context definition (Section 10.3).

**Definition 2.12** *A context definition is **uniform** if exactly the same context definition is applied to all mentions.*

A uniform context definition implies that the same information is required in the same extent to disambiguate or cluster all mentions. The assumption behind a uniform context definition is that the feature values account for the context relevant to disambiguate a mention. For instance, it is assumed that the relatedness scores are higher between the concepts that are denoted by the mentions. However, as we have discussed in the last section, this assumption often does not hold. Otherwise concept disambiguation, clustering and the identification of lexical chains would be trivial tasks and have been solved a long time ago (Section 2.2).

We argue that a uniform context definition is inappropriate, as it ignores the mutual relationship between the cohesive ties of a mention and its concept. Instead we argue that each mention requires an individual context definition that reflects its cohesive ties.

**Definition 2.13** *Context definitions are **individual** if they are tailored for each single mention.*

Such an individual context definition is difficult to implement, as it would require to identify the concept-level cohesive ties, which is a difficult task.

We propose to approximate the ideal context definition by a binwise approach. Instead of defining the context for each mention separately, we argue that mentions can be grouped into different bins according to their context behavior. For each bin, a separate context definition is used.

**Definition 2.14** *Context definitions are **binwise** if mentions are grouped into bins according to some criteria and a different context definition is used for each bin.*

Previous work that uses binwise context definitions assigns the considered lexical units before disambiguation to their respective bin, e.g. based on part-of-speech tags (Mihalcea & Csomai, 2005) or according to the similarity between words (Dhillon & Ungar, 2009). Most previous work that uses binwise context definitions determines the individual contexts type-based. For each lemma a separate model is applied. Different contexts can be differently weighted and by using feature selection methods different contexts can be selected. Supervised versions of this approach usually suffer from data sparsity: as training data is required

for all candidate concepts for each lemma, such supervised approaches are unscalable. But even if enough training data were available, such lemma-based approaches make inappropriate assumptions. Which context is relevant to disambiguate a mention depends on its denoted concept and its context-level cohesive ties and not on the lemma of a mention. The same lemma can trigger different context definitions depending on how it is embedded into discourse. If *tiger* highly contributes to the main topic as in our running example, it requires a completely different context definition than if *tiger* is mentioned casually in a text on a completely different topic. As we stated before, the context relevant to disambiguation or cluster a mention depends on its cohesive ties and vice versa, or more generally:

**Observation 2.10** *The context relevant to disambiguate a mention depends on how it is embedded into discourse and vice versa.*

Hence, the context relevant to disambiguate a mention does not only depend on its denoted concept, but also on the specific discourse context. That the discourse structure may play a role in disambiguation has also been stated by Gale et al. (1992d). They did some studies on the context window size and were surprised that for nouns large window sizes can be useful. They suggest that there might be a relation between disambiguation and the discourse structure, but they did not discuss this aspect any further.

Because concept-level cohesive ties and the concepts mutually depend on each other and are highly dependent on the specific constellation in a text, it is extremely challenging to approximate this ideal context definition. Methods that simultaneously disambiguate mentions and build lexical chains approach this ideal context definition. However, such methods exclusively maximize relatedness which can lead to problems as also candidate concepts that are not denoted by the mentions can be highly related (Section 2.2).

**Observation 2.11** *Due to the mutual relationship between denoted concept and context behavior, a mention should be simultaneously assigned to a bin and accordingly be disambiguated.*

As discussed in this chapter, cohesion is probably the most prevalent factor that determines the appropriate context to disambiguate a mention. In the next section, we thus propose a binwise approach based on cohesive scopes. This proposed approach is a discourse-aware approach, as it approximates the contextual behavior of mentions in discourse.

### 2.3.2 Context Definition Based on Cohesive Scopes

To identify the exact concept-level cohesive ties of a mention is extremely challenging. A less difficult task is to determine within which broader text parts a mention exhibits such cohesive

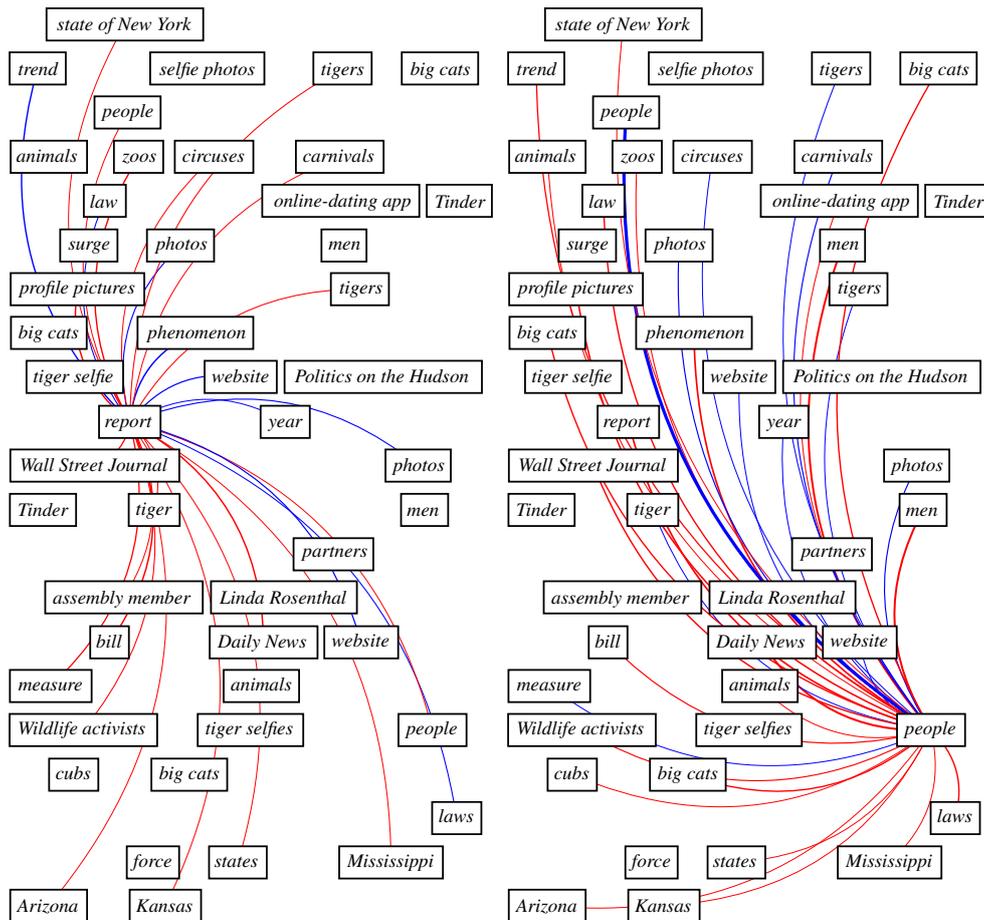


Figure 2.10: Two mentions with local cohesive scope. Blue lines indicate that the highest relatedness score between two mentions involve the concepts they actually denote; red lines indicate that the highest relatedness score obtained for two mentions involves candidate concepts that are not denoted in this context. We assume that the concepts of the context mentions are known.

ties. The idea of our bin-wise approach based on cohesive scopes is to approximate the exact concept-level cohesive ties by such broader text parts. We assume that each mention has a cohesive scope.

**Definition 2.15** *The cohesive scope of a mention is the text span within which a concept denoted by a mention shows cohesive ties.*

In the following, we use *cohesive scope* in the sense of *concept-level cohesive scope* unless it is indicated differently. Given the notion of cohesive scope, we can define different categories of cohesive scopes. These categories serve as our bins and each of them is associated with a different context definition. Each mention is classified into one of these bins and the context definition is activated accordingly. As the disambiguation of a mention depends on its

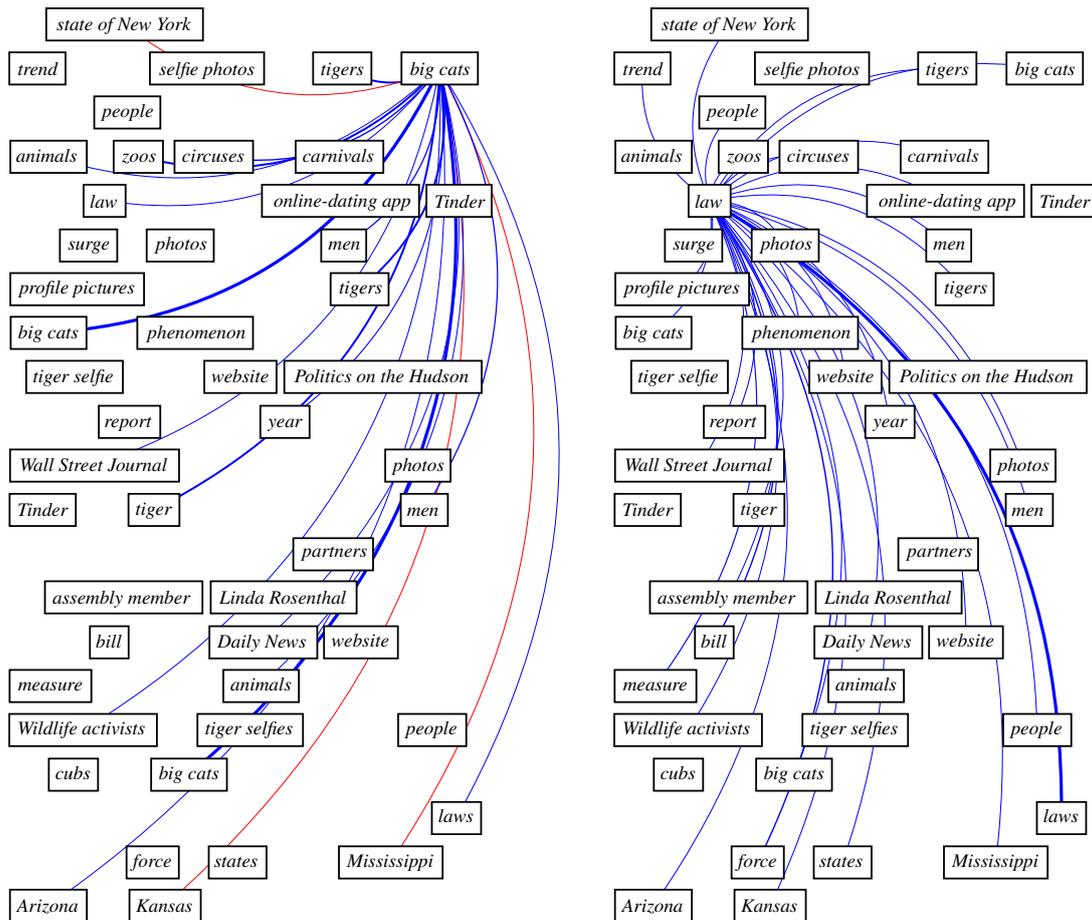


Figure 2.11: Two mentions with intermediate cohesive scope. Blue lines indicate that the highest relatedness score between two mentions involves the concepts they actually denote; red lines indicate that the highest relatedness score obtained for two mentions involves candidate concepts that are not denoted in this context. We assume that the concepts of the context mentions are known.

cohesive ties and vice versa, mentions are assigned to bins and disambiguated simultaneously. The notion of scope is consequently a means to define the appropriate context to disambiguate a mention.

We distinguish three broad categories of cohesive scopes, i.e. local, intermediate and global cohesive scopes. Figure 2.10, 2.11 and 2.12 show the cohesive ties for some examples for each scope: in Figure 2.10, some mentions with local scope are shown; in Figure 2.11, some mentions with intermediate scope are shown and in Figure 2.12, a mention with global scope is shown. As in former pictures, mention pairs are connected by blue lines in case the highest concept-level relatedness score involves the concepts that are denoted and by red lines otherwise.

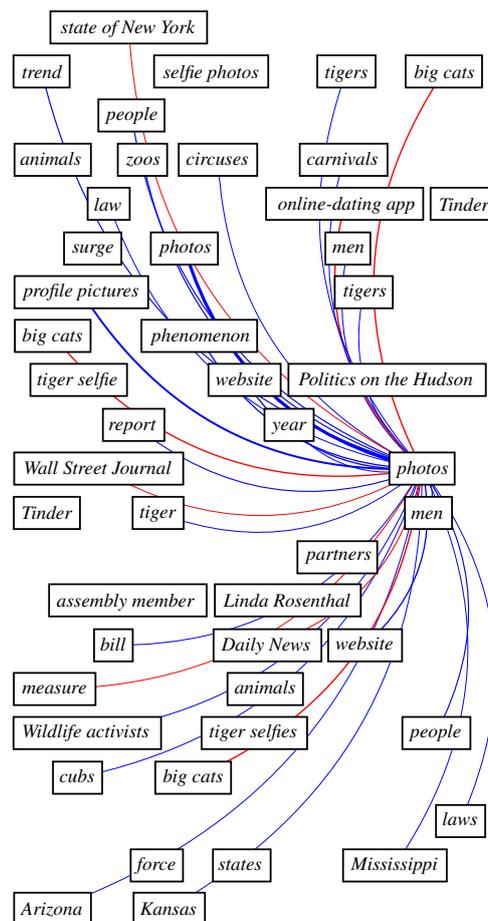


Figure 2.12: One mention with global cohesive scope. Blue lines indicate that the highest relatedness score between two mentions involve the concepts they actually denote; red lines indicate that the highest relatedness score obtained for two mentions involves candidate concepts that are not denoted in this context. We assume that the concepts of the context mentions are known.

**Definition 2.16** *Mentions with local cohesive scope exhibit cohesive ties with lexical units in the same sentence.*

A mention of local scope does not exhibit relations with lexical units outside its sentence. Hence, the global context does not help to disambiguate it or can even lead to the wrong disambiguation. In Figure 2.10, the cohesive ties for two mentions with local scope are shown: *report* and *people*. As indicated by the red lines, the global context is misleading in these examples.

**Definition 2.17** *Mentions with intermediate cohesive scope show cohesive ties both within the sentence and beyond.*

For a mention with intermediate scope both the local and global context are relevant. In Figure 2.11, *big cats* and *law* are of intermediate cohesive scope, as both the local and the global context are indicative.

**Definition 2.18** *Mentions with global cohesive scope form cohesive ties with mentions across sentence boundaries.*

For a mention with global scope, the global context is crucial, while the local context is not discriminative or even misleading. For instance, *photos* in Figure 2.12 is of global scope, as the local context is misleading. If we considered the local context, it might be disambiguated as PHOTO (FRENCH MAGAZINE) due to the high relatedness between PHOTO (FRENCH MAGAZINE) and WALL STREET JOURNAL. The global context is more indicative in this case.

Whether a mention has a local, intermediate or global scope is determined by its concept-level cohesive ties which are difficult to identify. However, we assume that other features can indicate the cohesive scope of a mention. We argue that features that approximate the salience of a mention – the less salient a mention is the more likely it has local cohesive scope – help to determine a mention’s scope (Chapter 5).

To obtain an idea how much scopes influence our running example, we applied the same method as described in Section 2.2, but this time via considering cohesive scopes. For each mention, we manually decided if it has a local, global or intermediate scope and only consider the mentions within this scope for its disambiguation. Instead of 73.5% (result without scope), 81.6% of the mentions are disambiguated correctly. However, these results are based on a small sample size with manually identified scopes and predetermined concepts of the context mentions. Nevertheless, it shows that a context definition based on cohesive scopes may improve the results and is worth to be studied.

## 2.4 Implications

In this chapter, we have analyzed concept disambiguation and clustering from a semiotic and discourse perspective. This analysis has led to some linguistic insights that affect the modeling of these tasks. In this section, we recapture these insights and summarize the requirements for a concept disambiguation and clustering approach from a linguistic perspective. In Chapter 3 and 4 we approximate these requirements in terms of available resources and state-of-the-art machine learning techniques.

### 2.4.1 Implications for Selecting an Inventory

The way how we define a concept affects the requirements for the inventory. Our concept definition comprises three main parts: (1) a concept is a cultural unit; (2) a concept consists of the denotation and shared connotations; (3) we assume that both proper names and common nouns denote concepts. These aspects – including the cross- and multilingual facet – must be reflected in the selected inventory.

**A Collaboratively Built Inventory.** Defining a concept as a culturally shared unit determines whose view should be reflected in the inventory. According to our definition, the inventory should mirror the concepts that are commonly shared in a language community. In this sense, an inventory should be built in a *collaborative* and *consensus-based* way. This questions if an inventory created by a few lexicographers reflects the view of a language community appropriately. We argue that the more people are collaboratively involved in the creation of an inventory the better it reflects what is actually shared. It is important to notice that defining a concept as a culturally shared unit implies that the denotation and connotation of a specific concept are shared. In addition, it also implies that the concept boundaries are shared. This affects the granularity of an inventory.

**An Inventory Accounting for Denotative and Connotative Aspects.** Although we mainly consider the denotation of a concept, an inventory that also approximates connotative aspects is preferable for automatic text processing. As we have shown in Chapter 2.1.2, concept-level cohesive ties – which are crucial for disambiguation – are often based on connotations. Hence, if an inventory accounts for connotative aspects they can be exploited for concept disambiguation.

**An Inventory for Common Nouns and Proper Names.** In order to explain cohesive ties between common nouns and proper names, a common representation for them is required. We therefore argue that an inventory should contain concepts denoted by both common nouns and proper names. Traditional inventories (e.g. dictionaries) often lack proper names.

**Conclusions.** Hence, the ideal inventory is a multilingual resource with concepts denoted by both common nouns and proper names. It is built by a larger community and does not only contain denotations, but also accounts for connotative potentials.

### 2.4.2 Implications for Modeling

The core insight gained from our semiotic- and discourse-oriented analysis is that the concept denoted by a mention depends on the concept-level cohesive ties of a mention and vice versa.

This interrelation between concepts and cohesive ties, or more generally between concepts and discourse structure, is the crux for concept disambiguation and explains at least partially why concept disambiguation is a difficult task. From this interrelation we can derive three implications for modeling concept disambiguation and clustering.

**Joint Modeling of Interrelated Mentions.** Given that two mentions share a concept-level cohesive tie, the concept denoted by one of the mentions depends on the concept denoted by the other mention and vice versa. For instance, if the mentions *tiger* and *big cat* in a text are connected by a concept-level cohesive tie and *tiger* denotes the animal tiger (TIGER (ANIMAL)), *big cat* most likely denotes the concept BIG CAT. In contrast, if *tiger* denotes TIGER ALI SINGH, a wrestler, *big cat* may denote CURTIS HUGHES, another wrestler. In terms of modeling, this implies that mentions that share a concept-level cohesive tie should be simultaneously disambiguated.

**Joint Disambiguation and Clustering of Concepts.** In concept disambiguation, the aim is to identify the concept a mention denotes, while concept clustering considers concept-level identity relations. If two mentions denote the same concept (disambiguation perspective), they are tied by a concept-level identity relation (clustering perspective) and vice versa. For instance, if *law* and *laws* are in a concept-level relation of type identity and *law* denotes a man-made law, *laws* must denote the same concept. In contrast, if *laws* denotes a physical law, *law* either has to denote the same concept or the identity relation between the two mentions is invalid. Because of this interdependency between disambiguation and clustering decisions, the two tasks can mutually benefit from each other and should be modeled jointly.

**Joint Context Selection and Concept Disambiguation.** The ideal context to disambiguate a mention is given by the syntactic context and the lexical units with which it shares concept-level cohesive ties. However, if two mentions share a concept-level cohesive tie depends on their denoted concepts. If for example *tiger* and *big cat* share a cohesive tie depends on their denoted concepts. If *tiger* denotes the animal (TIGER (ANIMAL)) and *big cat* the concept CURTIS HUGHES, they are not tied by a concept-level cohesive tie. In contrast, if *tiger* and *big cat* both denote animals, they share a concept-level cohesive tie. Hence, context selection and disambiguation are dependent on each other and should be modeled jointly.

**Conclusions.** Given these observations, we conclude that the ideal approach for concept disambiguation and clustering (1) models both tasks jointly, (2) models all mentions that are tied by concept-level cohesive ties jointly and (3) models the selection of the context and the concept disambiguation decision jointly.

Due to time efficiency reasons, it is not possible to model all these interrelations accordingly (Chapter 4). However, we can at least approximate them. To approximate context selection, we have already proposed a binwise approach (Section 2.3.1). In Chapter 4, we discuss how we actually approach concept disambiguation and clustering using state-of-the-art machine learning methods.



## Chapter 3

# Wikipedia as an Inventory

Concept disambiguation requires an inventory consisting of predefined concepts and their lexicalizations. From our concept definition, we derived that the inventory ideally meets four criteria (Section 2.4.1): First, it should reflect culturally shared concepts. An inventory that is collaboratively built by a larger community instead of a few experts is more likely to describe such shared concepts. Second, it should describe the denotation of a concept, but also account for its connotative potential, as this may help for disambiguation (Section 2.2). Third, as we aim to disambiguate both common nouns and proper names, it should cover both. Forth, it should be multilingual.

Based on these requirements, we justify our decision to use Wikipedia as an inventory. We discuss its adequacy and compare it to other available resources (Section 3.1). As Wikipedia has been designed for people and not as a concept inventory for a concept disambiguation algorithm, we need to adapt it accordingly and extract the relevant information. Inspired by *WikiNet* (Nastase & Strube, 2013), we derive a multilingual inventory from Wikipedia (Section 3.2). Some statistics for our derived inventory are presented in Section 3.3.

### 3.1 Selecting an Inventory

Inventories – or resources from which an inventory can be derived – have been built from different perspectives and for diverse purposes and applications. In this thesis, we aim to select an inventory that meets the requirements of multilingual concept disambiguation. We do not target a specialized domain such as the biomedical domain, but focus on the general language usage found e.g. in newspaper texts and web pages. Thus, the inventory must be a general purpose resource with high coverage and not a domain-specific repository such as the UMLS Metathesaurus for the biomedical and health care domains.<sup>1</sup>

---

<sup>1</sup><http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html>, 6.3.2014.

Candidate inventories range from dictionaries, thesauri, ontologies to encyclopedias. A few of these resources have been built for computer-based applications; however, most are primarily designed for humans, but have been exploited for natural language processing applications in the past. In the following, we analyze the adequacy of Wikipedia as an inventory (Section 3.1.1) and compare it to the most prominent inventories that have been used previously for concept disambiguation (Section 3.1.2).

### 3.1.1 Wikipedia

With 30 million articles in 287 languages, Wikipedia is the biggest multilingual encyclopedia and is continuously growing since it has been launched in 2001.<sup>2</sup> In 2005, the English Wikipedia, which is the biggest language version, consisted of about 0.4 million, in 2008 of 2.1 million and in January 2014 of 4.5 million articles.<sup>3</sup> The slogan of Wikipedia, “the free encyclopedia that everyone can edit”<sup>4</sup>, reflects its collaborative nature. In January 2014 more than 76,000 editors contributed to Wikipedia, including e.g. about 30,000 to the English, 6,800 to the German, 2,700 to the Italian, 2,200 to the Chinese, and more than 250 to the Bulgarian version.<sup>5</sup> Its collaborative design and the tremendous size makes Wikipedia an interesting resource for natural language processing (Gabrilovich & Markovitch, 2007b; 2007a; Finin et al., 2009; Ponzetto & Strube, 2011; Ratinov & Roth, 2012). In order to decide if it is suitable as a concept inventory, we need to clarify (1) what qualifies as an article, (2) what information is covered by an article (denotation, connotation) and (3) how linguistic realizations and articles are connected.

**What Qualifies as an Article.** According to the guidelines of the English Wikipedia, articles are about “a person, or a people, a concept, a place, an event, a thing etc. that their title can denote.”<sup>6</sup> This description reflects the concept-centric perspective of Wikipedia. Each article corresponds to one concept, i.e. “articles rarely, if ever, contain more than one distinct definition or usage of the article’s title.”<sup>7</sup> If the same title can denote different concepts, separate articles for the different concepts are used.<sup>8</sup> All synonyms and variations that denote

<sup>2</sup><http://en.wikipedia.org/wiki/Wikipedia>, 14.3.2014.

<sup>3</sup><http://stats.wikimedia.org/EN/TablesArticlesTotal.htm>, 14.3.2014. This number include all articles (redirects, category and disambiguation pages are excluded) that contain at least one internal link.

<sup>4</sup><http://en.wikipedia.org/wiki/Wikipedia>, 14.3.2014.

<sup>5</sup><http://stats.wikimedia.org/EN/TablesWikipediansEditsGt5.htm>, 14.3.2014. This number includes all Wikipedians who contributed at least 5 times in January 2014.

<sup>6</sup>[http://en.wikipedia.org/wiki/Wikipedia:Wikipedia\\_is\\_not\\_a\\_dictionary](http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_dictionary), 17.3.2014.

<sup>7</sup>[http://en.wikipedia.org/wiki/Wikipedia:What\\_Wikipedia\\_is\\_not](http://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not), 17.3.2014.

<sup>8</sup>[http://en.wikipedia.org/wiki/Wikipedia:Wikipedia\\_is\\_not\\_a\\_dictionary](http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_dictionary), 17.3.2014.

the same concept point to the same article.<sup>9</sup> The guidelines of Wikipedia<sup>10</sup> further state that only if a topic is *notable*, i.e. “has received significant coverage in reliable sources that are independent of the subject, it is presumed to be suitable for a stand-alone article or list.”<sup>11</sup> As reliable sources count for example peer reviewed publications, authoritative books or reputable media sources.<sup>12</sup> The notion of notability helps to ensure that only concepts shared by a larger community are included and that their verification is guaranteed. It is in line with our definition of concepts as culturally shared units. That the concepts are shared is further ensured by resolving disagreement among editors through “consensus-based discussions”.<sup>13</sup>

Wikipedia mainly contains concepts denoted by nouns. Adjectives and verbs often lack a separate article, but point to the article for the corresponding noun.<sup>14</sup> For instance, the adjective *excellent* points to the article EXCELLENCE, *invent* to INVENTION and *run* to RUNNING. Wikipedia often also lacks a separate entry for the agent of an activity: e.g. *runner* and *inventor* are linked to RUNNING and INVENTION respectively.<sup>15</sup>

Hence, with respect to what qualifies as a concept, Wikipedia is an appropriate resource for concept disambiguation. The guidelines that describe what an appropriate article topic constitutes are in line with our concept definition and also highlight that articles should be about shared topics. However, the guideline and also some examples imply that Wikipedia is more suitable to disambiguate nouns than verbs or adjectives.

**Information Covered by an Article.** In contrast to definitions in dictionaries, concept descriptions in encyclopedias such as Wikipedia are much more extensive and have a different purpose: the information found in a dictionary is often considered as reflecting linguistic knowledge, the information in an encyclopedia as reflecting factual knowledge or world knowledge. In the following, we investigate if the information covered by Wikipedia suits our requirements.

Each Wikipedia article usually starts with a concise definition of the respective concept.<sup>16</sup> These definitions include synonyms and transliterations – if applicable –, describe the essential and distinguishing properties and establish the context, e.g. by listing hyperonyms and

<sup>9</sup>[http://en.wikipedia.org/wiki/Wikipedia:Wikipedia\\_is\\_not\\_a\\_dictionary](http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_dictionary), 17.3.2014.

<sup>10</sup>We follow here the guidelines of the English Wikipedia.

<sup>11</sup><http://en.wikipedia.org/wiki/Wikipedia:N>, 14.3.2014.

<sup>12</sup><http://en.wikipedia.org/wiki/Wikipedia:N>, 14.3.2014.

<sup>13</sup>[http://en.wikipedia.org/wiki/Wikipedia:What\\_Wikipedia\\_is\\_not](http://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not), 14.3.2014., 17.3.2014.

<sup>14</sup>This is also postulated by the guidelines: [http://en.wikipedia.org/wiki/Wikipedia:Article\\_titles](http://en.wikipedia.org/wiki/Wikipedia:Article_titles) and <http://en.wikipedia.org/wiki/Wikipedia:Redirect>, 17.3.2014.

<sup>15</sup>Examples are from the 17.3.2014.

<sup>16</sup>[http://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style/Lead\\_section](http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section), 14.3.2014.

hyponyms.<sup>17</sup> The style of these definitions is uniform and usually adapts the recommendation of the style guidelines (e.g. it should be understandable by non-specialists and the title should be the subject).<sup>18</sup> Hence, these concept definitions are comparable to the descriptions in a dictionary or in WordNet and approximate the denotation. Apart from the definition, a Wikipedia article contains more details about what is known about the concept, its cultural significance and how it relates to other concepts. Depending on the concept, a Wikipedia article can consist of a few paragraphs to several pages of text. Phrases in article texts that are relevant to the current article or that need more explanations are usually hyperlinked to the corresponding Wikipedia articles.<sup>19</sup> Through these *internal hyperlinks* the articles are interlinked and build a network: a Wikipedia article is consequently not an isolated object, but is embedded into a network of related concepts. In addition, Wikipedia articles are typically linked to categories that are hierarchically organized and topically group articles.<sup>20</sup> The category hierarchy is not a strict taxonomy, but a *folksonomy* (Ponzetto & Strube, 2011; Nastase & Strube, 2013) reflecting the collaboratively created categorization. Furthermore, Wikipedia articles often contain so called infoboxes that summarize key features of the described concept.<sup>21</sup>

A Wikipedia article thus comprises a wealth of information accounting for the denotation (definition at the beginning) and connotative aspects (e.g. hyperlinks to related articles, categorial embedding). Compared to the short glosses in WordNet, a Wikipedia article is more extensive and the concept relations are much more dense (Navigli & Ponzetto, 2012a), although not labeled as in WordNet.

In contrast to a dictionary or WordNet that often contain example sentences to illustrate the usage of a word or a concept, Wikipedia articles lack such explicit example sentences. However, example sentences can be retrieved by extracting sentences with internal hyperlinks from article texts: the linguistic expression that is hyperlinked is the *anchor*, the article to which the hyperlink points is the *target* (Mihalcea & Csomai, 2007). According to the guidelines, a link should always point to the most specific topic<sup>22</sup>, which increases chances that the target is actually the denoted concept and not only an associated concept. Figure 3.1 shows an example for internal hyperlinks.

<sup>17</sup>[http://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style/Lead\\_section](http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section), 14.3.2014 and [http://en.wikipedia.org/wiki/Wikipedia:Wikipedia\\_is\\_not\\_a\\_dictionary](http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_dictionary), 17.3.2014.

<sup>18</sup>[http://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style/Lead\\_section](http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section), 14.3.2014.

<sup>19</sup><http://en.wikipedia.org/wiki/Wikipedia:Linking>, 18.3.2014.

<sup>20</sup><http://en.wikipedia.org/wiki/Wikipedia:Categorization>, 18.3.2014

<sup>21</sup>[http://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style\\_\(infoboxes\)](http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style_(infoboxes)), 18.3.2014.

<sup>22</sup>[http://en.wikipedia.org/wiki/Wikipedia:Linking#Units\\_of\\_measurement\\_that\\_aren.27t\\_obscure](http://en.wikipedia.org/wiki/Wikipedia:Linking#Units_of_measurement_that_aren.27t_obscure), 18.3.2014.

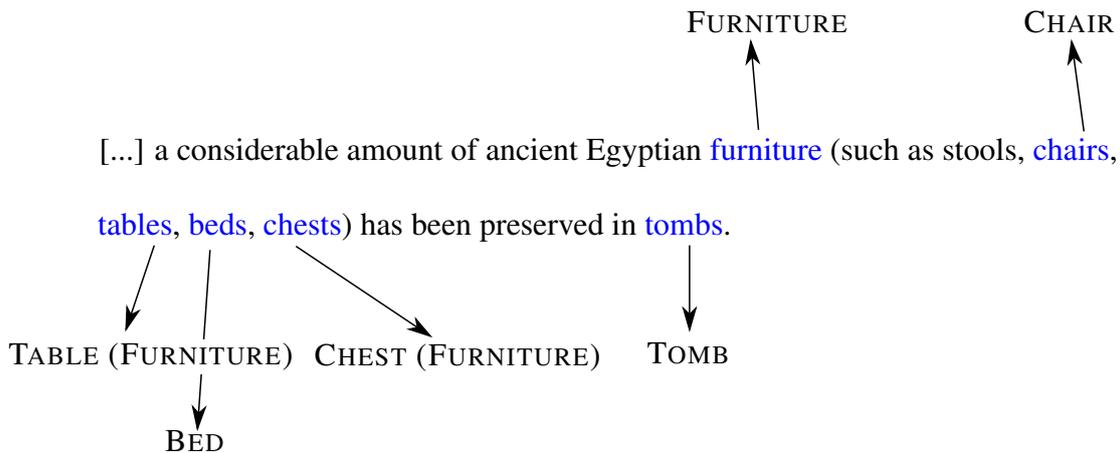


Figure 3.1: Sentence with an internal hyperlinks. The sentence is extracted from the Wikipedia article WOODWORKING, 19.9.2014. The anchors are highlighted in blue. The Wikipedia articles they point to are indicated by arrows.

**Relationship between Linguistic Realizations and Articles.** The aim of concept disambiguation is to link mentions in a text to concepts in an inventory. Hence, a concept inventory should not only contain information about the concepts, but also provide different linguistic realizations for the concepts.

Linguistic realizations can be harvested from different sources in Wikipedia. Each Wikipedia article has a title. This title can consist of a single word, usually a noun, or a whole phrase.<sup>23</sup> The name of an article should be a common linguistic realization and is determined consensus based.<sup>24</sup> If the title is ambiguous, a descriptor (e.g. *furniture* in *table (furniture)*) that allows to distinguish between different articles is added. Alternative linguistic realizations (synonyms) and morphological variants (e.g. plural forms) are linked to the corresponding article by *redirection pages*. These pages do not bear any information except a linguistic realization and have the only purpose to point or redirect to the corresponding article. *Disambiguation pages* are another source for linguistic realizations for articles. A disambiguation page lists candidate articles for a given ambiguous linguistic realization. Eventually, internal hyperlinks can also be exploited to obtain additional linguistic realizations of concepts (Figure 3.1). Linguistic realizations in other languages can be obtained by leveraging *cross-language links*. Cross-language links allow editors to point to Wikipedia articles in other languages about the same concepts.

**Example.** For illustration, we briefly describe the information which is available for *table* in Wikipedia. We will use this example also for other resources in the next section. Looking up *table* in Wikipedia leads us to the disambiguation page for *table* depicted in Figure 3.2. As

<sup>23</sup>[http://en.wikipedia.org/wiki/Wikipedia:Article\\_titles](http://en.wikipedia.org/wiki/Wikipedia:Article_titles), 17.3.2014.

<sup>24</sup>[http://en.wikipedia.org/wiki/Wikipedia:Article\\_titles](http://en.wikipedia.org/wiki/Wikipedia:Article_titles), 18.3.2014.

**Table** may refer to:

- [Table \(database\)](#)
- [Table \(furniture\)](#)
- [Table \(information\)](#), a data arrangement with rows and columns
- [Table \(landform\)](#)
- [Table \(parliamentary procedure\)](#)
- [Tables \(board game\)](#)
- [Calligra Tables](#), a spreadsheet application
- [The Table](#), a volcanic tuya in British Columbia, Canada

Figure 3.2: Disambiguation page for *table*, 14.3.2013.

we will see in next section, Wikipedia concepts are more coarse-grained than WordNet senses (Figure 3.3).

The article on TABLE (FURNITURE) is more than seven pages long. The first few sentences define TABLE (FURNITURE) by describing its characteristics (“flat horizontal upper surface”), its purpose (“to support objects of interest, for storage, show, and/or manipulation”) and by listing its hyperonym (“furniture”) and hyponyms (e.g. “dining room table”, “bedside table”).<sup>25</sup> The rest of the article elaborates on what is known about this concept and its cultural status. The concept is further associated with the two categories FURNITURE and TABLES (FURNITURE), which groups this concept with other similar concepts. Through cross-language links, the article is connected to concept descriptions in other languages.

**Advantages and Disadvantages.** Although it has been designed for humans and not for machines, Wikipedia, the biggest multilingual collaborative encyclopedia, is an attractive concept inventory. Beside its size, Wikipedia offers several other advantages compared to other resources.

First, Wikipedia is built in a *collaborative* and *consensus-based* way, not only by a few experts, but by thousands of people. It is therefore likely to reflect culturally shared units. Articles in Wikipedia that are not distinguishable tend to be merged leading to a less fine-grained inventory than e.g. WordNet (McCarthy, 2006a). As a consequence, the inter-annotator agreement for concept annotations with Wikipedia as an inventory is higher than with WordNet (see Table 8.4, Section 8.5). Second, Wikipedia articles contain a wealth of information including a description of the *denotation* and, by e.g. linking to related articles and associated categories, *connotative aspects*. Wikipedia articles are much more interlinked than WordNet synsets (Navigli & Ponzetto, 2012a). The internal hyperlink structure (Milne & Witten,

<sup>25</sup>[http://en.wikipedia.org/wiki/Table\\_\(furniture\)](http://en.wikipedia.org/wiki/Table_(furniture)), 14.3.2014

2008a), the article texts (Gabrilovich & Markovitch, 2007a) and the category hierarchy (Strube & Ponzetto, 2006; Ponzetto & Strube, 2007; Nastase & Strube, 2013) have been successfully exploited for calculating relatedness between articles, which is crucial for disambiguation. Third, Wikipedia covers both *common nouns* and *proper names*. Fourth, Wikipedia is *multi-lingual*: not only linguistic realizations of a concept are present in many languages, but whole articles are available in many languages interlinked by cross-language links. Each Wikipedia language version is an independent project inter-linked by the cross-language links. Hence, Wikipedia articles in different languages are usually written independently from other language version and account for cultural differences. In contrast, wordnets in other languages than English are often derived from the English WordNet (e.g. MultiWordNet) by e.g. translating the synset members. This strategy may lead to biases towards English. Fifth, Wikipedia is constantly updated. While in 2012, SELFIE was not part of it, it has been included in the meantime. In 2014, the article text of SELFIE is more than 5 pages long. Hence, Wikipedia accounts for the dynamic nature of language and culture. In contrast, WordNet is more static with few releases. It takes much longer until it accounts for a change or a new concept.

Wikipedia also has some shortcomings. First, it is suitable as an inventory for concepts denoted by nouns, but less for adjectives and verbs. As we focus on nouns in this thesis, this is not an issue, but when also other part-of-speech tags should be considered, one either has to combine Wikipedia with another resource such as WordNet (Navigli & Ponzetto, 2012a) or one links verbs and adjectives to the concepts denoted by the corresponding nouns (e.g. *excellent* to EXCELLENCE). Second, Wikipedia contains some blurs. For instance, the agent of an action is often not separated from the action (e.g. *runner* and *running* point to the same article). Finally, the dynamic nature of Wikipedia imposes difficulties with respect to benchmarking systems over the years. Concepts and linguistic realizations can appear, change or disappear. A test set that is manually annotated with respect to a specific Wikipedia version might be outdated given a later Wikipedia version. At the same time, the difficulty of the task can vary dependent on the used Wikipedia version. In an older version for instance, the average ambiguity tends to be lower than in a more recent version. These effects can be minimized by using the same version for all experiments.

### 3.1.2 Comparison with Other Resources

Wikipedia is only one out of many resources that can serve as an inventory. In the following, we discuss the most prominent resources that have been used as a concept inventory in the past and compare them to Wikipedia. Table 3.1 and Table 3.2 summarize the most important aspects of this comparison and also include a few more resources.

**Wiktionary.** Wiktionary<sup>26</sup> is a multilingual machine readable dictionary. As all machine readable dictionaries – such as *Webster’s 7th Collegiate*, the *Collins English Dictionary* or the *Oxford Advanced Learner’s Dictionary of Current English* –, it takes a lexeme-centric perspective: for each lexeme, the concepts it can denote are paraphrased by a short description or some synonyms.

Wiktionary has been collectively built. It follows the spirit of Wikipedia and allows everybody to contribute to its content. In the meantime, it is available in 122 language versions.<sup>27</sup> One language version is not restricted to words of one single language. On the contrary, the aim is to “describe all the words of all languages”<sup>28</sup> in each language. An entry contains information associated with a lemma in a specific part of speech, including hierarchically organized sense descriptions, synonyms, hypernyms, related terms, categories, etymological information and translations. For instance, the English entry for *table* describes the meanings that the word *table* has both in the English and the French language. The English Wiktionary is with 3,683,428 entries from 1,100 languages<sup>29</sup> the biggest language version. The different language versions are connected by inter-language links: the English entry for *table* points for example to the entry *table* in the German version, but not to the German equivalent *Tisch* or *Tabelle*.<sup>30</sup>

While machine readable dictionaries such as LDOCE have been used in early research on disambiguation, Wiktionary has been considered as an inventory only recently (Meyer & Gurevych, 2010; 2012; Miller et al., 2013). Although Wiktionary is appealing due to its collaborative design (Meyer & Gurevych, 2010), its size, its multilinguality and the wealth of information it contains, it comes with two disadvantages that are shared by all lexeme-centric resources. First, information that is actually related to concepts such as hyponym relations are associated with lexemes instead. For instance, in Wiktionary, *table* is considered as a hyponym of *furniture*. However, this hyponym relation only holds between the concepts TABLE (FURNITURE) and FURNITURE and for example not between the concepts TABLE (DATA GRID) and FURNITURE. Hence, although dictionaries often contain information such as hyponymy relations that might be valuable for concept disambiguation, it is difficult to exploit them properly. Second, lexeme-centric resources mainly account for ambiguity of words and less for variability. Usually, no common concept identifier is shared across entries and it is difficult to merge descriptions across entries that paraphrase the same concept. Meyer & Gurevych (2012) propose a method to automatically disambiguate semantic relations

<sup>26</sup><http://www.wiktionary.org/>, 7.3.2014

<sup>27</sup>This statistics is from the 25.2.2014 and only include languages with more than 10 articles and which received at least 10 edits within the previous month. <http://stats.wikimedia.org/wiktionary/EN/Sitemap.htm>, 7.3.2014.

<sup>28</sup>[http://en.wiktionary.org/wiki/Wiktionary:Main\\_Page](http://en.wiktionary.org/wiki/Wiktionary:Main_Page), 7.3.2014.

<sup>29</sup>[http://en.wiktionary.org/wiki/Wiktionary:Main\\_Page](http://en.wiktionary.org/wiki/Wiktionary:Main_Page), 7.3.2014

<sup>30</sup>The example reflects English and German versions from the 7.3.2014.

(such as synonymy and hyponymy) and translations in Wiktionary based on word overlap and hyperlink-based features. Their derived inventory contains more than 474,000 concepts and around 379,600 lexicalizations for English and 73,500 concepts and 85,500 lexicalizations for German. Although this might be a valuable resource, it relies on automatic disambiguation of semantic relations which lowers its quality.

Hence, we argue that machine readable dictionaries such as Wiktionary contain valuable information for concept disambiguation, but suffer from disadvantages due to their lexeme-centric perspective. In order to fully exploit such machine-readable dictionaries for concept disambiguation, an automatic disambiguation of semantic relations is required which can hamper the quality of the derived inventory.

**Roget's Thesaurus.** A thesaurus groups words that denote related concepts “without attempting to explicitly name each relationship” (Morris & Hirst, 1991, p. 28). These groups can contain words that denote similar or even the same concepts, but also words that denote related or associated concepts. We follow Roget's (1852) terminology and say that such a group of words represents an *idea*. These ideas are often organized in a taxonomy. A thesaurus can be considered as an inverse dictionary: “A dictionary explains the meaning of words, whereas a thesaurus aids in finding the words that best express an idea or meaning” (Morris & Hirst, 1991, p. 27). Consequently, a dictionary is lexeme-centric; a thesaurus is idea-centric.

Roget's Thesaurus is a general purpose thesaurus with the aim of “furnishing on every topic a copious store of words and phrases, adapted to express all the recognizable shades and modifications of the general idea under which those words and phrases are arranged” (Roget, 1964, p. 12). Words that denote similar or related concepts are clustered into groups and subgroups. Each group is labeled by a so called head and contains nouns, adjectives, adverbs and (idiomatic) phrases, but no proper names. As it is common for a thesaurus (Morris & Hirst, 1991, p. 28), the relations between the denoted concepts within a group are implicit. Roget's Thesaurus appeared for the first time in 1852. The 1987 version of Roget's Thesaurus contains 990 heads embedded in a taxonomic structure (Kennedy & Szpakowicz, 2008).<sup>31</sup> The resource is built by lexicographers and is only available for English.

Roget's Thesaurus has been used as an inventory for coarse-grained concept disambiguation (Yarowsky, 1992). Given a word to disambiguate, the assumption is that its coarse-grained denoted concepts can be derived from its corresponding groups in Roget's Thesaurus with the heads serving as labels. For instance, *table* is in the following groups: LIST, ARRANGEMENT, WRITING, SUPPORT, LAYER, FLATNESS and FOOD.<sup>32</sup> As this example illustrates,

<sup>31</sup>The version from 1911 is available from the Gutenberg project. The same version has been adapted for NLP applications and is available from here: <http://rogets.site.uottawa.ca/>. Kennedy & Szpakowicz (2008) compare the different versions of Roget's Thesaurus.

<sup>32</sup>Roget's Thesaurus (1911): <http://www.gutenberg.org/files/10681/10681-h-index-pos.htm>, 11.3.2014.

the partitions for a word given by the groups in Roget's Thesaurus roughly correspond to its (coarse-grained) concepts. However, there is no one-to-one relationship between these partitions and coarse-grained concepts. For instance, both the LIST and ARRANGEMENT group account for the grid sense of *table*.

Thesauri are a valuable resource for (coarse-grained) disambiguation (Yarowsky, 1992). The implicit relations between the members of a group encode both denotative and connotative aspects and go beyond synonymy and hyponymy that are sometimes present in a dictionary. Although the relations are unlabeled, they can be used to detect cohesive ties in texts (Morris & Hirst, 1991; Jarmasz & Szpakowicz, 2003) and to calculate similarities between concepts (Kennedy & Szpakowicz, 2008) – both tasks that are relevant for concept disambiguation (Section 2.2). The downsides of thesauri such as Roget's Thesaurus with respect to concept disambiguation are three-fold. First, although the partitions for a word based on the groups in Roget's Thesaurus roughly corresponds to coarse-grained concepts, it is not guaranteed that one concept corresponds to one group. The purpose of a thesaurus is to group topically related words and not to distinguish between different concepts a word can denote. Second, relations within groups are implicit and synonyms are thus not marked as such. Thus, it does not account for variability. Third, compared to dictionaries, thesauri are further detached from the language usage. Thesauri lack any examples that illustrate the usage of a word in context. Such information is valuable for a disambiguation system.

We conclude that thesauri contain topical relations that could be exploited for concept disambiguation. But as the basic units of thesauri such as Roget's Thesaurus are ideas from which concepts can only be derived with a remaining uncertainty, they are not optimal as inventories for concept disambiguation.

**WordNet.** WordNet is the inventory that has been used most frequently for concept disambiguation or – how the task is usually called in this context – word sense disambiguation. Starting in the early 2000's, several shared tasks have been organized for word sense disambiguation at SensEval and SemEval with WordNet as an inventory. WordNet's design is inspired by psycholinguistic theories of the human lexical memory and driven by the idea to search dictionaries conceptually rather than alphabetically (Miller et al., 1990). WordNet covers nouns, verbs, adjectives and adverbs and consists of more than 117,000 so called *synsets* (Fellbaum, 2012) that correspond to what we call concepts. Each synset is associated with lexicalizations, while “the criterion for joint synset membership is merely that the words denote the same concept” (Fellbaum, 2012). In addition, most synsets contain a short definition (a *gloss*) and often some example sentences for its members (Fellbaum, 2012). The synsets are connected by semantic and derivational relations (Miller et al., 1990). Noun synsets are for instance linked by hyponym and meronym relations. Hence, WordNet is concept-centric.

WordNet has been manually constructed by relatively few linguists and lexicographers

- Noun**
- **S: (n) table, tabular array** (a set of data arranged in rows and columns) "see table 1"
  - **S: (n) table** (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs) "it was a sturdy table"
  - **S: (n) table** (a piece of furniture with tableware for a meal laid out on it) "I reserved a table at my favorite restaurant"
    - **direct hyponym / full hyponym**
      - **S: (n) dining table, board** (a table at which meals are served) "he helped her clear the dining table"; "a feast was spread upon the board"
    - **direct hypernym / inherited hypernym / sister term**
  - **S: (n) mesa, table** (flat tableland with steep edges) "the tribe was relatively safe on the mesa but they had to descend into the valley for water"
  - **S: (n) table** (a company of people assembled at a table for a meal or game) "he entertained the whole table with his witty remarks"
  - **S: (n) board, table** (food or meals in general) "she sets a fine table"; "room and board"

Figure 3.3: Noun synsets for *table* in WordNet 3.1. Semantic relations to other synsets are only shown for one synset.

(Fellbaum, 2012) and is available for English. In the meantime, similar *wordnets* in almost 70 other languages have been constructed<sup>33</sup> and partially linked via an inter-lingual index: e.g. *EuroWordNet* (Vossen, 1998) covers some European languages (German, Dutch, Italian, French, Czech and Estonian); *BalkaNet* (Tufiş et al., 2004) covers Balkan languages (Bulgarian, Czech, Greek, Romanian, Turkish and Serbian); *MultiWordNet* focuses on Italian (Pianta et al., 2002); *AsianWordNet* (Robkop et al., 2010) comprises Asian languages (Thai, Lao, Japanese, Korean, Burmese, Indonesian, Vietnamese, Mongolian, Bengali and Sihala).<sup>34</sup> Although the number of covered languages is relatively high, most of these wordnets are much smaller than the English WordNet.

Figure 3.3 shows the noun synsets for *table* including the semantic relations of one of them. As this example illustrates, WordNet's sense distinctions are fine-grained.

WordNet is suitable as an inventory for concept disambiguation as it is designed from a concept-centric perspective. It allows to resolve ambiguities and to deal with variability. It is frequently used, as it is freely available and data sets annotated with WordNet senses exist (Navigli, 2006).

However, WordNet has been manually built by a few experts. It is thus questionable if it actually reflects shared cultural units and not only the view of a few experts. The sense distinctions in WordNet are partially so fine-grained that both systems and humans have problems to distinguish between them (Mihalcea & Moldovan, 2001a; Palmer et al., 2007; Passonneau et al., 2009). This difficulty is reflected in the low inter-annotator agreement scores for data sets manually annotated with WordNet senses. For instance, the inter-annotator agreement for the annotations used in the English all-words disambiguation task at SenseEval-2 is only

<sup>33</sup><http://globalwordnet.org/wordnets-in-the-world>, 12.3.2014.

<sup>34</sup>A comprehensive list of wordnets can be found here: <http://globalwordnet.org/wordnets-in-the-world/>, 12.3.2014.

72.5% overall and 74.9% for nouns (Snyder & Palmer, 2004). As a consequence, several algorithms have been proposed to automatically merge senses in WordNet and to perform coarse-grained disambiguation, which is sufficient for many NLP applications (Snow et al., 2007; Bhagwani et al., 2013; Navigli, 2006; McCarthy, 2006b; Chlovski & Mihalcea, 2002). A further shortcoming of WordNet is that the information associated with concepts (i.e. synsets) in WordNet is scarce. While the denotation is represented by a short gloss, connotative aspects are sparsely described by a limited number of semantic relations and only few examples are provided to illustrate the usage (Ng et al., 1999). All this information is crucial for concept disambiguation (Section 2.2). Thus, several approaches have been proposed to address this sparsity. For instance, Agirre et al. (2000) enrich synsets with so called topic signatures to overcome the lack of topical links in WordNet and Miller et al. (2012) explore distributional methods to expand word sense description.

Finally, WordNet suffers from coverage problems with respect to proper names (e.g. *Jacques Chirac* is missing)<sup>35</sup>, multiword expressions (e.g. no entry for *housewarming party* exists)<sup>36</sup> and novelisms (e.g. *selfie*). Although wordnets exist in many languages, these language versions are often fairly small. While the Chinese WordNet is with 115,400 synsets<sup>37</sup> almost as big as the English one, GermaNet (German) consists of around 84,000 synsets<sup>38</sup> and the Italian WordNet of around 58,000<sup>39</sup> corresponding to two-third and not even half of the size of the English WordNet respectively. Automatic extensions (de Melo & Weikum, 2009) are of lower quality and still relatively small with 800,000 words from more than 200 languages.

**Conclusions.** In the course of analyzing concept disambiguation from a linguistic perspective, we concluded that the ideal inventory for concept disambiguation is a multilingual, collaboratively built inventory for concepts denoted by common nouns and proper names and accounting for denotative and connotative aspects (Section 2.4.1).

In Table 3.1, the inventories for concept disambiguation that we previously discussed (including some additional ones) are compared with respect to these criteria. From the perspective of concept disambiguation, all inventories that are concept-centric are more appropriate than lexeme- and idea-centric inventories. While all resources catch ambiguities, only concept-centric repositories also account for variability leading to a text representation in which all mentions that denote the same concept are actually linked to the same concept. Given this criterion, WordNet and Wikipedia are preferable as inventories for concept disambiguation. Out of these remaining inventories, Wikipedia suits our requirements best: being collaboratively built, it may better reflect shared concepts; it better accounts for the connota-

<sup>35</sup>WordNet 3.1, <http://wordnetweb.princeton.edu/perl/webwn>, 13.3.2014.

<sup>36</sup>WordNet 3.1, <http://wordnetweb.princeton.edu/perl/webwn>, 13.3.2014.

<sup>37</sup><http://www.aturstudio.com/wordnet/windex.php>, 12.3.2014

<sup>38</sup><http://www.sfs.uni-tuebingen.de/GermaNet/>, 12.3.2014.

<sup>39</sup><http://multiwordnet.fbk.eu/english/licence.php>, 12.3.2014.

tive potential; it covers both common nouns and proper names; its language versions are built independently of each other allowing for cultural differences, while at the same time being interlinked among each other.

The two downsides of using Wikipedia as an inventory with respect to these criteria are its focus on concepts that are denoted by common nouns or proper names and its fuzziness with respect to some concepts (e.g. *runner* and *running* both point to the same concept RUNNING).

Besides our criteria derived from our concept definition, other aspects related to feasibility need to be considered. The final goal is to design a concept disambiguation method with high accuracy and coverage leading to a text representation that is useful for downstream applications. These aspects also impose constraints on selecting an inventory. From this perspective, the size of an inventory both with respect to concepts and linguistic realizations is essential. As the method needs to be developed and evaluated, the availability of sufficient data annotated with concepts is crucial. Finally, the usefulness of the inventory for downstream tasks is also relevant. In Table 3.2, we compare the discussed inventories with respect to these criteria. To evaluate the availability of annotated data, we list for each inventory available data sets and their size. The usability of an inventory for downstream tasks can be assessed by consulting the literature. In addition, we can glimpse the importance of an inventory by checking how often it has been used within the natural language processing community. To approximate this importance, we compare hit counts for the respective inventory in the ACL anthology<sup>40</sup>.

According to the hit counts from the ACL anthology, WordNet is the most frequently used resource. It has been used for tasks such as information retrieval, textual entailment or coreference resolution, although with mixed results. The second most cited resource is Wikipedia. Since Wikipedia has been launched in 2001, it has been widely used in natural language processing as a standalone resource or in combination with e.g. WordNet. While it has been useful for semantic distance calculations (Milne & Witten, 2008b; Ponzetto & Strube, 2011) and for information retrieval (Santamaría et al., 2010), it is ambivalent if Wikipedia can be exploited to improve textual entailment (Bentivogli et al., 2011).

Wikipedia exceeds the other inventories in terms of size and annotated data sets by far. Its English version contains with 4.5 million concepts more than 38 times the number of concepts in WordNet covering much more concepts that are denoted by proper names. The number of lexicalizations in the English Wikipedia given by article titles and redirects is almost 70 times higher than in WordNet including much more multiword expressions and proper names. However, these numbers are not comparable: the Wikipedia counts include morphological variants (e.g. plural forms), which is not the case for the WordNet counts.

---

<sup>40</sup>We used the name of the respective inventory as search term in <http://www.aclweb.org/anthology/>.

Inventory	Design	Contributors	Denotation	Connotative Aspects	Covered POS	Language Coverage
<b>Dictionaries</b>						
<b>LDOCE</b>	Lexeme-centric: for each lexeme its different senses are described using a controlled vocabulary; for humans	Lexicographers	Descriptions	Can be derived by calculating similarity between descriptions that use the controlled vocabulary	Unrestricted	English
<b>Wiktionary</b>	Lexeme-centric: definitions and descriptions of words in all languages; an article corresponds to one word in one part of speech with all its corresponding definitions; for humans	Collaboratively built; everyone can edit it	Descriptions of concepts in short sentence(s)	Wikisaurus with synonyms, hyponyms and other relations, but word-, not concept-based; categories	Unrestricted	122 languages; 32 language versions with 100,000 or more entries <sup>41</sup>
<b>Thesauri</b>						
<b>Roget's Thesaurus</b> <sup>42</sup>	Idea-centric: groups words that denote related or similar concepts; for humans	Lexicographers	Implicit relations, no descriptions, definitions or examples	Implicit relations	common nouns, verbs, adjectives, adverbs, (idiomatic) phrases	English
<b>Semantic Networks and Ontologies</b>						
<b>WordNet</b>	Concept-centric: resource for linguistic and psycholinguistic research; dictionary for humans	Linguists, lexicographers for core; recently experiments with cloud	Given by the <i>gloss</i> , a short definition	Relations to other synsets (for nouns: hyperonymie, meronymie); recently efforts to add more connotations	Nouns (mainly common nouns, proper names), adjectives, adverbs, verbs	Wordnets for many languages connected by the inter-lingual index
<b>Cyc</b>	Concept-centric: ontology of all commonsense knowledge: knowledge needed in order to understand encyclopedic articles (Lenat, 1995).	Built by a few experts	Natural language definition as comments; partially through relations	Formally defined through relations (e.g. is-a-relations, specific relations such as OLDERTHAN) and attributes (e.g. HARDASAROCK)	Mainly common nouns and proper names	English
<b>Encyclopedias</b>						
<b>Wikipedia</b>	Concept-centric: Encyclopedia for humans	Collaboratively built; everyone can edit it	Usually given by the first few sentences in the article	Categories, internal hyperlink structure, infoboxes	Mainly nouns (common nouns, proper names)	286 languages with 50 languages with 100,000 or more articles <sup>43</sup>

Table 3.1: Comparison of the most prominent inventory used for concept disambiguation based on the requirements from Section 2.4.1.

<sup>41</sup><http://stats.wikimedia.org/wiktionary/EN/Sitemap.htm>, considering only languages with more than 10 articles and which received at least 10 edits within the previous month, 25.2.2014.

<sup>42</sup>Version from 1987.

<sup>43</sup>Statistics from [http://meta.wikimedia.org/wiki/List\\_of\\_Wikipedias](http://meta.wikimedia.org/wiki/List_of_Wikipedias), 5.3.2014.

Inventory	Size		Annotated Instances			Inventory for Downstream Tasks	
	Concepts	Lexicalizations	Data Set	Size	Reference	Hits <sup>44</sup>	Applications
LDOCE <sup>45</sup>	53,838	35,899	Interest Data Set 50 sentences 166 sentences 20 sentences	2,369 occurrences of the noun <i>interest</i> 50 sentences 166 sentences 20 sentences	Bruce & Wiebe (1994) Cowie et al. (1992) Liddy & Paik (1992) Demetriou (1993)	279	In Information Retrieval with unclear results (Krovetz & Croft, 1989)
Wiktionary	EN: 474,128 <sup>46</sup> DE: 73,500	EN: 379,694 <sup>47</sup> DE: 85,574	WK Example	sentences in Wiktionary that illustrate usage of the words		235	<i>Semantic distance calculation</i> (Meyer & Gurevych, 2012; Zesch et al., 2008); <i>question answering</i> (Bernhard & Gurevych, 2009)
Roget's Thesaurus <sup>48</sup>	990	100,470				203	<i>Semantic distance</i> (Kennedy & Szpakowicz, 2008); <i>lexical chains</i> (Morris & Hirst, 1991; Jarmasz & Szpakowicz, 2003)
WordNet <sup>49</sup>	117,659	155,287	WN-Glosses, WN-Examples SemCor Ontonotes DSO Corpus Open Mind Word Expert SensEval & SemEval (all-words task)  SensEval & SemEval (lexical sample task)  SemEval-2007 (coarse-grained all-words task) SemEval-2007 (coarse-grained lexical sample task) Semeval-2010 (domain-specific WSD task) Line-Hard-Serve corpus	in 9,257 glosses 15,179 annotations (XWN project, SensEval-3) 352 texts, 234,136 annotations  192,800 instances of 121 nouns and 70 verbs 29,431 occurrences of 288 nouns annotated by web user 4,979 annotated tokens  24,743 annotated occurrences of 130 nouns, verbs and adjective for training and testing respectively 2,269 annotated tokens  27,132 occurrences for 100 nouns and verbs from OntoNotes 5,342 annotated tokens  around 4000 instances of the noun <i>line</i> , the adjective <i>hard</i> and the verb <i>serve</i>	Moldovan & Novischi (2004); Litkowski (2004) Miller et al. (1993)  Ng & Lee (1996) Chlovski & Mihalcea (2002) (Palmer et al., 2001; Snyder & Palmer, 2004; Pradhan et al., 2007) (Kilgarriff, 2001; Palmer et al., 2001)  (Navigli et al., 2007)  (Pradhan et al., 2007)  (Agirre et al., 2010b)  (Leacock et al., 1998; 1993))	5,070	WordNet is widely used in NLP (Morato et al., 2004), although often with mixed results: <i>semantic distance computations</i> (Resnik, 1993); Pedersen et al., 2004); <i>information retrieval</i> (Moldovan & Mihalcea, 2000; Voorhees, 1999; Gonzalo et al., 1999; Liu et al., 2004; Varelas et al., 2005); <i>question answering</i> (Moldovan & Rus, 2001; Moldovan et al., 2003); <i>text classification</i> (Scott & Matwin, 1998; Kehagias et al., 2001; Moschitti & Basili, 2004); <i>textual entailment</i> (Castillo, 2011; Bentivogli et al., 2011), <i>coreference resolution</i> (Ponzetto & Strube, 2006; Pradhan et al., 2012; Lee et al., 2013); <i>summarization and lexical chains</i> (Hirst & St-Onge, 1998; Barzilay & Elhadad, 1999); <i>semantic role labeling</i> (Gildea & Jurafsky, 2002)

<sup>44</sup><http://aclweb.org/anthology>, 24.3.2014.

<sup>45</sup>The Longman Dictionary of Contemporary English in the version of 1987 (Liddy & Paik, 1992).

<sup>46</sup>Number of concepts extracted by Meyer & Gurevych (2012).

<sup>47</sup>Number of lexicalizations extracted by Meyer & Gurevych (2012).

<sup>48</sup>Version of 1987.

<sup>49</sup>WordNet 3.0: <http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>.

Inventory	Size		Annotated Instances			Inventory for Downstream Tasks	
	Concepts	Lexicalizations	Data Set	Size	Reference	Hits	Applications
			Some annotated corpora in other languages: 10,750 tokens annotated with GermaNet senses: (Henrich et al., 2012); MultiSemCor (Italian translation of SemCor is aligned to English version, 92,420 annotated Italian tokens (Bentivogli & Pianta, 2005)); Estonian all-words task with 5,854 and 5,650 annotated tokens for training and testing respectively (EuroWordNet) (Kahusk et al., 2001)				
Cyc	500,000 <sup>50</sup>	239,000 <sup>51</sup>				196	Cyc has only been integrated in a few research studies – often with negative results, e.g. for <i>textual entailment</i> (Cox, 2005), <i>coreference resolution</i> (Mahesh et al., 1996)
Wikipedia <sup>52</sup>	EN: 4,500,000 DE: 1,700,000	EN: 10,800,000 <sup>53</sup> DE: 2,900,000 <sup>54</sup>	Internal hyperlinks Aquaint50 MSNBC IITB Turdakov ACE 2004 ACE 2005 CoNLL 2003 Wikilinks TAC 2009-2014  Multilingual and cross-lingual annotated corpora are e.g. available from two shared task – TAC for Spanish and Chinese and from NTCIR for Japanese, Korean and Chinese Tang et al. (2011) – and can be derived from internal hyperlinks in other language versions in Wikipedia	more than 92,000,000 linguistic expressions that are linked to another article <sup>55</sup> 50 documents from Aquaint corpus with 727 annotated key phrases 756 annotated proper names 19,751 annotated mentions in 104 documents 8236 annotations in 131 documents 306 annotated mentions 29,300 annotated mentions in 597 documents 34,956 named entities in 1,393 documents 40,323,863 annotated mentions extracted from 10,893,248 web pages that contain hyperlinks to Wikipedia articles 14,320 annotated English proper names after 5 years of TAC	  Milne & Witten (2008b) Cucerzan (2007) Kulkarni et al. (2009) Turdakov & Lizorkin (2009) Ratinov et al. (2011) Bentivogli et al. (2010)  Hoffart et al. (2011b) Singh et al. (2012)  McNamee & Dang (2009), Ji et al. (2010), Ji et al. (2011)	2,810	Widely used in conjunction or as a substitution of WordNet: <i>semantic distance</i> (Gabrilovich & Markovitch, 2007a; Milne & Witten, 2008a; Ponzetto & Strube, 2011; Nastase et al., 2012); <i>information retrieval</i> (Cimiano et al., 2009; Santamaría et al., 2010; Egozi et al., 2011), <i>question answering</i> (Chu-Carroll & Fan, 2011; Ferrández et al., 2007), <i>textual entailment</i> (Bentivogli et al., 2011); <i>coreference resolution</i> (Ponzetto & Strube, 2006; Rahman & Ng, 2011; Ratinov & Roth, 2012; Finin et al., 2009); <i>summarization and topic identification</i> (Zhou et al., 2010b; Nastase, 2008), <i>text categorization and topic identification</i> (Gabrilovich & Markovitch, 2006; Hu et al., 2008; Coursey & Mihalcea, 2009)

Table 3.2: Comparison of resources based on the size, the available annotated data sets and the usage in down-stream applications.

<sup>50</sup>ResearchCyc: <http://www.cyc.com/platform/researchcyc>, 12.3.2014.

<sup>51</sup>OpenCyc: <http://www.cyc.com/platform/opencyc>, 12.3.2014.

<sup>52</sup>Statistics from January 2014: <http://stats.wikimedia.org/EN/TablesArticlesTotal.htm>, 14.3.2014.

<sup>53</sup>This number only includes articles titles and redirects.

<sup>54</sup>This number only includes articles titles and redirects.

<sup>55</sup>This number is derived from the English Wikipedia dump from from January the 4th 2014 and must be higher in the meantime.

For both WordNet and Wikipedia, several data sets of reasonable size are available to test systems, partially due to shared tasks. However, for each language version covered by Wikipedia, annotated data can be derived from the internal hyperlinks, but also by harvesting the web for links to Wikipedia articles (e.g. as in the Wikilinks data set, Table 8.4, Section 8.5). Although this data may be biased to a certain genre and can be erroneous, it is still valid data to extract some statistics (e.g. concept distributions) and to use it as training data (Mihalcea, 2007).

This comparison reveals that Wikipedia is the resource which best fits our requirements. Hence, in this thesis we use Wikipedia as an inventory for concept disambiguation. This is also in accordance with recent developments in the field. The idea to use Wikipedia as an inventory (Bunescu & Paşca, 2006; Cucerzan, 2007; Csomai & Mihalcea, 2008; Milne & Witten, 2008b) has given concept disambiguation a new spin reflected in the numerous shared tasks that have been organized in recent years for disambiguation with respect to Wikipedia at TAC<sup>56</sup>, NTCIR<sup>57</sup>, INEX<sup>58</sup> and SIGIR<sup>59</sup>.

## 3.2 From Wikipedia to an Inventory

Wikipedia has been designed for humans. To use it as an inventory for concept disambiguation or for another natural language processing task the relevant information needs to be extracted and converted into a suitable format.

In the meantime, extracting information from Wikipedia to derive new or to enrich already existing resources has become a research field on its own (Hovy et al., 2013). Ponzetto & Strube (2011) and Nastase & Strube (2013) exploit the categorial backbone of Wikipedia to derive a concept network. Medelyan & Legg (2008) integrate Cyc and Wikipedia. Freebase<sup>60</sup> is a collaborative knowledge base and combines data from Wikipedia with other sources such as MusicBrainz. DBPedia (Auer et al., 2007; Lehmann et al., 2014) consists of information extracted from the structured parts of Wikipedia and contains pointers to other resources such as Freebase or the Project Gutenberg. Yago (Suchanek et al., 2008), BabelNet (Navigli & Ponzetto, 2012a) and UBY (Gurevych et al., 2012) combine WordNet with information and concepts extracted from Wikipedia.

In this thesis, we exclusively rely on Wikipedia. For concept disambiguation, it is important to have a clean inventory. By merging Wikipedia with other resources, noise can be

<sup>56</sup><http://www.nist.gov/tac/2013/KBP/EntityLinking>, 19.3.2014.

<sup>57</sup><http://ntcir.nii.ac.jp/CrossLink/> and <http://ntcir.nii.ac.jp/CrossLink-2/> 19.3.2014.

<sup>58</sup><http://www.inex.otago.ac.nz/tracks/wiki-link/wiki-link.asp>, 19.3.2014.

<sup>59</sup><http://web-ngram.research.microsoft.com/ERD2014/Docs/CFP%20ERD%202014.pdf>, 14.3.2014.

<sup>60</sup>[www.freebase.com](http://www.freebase.com), 25.3.2014.

introduced. For instance, Navigli & Ponzetto (2012a) report a precision for the mapping of Wikipedia pages to WordNet synsets of 0.897 and an F1 measure of 0.771. While such a mapping might be of sufficient quality for some tasks, it may hamper the performance of a concept disambiguation algorithm. However, by linking mentions in a text to Wikipedia articles, down-stream applications can still benefit from such mappings, as the Wikipedia articles can serve as entry points to many resources including Yago, BabelNet, WikiNet or Freebase.

In this section, we describe how we derived our concept inventory from Wikipedia. Our approach consists of two steps. First, we build a separate concept inventory for each language to be considered (Section 3.2.1). Then, we map these separate concept inventories obtained for different languages by exploiting the cross-language links in Wikipedia (Section 3.2.2). The output of this second step is a multilingual resource.

### 3.2.1 Deriving a Concept Inventory from Wikipedia

In this section, we describe how we derive a concept inventory from Wikipedia for a specific language. If this procedure is repeatedly applied to multiple languages, a separate concept inventory is obtained for each of them. In the next section, we will explain how these separate concept inventories extracted for different languages can be mapped to build a multilingual resource.

**Extracting the Concepts.** We consider each Wikipedia article as a concept. To obtain the concepts, we first download the Wikipedia dump for the respective language from *WikiMedia*<sup>61</sup> and extract all articles from this dump. Each of these articles is associated with a unique article identifier. These identifiers serve as our concept identifiers. For instance, the English article TABLE (FURNITURE) has the concept identifier 1147423. The output of this step is a list of concepts for the language under consideration.

**Extracting the Lexicon.** To harvest lexicalizations for the concepts obtained in the last step and to build our lexicon, we leverage the sources described in the following.

*Article name.* We extract all article names from the Wikipedia dumps and associate them with their corresponding concept identifiers. As article names are easier to read than arbitrary numbers, we use them in this thesis to name the concepts.

*Redirects.* The titles of redirects constitute another source for lexicalizations. We extract all these titles and associate them with the identifier of the concept they point to.

*Anchor texts.* More lexicalizations can be obtained from the internal hyperlinks (Figure 3.1). Given an internal hyperlink, the anchor text can be considered as a lexicalization of the Wikipedia article the internal hyperlink points to. We extract all anchor texts and their

<sup>61</sup><http://dumps.wikimedia.org/backup-index.html>, 25.3.2014.

corresponding concepts from the dumps. These anchor texts can be noisy. To reduce the noise, we follow Milne & Witten (2008b) and only consider an anchor as a lexicalization of a concept if it serves at least two times as an anchor for this concept. While this strategy works for English, it is too restrictive for the other languages, as these dumps are smaller. We thus only apply this filtering to English.

From these sources we compile a lexicon. To increase the coverage, we lowercase all lexicalizations. We also experimented with other sources to obtain lexicalizations such as disambiguation pages and bold terms extracted from the first sentence of an article. However, the lexicalizations extracted from these sources are extremely noisy and are not included in our lexicon.

**Extracting Concept Information.** Besides the lexicalizations, we extract for each concept the following information from the Wikipedia dump. This information is used to derive features for concept disambiguation (Section 5.1).

*Inlinks.* For each concept, we extract all inlinks. The set of inlinks for a Wikipedia article (and thus concept) consists of all Wikipedia articles that point to this Wikipedia article via an internal hyperlink (Milne & Witten, 2008a). Figure 3.1 shows a sentence extracted from the Wikipedia article WOODWORKING. This sentence contains internal hyperlinks to the Wikipedia articles FURNITURE, CHAIR, TABLE (FURNITURE), BED, CHEST (FURNITURE) and TOMB. Hence, WOODWORKING is an inlink of FURNITURE, CHAIR, TABLE (FURNITURE), BED, CHEST (FURNITURE) and TOMB. Inlinks indicate connections between Wikipedia articles (and thus concepts) and account for the connotative aspects. If two concepts share an inlink, they co-occur in a text in Wikipedia. Thus, from the inlinks, we can derive concept-level co-occurrence information.

*Sentences.* For each concept, we retrieve all sentences in Wikipedia that link to this concept via an internal hyperlink. For instance, for the concept TABLE (FURNITURE), we store the sentence depicted in Figure 3.1. These sentences serve as annotated examples for the concepts and are used to derive string-level co-occurrence information (Section 5.1).

### 3.2.2 Building a Multilingual Concept Inventory

Our goal is to build a multilingual concept inventory in which concepts are shared across languages. To derive such a multilingual concept inventory, we map the concept inventories extracted from the different language versions of Wikipedia (Section 3.2.1) by adapting the strategy of Nastase & Strube (2013).

In Wikipedia, articles in the different language versions that describe the same concept are interlinked via cross-language links. For instance, the English article TABLE (FURNITURE) is

Language	Concepts	Lexicalizations		
		Articles Names	Redirects	Anchor Texts
en	3,641,010	3,641,010	5,204,184	2,928,125
es	907,397	907,397	1,304,911	151,491
zh	485,887	485,887	383,325	80,271
ja	715,886	715,886	not extracted	not extracted
ko	201,487	201,487	not extracted	not extracted

Table 3.3: Number of concepts and lexicalizations for each language. For Japanese and Korean, we only report the number of concepts and articles names, as we did not extract any other information.

interlinked to the German article TISCH via such a cross-language link. These cross-language links can be leveraged to map the separate concept inventories derived from the different language versions (Section 3.2.1) and to build a multilingual inventory (Nastase & Strube, 2013).

Following Nastase & Strube (2013), we extract all cross-language links from the Wikipedia dumps in different language versions.<sup>62</sup> Given these cross-language links, we map the concepts we extracted for languages other than English to the concepts we extracted for English. To increase the coverage, we apply the triangulation strategy proposed by Wentland et al. (2008): given that a cross-language link is established between concept  $c_a$  and concept  $c_b$  and between concept  $c_b$  and concept  $c_c$ , we insert a cross-language link between  $c_a$  and  $c_c$ .

Hence, after this step, our separate concepts inventories for the different languages are mapped and form together a multilingual resource.

### 3.3 Statistics

In this section, we present some statistics concerning the information we extracted from Wikipedia. We show the statistics for English, Spanish, Chinese, Japanese and Korean.<sup>63</sup> These are the languages we focus on in this thesis (Section 7.4).

**Concepts and Lexicalizations.** Table 3.3 shows for each language the number of concepts and the number of lexicalizations. As we consider Japanese and Korean only as target languages, but not as source languages, we did not extract any lexicalizations apart from the

<sup>62</sup><http://dumps.wikimedia.org/backup-index.html>, 25.3.2014.

<sup>63</sup>We extract the information from the following Wikipedia dumps: English (2012/01/04), Chinese (2012/08/22), Spanish (2012/07/28), Japanese (2010/11/02) and Korean (2011/06/21). To obtain more cross-language links, we additionally considered the following dumps: German (2012/01/16), Italian (2012/01/26), Dutch (2012/01/19), French (2011/02/01), Russian (2011/07/16).

Language	Concepts	Concepts Mapped to English	
en	3,641,010	3,641,010	(100.0%)
es	907,397	545,728	(60.1%)
zh	485,887	240,080	(49.4%)
ja	715,886	294,166	(41.1%)
ko	201,487	101,666	(50.5%)

Table 3.4: Number of concepts that are mapped to English: we report the number of concept in the respective concept inventory and the number of concepts that are mapped to an English concept (absolute and relative).

article names for these two languages. The number of concepts and article names is identical, because each article has exactly one title.

As the numbers show, the concept inventory we derived for English is much bigger than the inventory we obtained for the other languages. It contains more than 3.6 million concepts and millions of lexicalizations, while the Spanish inventory contains 0.9 million concepts and less than two million lexicalizations. The smallest inventory with 0.2 million concepts is the inventory extracted for Korean. The Japanese and Chinese inventories contain 0.71 and 0.48 million concepts respectively.

**Mapping between the Languages.** In Table 3.4, we present some statistics for the mapping of concepts across languages. For each language, we show how many concepts are in its inventory (*Concepts*) and how many of them we could map to a corresponding concept in the English Wikipedia (*Concepts Mapped to English*). For the latter, we report the absolute number and the relative portion.

For Spanish, we were able to map more than 60% of its concepts, while for Chinese and Korean approximately 50% of the concepts is mapped. For Japanese, slightly more than 40% of its concepts is mapped. Some differences between the languages can be explained by the date of the respective Wikipedia dump we used to derive the information. For instance, the Wikipedia dump we used for Japanese is from November 2010, while the Spanish and Chinese dumps are from July and August 2012 respectively. As Wikipedia is a growing resource, it is likely that also the number of cross-language links grows over time. However, some differences between the languages might be due to cultural differences. Some concepts that are for instance relevant in Japanese might not be present in English or – at least – they might be less relevant, so that no article exists.

To measure the impact of the triangulation technique (Wentland et al., 2008), we did some additional comparisons for Japanese and Korean. If we only use the Japanese and the English dump to map the Japanese concepts, 39.0% (corresponding to 279,443 concepts) of the Japanese concepts can be mapped. If we only use the Korean and the English dump to map

Language	Concepts	Concepts with Inlinks			
		in L		in EN	
en	3,641,010	3,158,089	(86.7%)	3,158,089	(86.7%)
es	907,397	735,272	(81.0%)	469,536	(51.7%)
zh	485,887	369,930	(76.1%)	203,454	(41.9%)

Table 3.5: Number of concepts with at least one inlink: we report the number of concepts in the respective language version, the number of concepts with at least one inlink in the respective language version (*in L*) and the number of concepts with at least one inlink in English (*in EN*).

the Korean concepts, 48.6% (corresponding to 97,987 concepts) of the Korean concepts can be mapped. Hence, using information from more language versions via triangulation leads to an increase within the range of 2% (absolute).

**Inlinks.** The inlinks we extracted for each concept provide highly relevant information for our concept disambiguation approach. Multiple features used by our approach depend on the inlinks (Section 5.1). Table 3.5 shows for each language the number of concepts in the inventory and the number of concepts with at least one inlink in the respective language version (*in L*) and in English (*in EN*). Table 3.6 presents the number of inlinks extracted from the respective language version (*in L*) and the number of inlinks that are available for the respective concepts in the English version (*in EN*).

As the Table 3.5 indicates, the majority of concepts is associated with at least one inlink in the respective language version. For 81% of the Spanish concepts, at least one inlink is available in the Spanish version. In the English version, only for slightly more than 51% of the Spanish concepts an inlink is available. A similar behavior can be observed for Chinese: for more than 76% of the Chinese concepts at least one inlink is available in the Chinese version, while only for 41% of them at least one inlink is available in the English version. However, if we consider the total number of inlinks (Table 3.6), much more inlinks are available in the English version for all languages. For Spanish, we extracted approximately 20.4 million inlinks from its language version. The number of inlinks that are available for the Spanish concepts in the English version is more than twice as big. For Chinese, we extracted around 8.9 million inlinks from its language version, while more than 34.8 million inlinks are available for Chinese concepts in the English version.

Hence, for both the Spanish and the Chinese concepts, more inlinks are available in the English Wikipedia than in the respective language version, although for fewer concepts. This indicates that it might be beneficial to use the inlinks extracted from the English Wikipedia to derive disambiguation features for Spanish or Chinese (Section 9.4.3).

Language	Inlinks	
	in L	in EN
en	70,411,071	70,411,071
es	20,141,902	41,073,069
zh	8,913,677	34,835,740

Table 3.6: Number of inlinks: we report the number of inlinks extracted from the respective language version (*in L*) and the number of inlinks that are available for the respective concepts in the English version (*in EN*).

## 3.4 Summary

In this section, we discussed how Wikipedia can be used as an inventory for concept disambiguation. We compared it to other inventories and summarized its advantages and disadvantages. Furthermore, we described the information we extracted from Wikipedia and provided some statistics for the extracted inventory.



# Chapter 4

## Method

In this chapter, we present the core of our approach for concept disambiguation and clustering. Building upon our linguistic insights (Chapter 2), we discuss how we can implement them and model concept disambiguation and clustering in a linguistically sound way using state-of-the-art machine learning methods.

In the following, we first revisit the requirements rooted in linguistic insights (Section 4.1) and outline the consequences in terms of modeling. The aim of this section is to bridge between linguistics and machine learning methods. We then introduce Markov Logic Networks, the relational inference method we have chosen for our approach, present the inference and learning method and discuss its relationship to other relational inference techniques (Section 4.2). Given this background, we explain how we approach concept disambiguation and clustering using Markov Logic Networks. We first model each task in isolation (Section 4.3.1 and Section 4.3.2). Then we integrate the two tasks, while accounting for the interdependencies between them (Section 4.3.3). However, this joint model applies the same context definition for all mentions. In order to turn it into a discourse-aware approach using binwise context modeling (Section 4.4), we need to extend both the model (Section 4.4.1) and the learning method to deal with latent variables (Section 4.4.2).

This chapter focuses on the backbone of our approach. The features are presented in Chapter 5.

### 4.1 Motivation and Prerequisites

Our linguistic analysis of concept disambiguation and clustering (Chapter 2) revealed that the concepts denoted by mentions and the concept-level cohesive ties of mentions mutually depend on each other. Based on this interrelation between denoted concepts and discourse structure – the crux of concept disambiguation and clustering –, we derived three implications for modeling for concept disambiguation and clustering (Section 2.4.2): (1) joint modeling of

interrelated mentions; (2) joint disambiguation and clustering of concepts; and (3) joint context selection and concept disambiguation. These implications are still linguistic formulations. To actually implement them, we need to clarify how we can transfer them into a working disambiguation and clustering algorithm. The key to transforming these linguistics insights into a concrete system lies in implementing the notion of joint modeling.

Joint modeling has become a rapidly growing research field in recent years (Domingos & Richardson, 2007). Various frameworks and approaches have been proposed in the field of network analysis (Newman, 2010), structured prediction (Bakir et al., 2007; Smith, 2011) and statistical relational learning (Getoor & Taskar, 2007b). The idea behind such approaches – usually characterized as *global* (Roth & Yih, 2004), *collective* (Jensen et al., 2004), *relational* (Cheng & Roth, 2013) or *joint* (McCallum, 2009) – is to leverage dependencies between objects (Getoor & Taskar, 2007a). While most work in machine learning assumes that the instances are independent and identically distributed, such approaches drop the independence assumption and model interdependencies between objects (Getoor & Taskar, 2007a). Such interdependencies can occur on different levels. For example, objects can show interdependencies between values of the same attribute such as in concept disambiguation where disambiguation decisions for different mentions influence each other. In this case, interdependencies occur within one single task. Interdependencies can also arise across tasks, i.e. between different attributes of the same or different objects. For instance, concept disambiguation is influenced by concept clustering and vice versa. Similarly, the task of determining the relevant context and the task of disambiguating mentions are interrelated.

The terms *global*, *collective*, *relational* or *joint inference* have become buzz words and are often used interchangeably in the literature. While *collective inference* tends to be used for considering interdependencies within one single task (Jensen et al., 2004), *joint inference* is commonly used if interdependencies between different tasks are modeled (McCallum, 2009). We stick to the term *joint inference* for both cases.

In the following, we discuss three principled ways to implement interdependencies, namely aggregative, iterative and joint approaches. While aggregative and iterative approaches only approximate interdependencies, joint approaches model the joint distribution and directly optimize for the outcome of interdependent objects.

#### 4.1.1 Aggregative Approaches

We summarize under *aggregation* all approaches that approximate interdependencies between objects by consolidating the relational information into local features. The idea of aggregation has been adopted from database theory (Friedman et al., 1999). Given an instance  $X_i$  which is interrelated with a disjoint multiset  $\{X_1, \dots, X_n\}$ , we can aggregate the properties of this multiset into local features using aggregation functions such as for example the mode of the

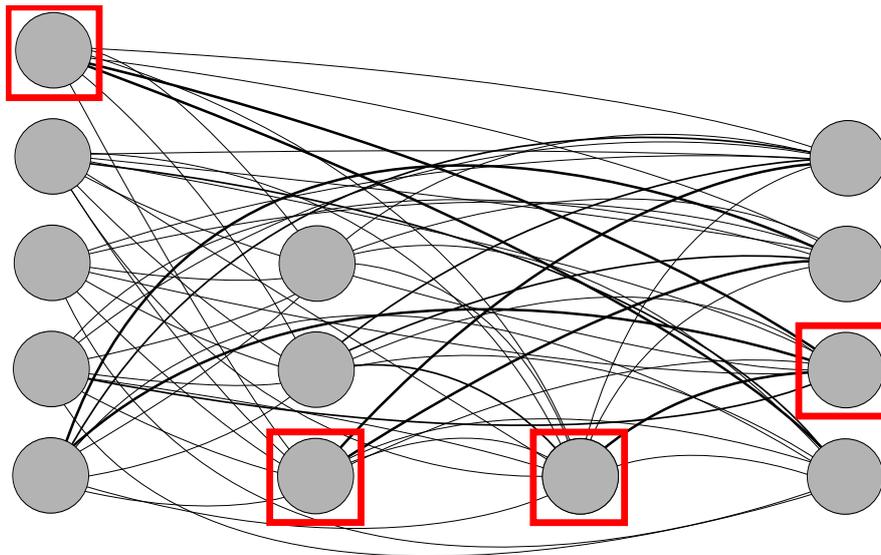


Figure 4.1: Aggregative approaches can be represented as a graph. The vertices represent concepts and the edges relations between them. Each partition corresponds to one mention.

set (e.g. the most frequently occurring value), the mean, median, maximum or minimum or the cardinality of the multiset (Friedman et al., 1999). The local features can be understood as a summary of the multiset (Friedman et al., 1999). Aggregation approximates dependencies between instances and is useful if it is infeasible to explicitly model the interdependencies between instances (Getoor & Taskar, 2007a).

Such aggregative approaches can be represented as an  $n$ -partite graph (Figure 4.1). Each partition corresponds to one object (in our case a mention). Each attribute value is represented as a vertex and interdependencies between attribute values are indicated by directed or undirected edges. The strength of the interdependency is indicated by the edge weight. Given such a graph, centrality measures that rank the vertices according to their centrality in the graph offer possibilities to aggregate the interdependencies into local feature values. For instance, by using measures such as *degree*, *betweenness* or *PageRank* each vertex is assigned a score (e.g. Newman (2010, p. 168-193)). This score takes into account the topology of the graph given by the interconnections between the vertices and can be interpreted as an aggregated value of a multiset.

In word sense or concept disambiguation, such graph-based approaches are popular to model interdependencies between senses or concepts (Navigli & Lapata, 2010; Agirre & Soroa, 2009; Mihalcea et al., 2004b). The approach we used in Section 2.2.2 is also such a graph-based approach. In such approaches, each vertex represents a concept, while the edge scores are usually given by a relatedness or similarity measure. Given this graph representation, words are disambiguated in two steps. In the first step, each vertex is assigned a score

based on its centrality. Popular measures to score vertices for concept disambiguation are *degree* and *PageRank*. In the second step, for each word the concept with the highest score is selected and considered as its denoted concept. Although such approaches model the interconnection between related concepts in the aggregation step, they fail to directly optimize for the best overall disambiguation decision. Most centrality measures have been developed in social network analysis (Newman, 2010, p. 168), where all vertices represent relevant objects. In contrast, in concept disambiguation only few vertices represent correct candidate concepts, while the majority of vertices are wrong candidate concepts that are irrelevant or even misleading. Aggregating information from all these wrong candidate concepts biases the centrality scores of the vertices and can lead to wrong decisions (Section 2.2.2).

Aggregation methods can also be used to model interdependencies between two different tasks, e.g. between context selection and disambiguation. However, as in case of disambiguation, the interdependencies are approximated, but decisions are not taken jointly.

In summary, aggregations are a way to approximate interdependencies and are helpful if it is infeasible to model the interdependencies jointly.

### 4.1.2 Iterative Approaches

Iterative approaches are another way to approach interdependencies between instances. The intuition behind iterative approaches is that for some instances the solution is known a priori (Macskassy & Provost, 2003; Chakrabarti et al., 1998) or can be more easily obtained than for others (Milne & Witten, 2008b). The unknown or more difficult instances are then solved given the solutions of the known or easily solvable instances. Hence, interdependencies are not modeled between all instances, but only with respect to a subset of previously known or solved instances. Instead of solving all instances in two steps, more iterations are possible as for instance in relaxation labeling (Chakrabarti et al., 1998).

Initialization is crucial for such iterative approaches. If the instances in the first step are incorrectly solved, all other instances are likely to also be incorrectly solved. Beside the quality of the initialization, the relevance of presolved instances to solve other instances is critical. If a *high quality* initialization of instances that are *relevant* to solve others is provided, iterative approaches are successful (Milne & Witten, 2008b). If this is not the case, iterative approaches might suffer from error propagation.

Iterative approaches only approximate interdependencies without directly optimizing for the overall joint decision. However, iterative approaches are efficient as they reduce the complexity of the joint optimization problem by solving instances given presolved instances. The interdependencies between an unsolved instance and a set of presolved instances are often aggregated by using a weighted average score or the maximum value (Milne & Witten, 2008b; Ratinov et al., 2011) (Section 4.1.1). Hence, iterative approaches are usually at the same time

also aggregative.

In concept disambiguation, iterative approaches have been successful to model interdependencies between disambiguation decisions. For instance, Milne & Witten (2008b) make use of the long-tailed Zipfian distribution over candidate concepts of a mention which implies that given a mention and its candidate concepts, a few candidate concepts are much more frequent than all others. This unequal frequency distribution of candidate concepts given a mention is also responsible for the high performance of the *most frequent concept baseline* (Section 9.1.1). Milne & Witten (2008b) assume that for some mentions the frequency distribution over candidate concepts is extremely biased towards one single concept, such that this concept is the correct disambiguation in at least 95% of the cases. They further assume that mentions with such a frequency distribution occur sufficiently frequent in all texts and are at the same time relevant to disambiguate all other mentions. Hence, they first disambiguate mentions with such biased frequency distributions and then use interdependencies to these disambiguated mentions to solve all other mentions. Ratnov et al. (2011) drop these assumptions and first disambiguate all mentions without considering interdependencies. In a second step, all mentions are resolved again, but this time interdependencies between concepts are taken into account: given a mention to solve, interdependencies with the concepts that have been selected for all other mentions in the first step are modeled using aggregation.

Iterative approaches only converge to a local optimum and not necessarily to the overall optimal joint decision (Chakrabarti et al., 1998). If the initialization is close enough to the correct solution, such iterative approaches are an efficient way to model interdependencies.

### 4.1.3 Joint Inference

Joint inference approaches are genuinely joint. Their objective function directly considers the joint outcome of interrelated instances. Such approaches are often directed or undirected probabilistic graphical models combining statistics with a logic- or frame-based representation (Getoor & Taskar, 2007a). Most approaches are, at their core, linear models. In linear models, features are combined linearly. Each possible assignment  $\mathbf{y}$  to the instances  $\mathbf{x}$  is scored by (Smith, 2011, p. 24)

$$\begin{aligned} score(\mathbf{x}, \mathbf{y}) &= \sum_{j=1}^d w_j g_j(\mathbf{x}, \mathbf{y}) \\ &= \mathbf{w}^T \mathbf{g}(\mathbf{x}, \mathbf{y}), \end{aligned}$$

i.e. by summing over all  $d$  feature functions  $g_j(\mathbf{x}, \mathbf{y})$  weighted by a weight  $w_j$ . Note that  $\mathbf{x}$  and  $\mathbf{y}$  denote groups of instances, i.e.  $g(\mathbf{x}, \mathbf{y})$  comprises both local and global features. The solution is then given by (Smith, 2011, p. 24):

$$\mathbf{y} = \arg \max_{\mathbf{y} \in \mathcal{Y}_{\mathbf{x}}} \mathbf{w}^T \mathbf{g}(\mathbf{x}, \mathbf{y}).$$

Among all possible assignments  $\mathcal{Y}_{\mathbf{x}}$  for the instances  $\mathbf{x}$ , the solution  $\mathbf{y}$  is selected that maximizes the score of the linearly combined feature values.

Similar to aggregative and iterative approaches, joint approaches can usually be represented using a graph. In contrast to aggregative approaches, the vertices do not represent attribute values, but (discrete) random variables. Edges between these vertices model dependencies between the random variables.

Joint inference during prediction does not necessarily imply that interdependencies are also modeled during learning (Chang et al., 2012; Roth & Yih, 2004). Constraint conditional models (Chang et al., 2012) allow to train separate models which are then combined at inference time using soft or hard constraints. In contrast, other approaches such as Markov Networks always model interdependencies both in the learning and the training phase. Both approaches have advantages and disadvantages. Ignoring interdependencies in the learning phase and only modeling them at inference time has the advantage that parameter learning is less expensive and models for different tasks can be trained on different data sets which is not possible if interdependencies are also modeled during learning. However, joint learning allows to optimize the weights of global features that model interdependencies between instances in a statistically sound way.

Compared to aggregative and iterative approaches, joint inference is more expensive. Depending on the complexity of the problem, i.e. the number and the density of interdependencies, exact inference might be unfeasible. This leads to the need for approximate inference. Such approximate inference methods are well studied and often criteria are formulated under which they are successful. In contrast, aggregative and iterative approaches, which also approximate the notion of joint inference, are more ad hoc, in the sense that their success criteria are less understood.

#### 4.1.4 Discussion

We have presented three general strategies to implement the notion of joint modeling. While aggregative and iterative approaches are less expensive than genuinely joint methods, they approximate the notion of interdependencies between instances in a less statistically founded way. Only approaches based on joint inference directly optimize the overall outcome of interdependent instances. In contrast to aggregative and iterative approaches, joint approaches allow to explicitly state and formalize the interdependencies. Hence, approaches based on joint inference comply best with our linguistic requirements.

In this thesis, we therefore propose to approach concept disambiguation and clustering in

a framework that allows for joint parameter learning and joint inference. Such a framework is most appropriate for our requirements. However, as joint approaches are expensive, one challenge is to balance between appropriateness and tractability. To keep our approach tractable, we need to model some interdependencies via an aggregative approach. Such a hybrid strategy is a common practice to keep inference tractable (Getoor & Taskar, 2007a).

We chose Markov logic as our framework. Markov logic has been proposed as a unifying framework for statistical relational learning (Domingos & Richardson, 2007) and fits our requirements best. One of the big advantages of Markov logic is that it combines the expressiveness of first-order logic with probabilities. It allows us to concisely model and formalize the interdependencies – we derived in the last chapter – in first-order logic. First-order logic not only allows for a compact presentation of our model, but it is also widespread and easy to read. At the same time, Markov logic is based on Markov networks and thus statistically well founded. Markov networks can be represented as log-linear models, which in turn are well understood. These properties of Markov logic made it a successful framework for many natural language processing tasks.

Domingos & Richardson (2007) discuss the relationship to other statistical relational learning approaches and show that many of them such as *probabilistic relations models* (Friedman et al., 1999) or *stochastic logic programs* (Muggleton, 1995) can be converted to Markov logic. Markov logic is closely related to the *constrained conditional model* framework (Chang et al., 2012). Similar to Markov logic, constrained conditional models allow to inject declarative knowledge or relational dependencies via soft or hard constraints into the inference process. However, while in Markov logic this global information is used both during learning and inference, constrained conditional models allow to learn different parts of the model separately that are then joined via soft or hard constraints during inference. This decoupling in constrained conditional models leads to a higher flexibility, but at the same time also opens the question how the different parts should be combined. Hence, the constrained conditional model framework would be an alternative to Markov logic, in particular, if the different parts of the model are coupled via hard constraints. As we aim to learn the weights jointly, constrained conditional models would not add any additional benefits. We therefore decided to use Markov logic. Moreover, as explained above, the modeling language in Markov logic is based on first-order logic, which is easily to read, whereas in constrained conditional models a special language, *Learning Based Java*, is used (Rizzolo & Roth, 2010).

In the following we discuss Markov logic and show how we implement our linguistic insights in this framework.

## 4.2 Markov Logic

*Markov Logic (ML)* combines first-order logic with probabilistic graphical models (Domingos & Lowd, 2009) and has been proposed as a unifying framework for statistical relational learning to promote research in this field (Domingos & Richardson, 2007, p. 340). By decoupling the inference and learning algorithms from the representation language, algorithms and applications can be iteratively improved leading to faster progress (Domingos & Lowd, 2009, p. 1–4). Benefiting from the expressiveness and flexibility of first order logic and the robustness of graphical models, different state-of-the-art applications can be implemented in the same framework (Domingos & Lowd, 2009, p. 3–4) allowing to exploit relations between them. For instance, Markov Logic has been recently used for various natural language processing tasks such as coreference resolution (Poon & Domingos, 2008; Bögel & Frank, 2013), sentiment analysis (Zirn et al., 2011) and joint semantic role labeling and word sense disambiguation (Meza-Ruiz & Riedel, 2009; Che & Liu, 2010). The core of Markov Logic is a log-linear model. It therefore belongs to a class of models that is well understood (Smith, 2011). In the meantime, several implementations have been developed for Markov Logic, e.g. *Alchemy* (Kok et al., 2005), *thebeast* (Riedel, 2008), *RockIt* (Noessner et al., 2013) or *Tuffy* (Niu et al., 2011), enabling us to build upon previous work. In the following, we first present the basic ideas of Markov Logic. We then show different inference and learning strategies (Section 4.2.3 and 4.2.4) and compare different available implementations (Section 4.2.5).

As Markov Logic integrates two research paradigms, it can be analyzed from two different perspectives (Domingos & Lowd (2009, p. 5) and Riedel (2009, p. 39–40)): it can be understood as a probabilistic extension of first-order logic accounting for uncertainty (Section 4.2.1) or as a compact representation of Markov networks (Section 4.2.2).

### 4.2.1 Markov Logic as a Probabilistic Extension of First-order Logic

In this section, we discuss Markov logic from the perspective of first-order logic. Extending first-order logic by integrating probabilities, Markov logic allows to account for uncertainty. Before discussing the benefits of Markov logic from the perspective of first-order logic, we introduce some basic concepts and notations of first-order logic that are essential to understand Markov logic. We assume that the reader has some basic understanding of logic.

**First-order logic.** *First-order logic* is an expressive formal language that models a world or a domain by its objects, i.e. its individuals and properties, and the relations between them (Russell & Norvig, 1995, p. 185–186). Despite its limitations – given by modeling a domain only in terms of objects and relations and by its restrictions with respect to quantification –, first-order logic is still expressive, at the same time well understood and therefore highly

ID	Formula	Description
$f_1$	$\forall m, n, c$ <i>hasSameConcept</i> ( $m, n$ ) $\wedge$ <i>hasConcept</i> ( $m, c$ ) $\implies$ <i>hasConcept</i> ( $n, c$ )	If two mentions $m$ and $n$ denote the same concept ( <i>hasSameConcept</i> ( $m, n$ )) and mention $m$ denotes the concept $c$ ( <i>hasConcept</i> ( $m, c$ )), the other mention also needs to denote the concept $c$ .

Table 4.1: Example knowledge base in first-order logic with one formula  $f_1$ .  $m$  and  $n$  are variables that range over mentions, whereas  $c$  is a variable that ranges over concepts.

popular in artificial intelligence (Russell & Norvig, 1995, p. 186).

Table 4.1 shows a formula in first order logic. This formula states that if two mentions  $m$  and  $n$  share the same concept and mention  $m$  denotes concept  $c$ , mention  $n$  also needs to denote concept  $c$ . In the following, we write the first letter in upper-case to indicate that a string is a constant. For variables, we use lower-case letters. In the formula in Table 4.1, the two variables  $m$  and  $n$  range over mentions, while the variable  $c$  ranges over concepts. Predicates (e.g. *hasSameConcept*) are written in italics.

For the following discussion, it is important to understand the concept of *grounding*.

**Definition 4.1** A *ground term* is a term without any variables (Russell & Norvig, 1995, p. 190). A *ground predicate* is a predicate that only contains ground terms as arguments (Richardson & Domingos, 2006, p. 4). If we say that we ground some formulas with a set of constants  $K$ , we mean that each formula is instantiated with all possible combinations of constants.

For instance, if we ground the formula in Table 4.1 with the constants *Tiger* and *Tigers* (mentions) and TIGER (ANIMAL) (concept), we would obtain the ground formulas shown in Table 4.2.

Dependent on an interpretation, i.e. the correspondence between the symbols and the domain objects, and the actual state of the respective domain, a ground predicate can either be true or false (Russell & Norvig, 1995, p. 163). A *possible world* or a *Herbrand interpretation* assigns to each ground predicate a truth value (Richardson & Domingos, 2006, p. 4). The truth value of a ground formula can be recursively determined by considering the truth values of its parts (Russell & Norvig, 1995, p. 168).

First-order logic lacks a possibility to express degrees of truthness or uncertainty (Russell & Norvig, 1995, p. 165), e.g. a ground predicate can only be true or false, but not 70% true.

The formula in Table 4.1 can be considered as a knowledge base consisting of one single formula. More generally, we define a *first-order logic knowledge base* as a collection of (implicitly) conjuncted first-order logic formulas (Domingos & Richardson, 2007, p.342).

We now discuss Markov logic from the perspective of first-order logic.

ID	Formula	Ground Formula
$f_1$	$\forall m, n, c$ $hasSameConcept(m, n)$ $\wedge hasConcept(m, c)$ $\implies hasConcept(n, c)$	$hasSameConcept(Tiger, Tigers)$ $\wedge hasConcept(Tiger, TIGER (ANIMAL))$ $\implies hasConcept(Tigers, TIGER (ANIMAL))$  $hasSameConcept(Tigers, Tiger)$ $\wedge hasConcept(Tigers, TIGER (ANIMAL))$ $\implies hasConcept(Tiger, TIGER (ANIMAL))$  $hasSameConcept(Tiger, Tiger)$ $\wedge hasConcept(Tiger, TIGER (ANIMAL))$ $\implies hasConcept(Tiger, TIGER (ANIMAL))$  $hasSameConcept(Tigers, Tigers)$ $\wedge hasConcept(Tigers, TIGER (ANIMAL))$ $\implies hasConcept(Tigers, TIGER (ANIMAL))$

Table 4.2: Grounding of Formula  $f_1$  with the constants *Tiger* and *Tigers* (mentions) and TIGER (ANIMAL) (concept).

**Markov Logic.** While first-order logic allows to describe rich connections in a domain, it lacks a means to express uncertainty. As a consequence, a world that violates one single formula is impossible and has zero probability (Domingos & Richardson, 2007, p. 344). At the same time, a world that only violates one single formula is equally impossible than a world that violates all formulas. This is a major drawback if first-order logic is used to model domains in which uncertainty is highly present. For instance, to describe linguistic phenomena, formulas that are always true or false are of limited applicability. In fact, in such domains formulas are true with a certain probability, e.g. they are hardly ever, sometimes, often or mostly true.

Markov logic addresses this discrepancy between modeling requirements and first-order logic and extends first-order logic by a means to deal with uncertainty. Each formula is associated with a weight. The weight indicates how expensive it is to violate the corresponding formula (Domingos & Richardson, 2007, p. 344). Consequently, a world that violates some formulas is not impossible as in first-order logic, but only less probable than a world in which all formulas are true (Domingos & Richardson, 2007, p. 344).

A Markov logic network is a knowledge base in which each formula is assigned a weight or more formally:

**Definition 4.2** A *Markov logic network (MLN)* consists of pairs  $(F_i, w_i)$  with  $F_i$  being a formula in first-order logic and  $w_i \in \mathbb{R} \cup \{\pm\infty\}$  is the weight assigned to it (Domingos & Richardson, 2007, p. 344).

Hence, a first-order logic knowledge base can be transformed into a MLN by assigning

each formula a weight. These weights can be either manually determined, for example by a domain expert, or learned from data (Domingos & Lowd, 2009, p. 5). Table 4.3 shows a MLN with one single formula. In the following, we distinguish between hard and soft constraints depending on the weight of the formula. While hard constraints are useful to model rules that must be satisfied, soft constraints account for uncertainty.

**Definition 4.3** A *hard constraint* is a first-order logic formula with infinite weight. A world that violates a hard constraint with weight  $+\infty$  is impossible. A world that does not violate a hard constraint with weight  $-\infty$  is also impossible (Domingos & Lowd, 2009, p. 4-5).

**Definition 4.4** A *soft constraint* is a first-order logic formula with non-infinite weight and can be violated (Domingos & Lowd, 2009, p. 4-5).

A knowledge base in first-order logic only consists of hard constraints and categorizes the possible worlds into two classes: worlds that satisfy the constraints and worlds that violate constraints (Domingos & Lowd, 2009, p. 17). In contrast, a MLN that also contains soft constraints partitions the possible worlds in more fine-grained subsets and assigns each of them a different probability. In fact, MLNs define probability distributions over possible worlds (Domingos & Richardson, 2007, p. 344). Hence, instead of classifying possible worlds as violating or non-violating, they are ranked according to their probabilities. By adding more formulas to a MLN, its granularity is increased leading to a more fine-grained ranking over possible worlds (Domingos & Lowd, 2009, p. 17).

## 4.2.2 Markov Logic as a Compact Representation of Markov Networks

A Markov logic network defines a probability distribution over possible worlds. While we discussed the benefits of assigning a probability to a possible world in Section 4.2.1, we focus in this section on how this probability distribution over possible worlds is defined.

From the perspective of probabilistic graphical models, Markov logic is a way to compactly represent Markov networks. In the following, we first introduce Markov networks, before we discuss the relationship between Markov logic and Markov networks.

**Markov networks.** A Markov network, also known as a Markov random field, defines a joint probability distribution over a set of random variables (Smith, 2011, p. 27).<sup>1</sup> In the following,  $\mathbf{X} = \langle X_1, X_2, \dots, X_n \rangle$  denotes a collection of random variables. Value assignments to random variables are indicated by lower-case letters, e.g.  $X_1 = x_1$ , while bold letters denote collections of variables or values. The space of all possible assignments to  $\mathbf{X}$  is represented by  $\mathcal{X}$ . The probability of a specific assignment  $\mathbf{x} \in \mathcal{X}$  to a set of variables  $\mathbf{X}$  is then written as

<sup>1</sup>A more concise description of Markov networks can be found in (Koller et al., 2007, p. 103-156).

ID	Weight	Formula	Ground Formula
$f_1$	0.7	$\forall m, n, c \text{ hasSameConcept}(m, n)$ $\wedge \text{hasConcept}(m, c)$ $\implies \text{hasConcept}(n, c)$	$\text{hasSameConcept}(Tiger, Tigers)$ $\wedge \text{hasConcept}(Tiger, \text{TIGER (ANIMAL)})$ $\implies \text{hasConcept}(Tigers, \text{TIGER (ANIMAL)})$  $\text{hasSameConcept}(Tigers, Tiger)$ $\wedge \text{hasConcept}(Tigers, \text{TIGER (ANIMAL)})$ $\implies \text{hasConcept}(Tiger, \text{TIGER (ANIMAL)})$  $\text{hasSameConcept}(Tiger, Tiger)$ $\wedge \text{hasConcept}(Tiger, \text{TIGER (ANIMAL)})$ $\implies \text{hasConcept}(Tiger, \text{TIGER (ANIMAL)})$  $\text{hasSameConcept}(Tigers, Tigers)$ $\wedge \text{hasConcept}(Tigers, \text{TIGER (ANIMAL)})$ $\implies \text{hasConcept}(Tigers, \text{TIGER (ANIMAL)})$

Table 4.3: Formula with its associated weight. On the right-hand side, the Formula  $F_1$  is grounded with the constants *Tiger* and *Tigers* (mentions) and *TIGER (ANIMAL)* (concept). Each ground formula of  $F_1$  is associated with the same weight.

$$P(\mathbf{X} = \mathbf{x}) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n).$$

Markov networks can be represented as log-linear models. In the rest of this thesis we use this log-linear model representation defined as

$$P_w(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp \left( \sum_i w_i \cdot f_i(\mathbf{x}_i) \right) \quad (4.1)$$

where  $f_i(\mathbf{x}_i)$  is a feature and  $w_i$  is a weight. In this context, a feature is a function that takes the values  $\mathbf{x}_i$  of some random variables  $\mathbf{X}_i \subset \mathbf{X}$  and maps them to a value in  $\mathbb{R}$  (Koller et al. (2007, p. 104;p. 125)). Such a value is not normalized and can take any value. It can be interpreted as the affinity or compatibility of a specific assignment to a set of random variables, i.e. the higher the value is, the higher is the affinity of the corresponding assignment (Koller et al., 2007, p. 104-107). The set of variables  $\mathbf{X}_i$  considered by a feature  $f_i$  is called the scope of a feature (Koller et al., 2007, p. 104).

As the scores returned by the features are not normalized and can take any value, the overall sum is normalized by the so-called partition function  $Z$ . The partition function sums over all possible assignments in  $\mathcal{X}$  and is given by

$$Z = \sum_{\mathbf{x} \in \mathcal{X}} \exp \left( \sum_i w_i \cdot f_i(\mathbf{X}_i) \right).$$

By normalizing with  $Z$  it is guaranteed that the result is a normalized joint distribution (Koller et al., 2007, p. 104-105). Without this normalization, Formula 4.1 would not describe a probability distribution.

**Markov logic networks.** From the perspective of Markov networks, Markov logic allows to compactly define the skeleton of a Markov network. Let  $F_k$  be a formula in a MLN. Let  $F_{k,m}$  be the  $m^{\text{th}}$  grounding of the formula  $F_k$ . Let  $w_k$  be the weight associated with the formula  $F_k$ . We can now define how the Markov network corresponding to a MLN can be constructed.

**Definition 4.5** *Given a MLN  $L$  and a finite set of constants  $\mathbf{K}$ , a Markov network  $M_{L,\mathbf{K}}$  can be constructed in the following way:*

- (i) *For each ground predicate a binary variable  $X_i$  is introduced.*
- (ii) *For each ground formula  $F_{k,m}$  a feature  $f_{k,m}$  with scope  $\mathbf{X}_{k,m}$  is introduced.  $\mathbf{X}_{k,m}$  is the set of binary variables that correspond to the ground predicates present in the ground formula  $F_{k,m}$ .*
- (iii) *The weight for a feature  $f_{k,m}$  corresponds to the weight  $w_k$  that is associated with the corresponding formula  $F_{k,m}$ , i.e. all ground formulas  $F_{k,m}$  have the same weight  $w_k$ .*

The assignment of a binary variable  $X_i$  corresponding to a ground predicate depends on the truth value of the ground predicate as follows

$$X_i = \begin{cases} 1 & \text{if the corresponding ground predicate is true} \\ 0 & \text{otherwise.} \end{cases}$$

Each feature  $f_{k,m}$  is an indicator function that maps each possible assignment of the variables in its scope to 1 or 0 depending on the truth value of the corresponding ground formula  $F_{k,m}$ . It is defined as

$$f_{k,m} = \begin{cases} 1 & \text{if the corresponding ground formula } F_{k,m} \text{ is true} \\ 0 & \text{otherwise.} \end{cases}$$

Representing the Markov network  $M_{L,\mathbf{K}}$  as a log linear model, the probability distribution is given by

$$P_w(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp \left( \sum_k w_k n_k(\mathbf{x}) \right)$$

where  $n_k(\mathbf{x})$  is the number of true groundings of formula  $F_k$  in  $\mathbf{x}$ .

MLNs also allow to use numeric features, instead of binary ones. In this case a feature  $f_{k,m}$  also includes a score  $q_{k,m}$  and is defined as

$$f_{k,m} = \begin{cases} q_{k,m} & \text{if the corresponding ground formula } F_{k,m} \text{ is true} \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the probability distribution is given by

$$P_w(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp \left( \sum_k w_k q_k(\mathbf{x}) \right)$$

where  $q_k(\mathbf{x})$  is obtained by

$$q_k(\mathbf{x}) = \sum_m f_{k,m}(\mathbf{x}_{k,m}).$$

Grounding a MLN  $L$  with different sets of constants  $\mathbf{K}_i$  leads to different Markov networks  $M_{L,\mathbf{K}_i}$ . Although these derived Markov networks vary in size determined by the respective finite set of constants  $\mathbf{K}_i$ , they show structural regularities and share parameters as defined by the MLN  $L$ . Hence, a MLN is a “template for constructing Markov networks” (Domingos & Richardson, 2007, p. 344).

### 4.2.3 Inference

Given a MLN  $L$  and a set of constants  $\mathbf{K}$ , we can perform maximum a posteriori inference. We assume that we have observed and hidden predicates in our MLN.

**Definition 4.6** A predicate is **observed** if the truth values of the corresponding ground predicates are given during prediction. We denote the variables in the Markov network denoting observed ground predicates as  $\mathbf{O} = (O_1, \dots, O_o)$ .

For instance the predicate *shareSameLemma*( $m, n$ ) is observed, as we can extract from the data for which mention pairs this relation holds.

**Definition 4.7** A predicate is **hidden** if the true values of the corresponding ground predicates are unknown during prediction. We denote the variables in the Markov network denoting hidden ground predicates as  $\mathbf{H} = (H_1, \dots, H_h)$ .

Hidden predicates describe the relations we aim to predict. For instance, the predicate *hasSameConcept*( $m, n$ ) that is true if two mentions denote the same concept is hidden in our data and part of our prediction.

**Definition 4.8** *The maximum a posteriori inference (MAP inference) over a MLN  $L$  and a set of constants  $\mathbf{K}$  determines the most probable truth values of the hidden predicates given the truth values of the observed predicates.*

In the log-linear model, MAP inference is equivalent to find the assignment  $\mathbf{h}$  for the hidden predicates  $\mathbf{H}$  given  $\mathbf{O} = \mathbf{o}$  that maximizes the probability score, i.e.

$$\arg \max_{\mathbf{h} \in \mathcal{H}} P(\mathbf{H} = \mathbf{h} | \mathbf{O} = \mathbf{o}) = \arg \max_{\mathbf{h} \in \mathcal{H}} \frac{P(\mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o})}{\sum_{\mathbf{h}' \in \mathcal{H}} P(\mathbf{H} = \mathbf{h}', \mathbf{O} = \mathbf{o})}.$$

As the denominator is constant for all assignments, it can be dropped leading to

$$\begin{aligned} \arg \max_{\mathbf{h} \in \mathcal{H}} P(\mathbf{H} = \mathbf{h} | \mathbf{O} = \mathbf{o}) &= \arg \max_{\mathbf{h} \in \mathcal{H}} P(\mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o}) \\ &= \arg \max_{\mathbf{h} \in \mathcal{H}} \frac{1}{Z} \exp \left( \sum_k w_k n_k(\mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o}) \right). \end{aligned}$$

This can be further simplified by ignoring  $Z$  as  $Z$  is constant for all possible assignments to  $H$  and by using the log-probability  $\log P(\mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o})$  instead of the probability, which is maximal for the same assignments because the log function is monotonous. This leads to

$$\begin{aligned} \arg \max_{\mathbf{h} \in \mathcal{H}} P(\mathbf{H} = \mathbf{h} | \mathbf{O} = \mathbf{o}) &= \arg \max_{\mathbf{h} \in \mathcal{H}} P(\mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o}) \\ &= \arg \max_{\mathbf{h} \in \mathcal{H}} \log P(\mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o}) \\ &= \arg \max_{\mathbf{h} \in \mathcal{H}} \sum_k w_k n_k(\mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o}). \end{aligned}$$

Finding the MAP solution in a MLN is a NP-hard problem (Roth, 1996). Domingos & Richardson (2007, p. 358) propose to approximate MAP inference in MLNs using a MaxWalk-Sat algorithm (Kautz et al., 1996), but also other approximation techniques such as simulated annealing or belief propagation are possible (Riedel, 2009, p. 31-33). In this thesis, we follow the approach of Riedel (2008) in which the MAP inference problem is transformed into an *Integer Linear Program (ILP)*, i.e. a constrained optimization problem. More precisely, as the random variables in MLNs are binary, it can be mapped to a 0-1 Linear Program (Riedel, 2009, p. 28–31). To solve such an ILP an off-the-shelf commercial or non-commercial ILP solver can be used such as *Gurobi* (Gurobi Optimization, 2014) or *lp\_solve*<sup>2</sup>.

In order to speed up the inference process, Riedel (2009) proposes to solve the ILP with a *Cutting Plane Inference* technique proposed for constrained optimization problems in Operations Research (Dantzig et al., 1954). The idea is to first solve a partial problem by only

<sup>2</sup><http://sourceforge.net/projects/lpsolve/>, 14.5.2014.

considering a few constraints. Then iteratively more and more constraints are added and the problem is solved accordingly until all constraints are included. For a complete discussion of the algorithm and the transformation of a MLN problem to an ILP, the reader is referred to Riedel (2008) and Riedel (2009).

We follow the inference approach of Riedel (2008) as it has been shown to be more accurate, while being faster at the same time than a MaxWalkSat algorithm (Riedel, 2008). If the ILP solver terminates, the algorithm returns the optimal MAP solution for a ground MLN. In addition, current state-of-the-art ILP solvers – as for instance *Gurobi* – have recently shown good performance in solving complex NLP problems (Cheng & Roth, 2013; Do et al., 2012).

#### 4.2.4 Learning

The weights of a MLN can be learned generatively (e.g. Richardson & Domingos (2006)) or discriminatively (e.g. Singla & Domingos (2005) and Lowd & Domingos (2007)). Generative weight learning maximizes the likelihood of a data set, i.e. the joint distribution of all variables. However, if it is known a priori which variables are hidden and which are observed as in our case, it is more efficient and often more accurate to optimize the conditional likelihood of the hidden variables given the observed ones (e.g. Singla & Domingos (2005)), i.e. to learn the weights discriminatively. In this thesis, we therefore use a standard discriminative learning approach based on gradient ascent.

*Gradient ascent* is a standard approach to find the global optimum of a function  $Q_w$  given that it is concave as in the case of MLNs. The weight vector  $\mathbf{w}$  is updated at each iteration  $t$  by using the gradient  $\nabla_w Q_w$  of the function to optimize multiplied by a learning rate  $\eta$  (Lowd & Domingos, 2007), i.e.

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta \nabla_w Q_w.$$

In the case of MLNs, the aim is to find the weight vector  $\mathbf{w}$  that maximizes the conditional log-likelihood  $\log P_w(\mathbf{H} = \mathbf{h} | \mathbf{O} = \mathbf{o})$ , i.e.

$$Q_w = \log P_w(\mathbf{H} = \mathbf{h} | \mathbf{O} = \mathbf{o}).$$

Under the assumption that annotated data is available for all hidden predicates during learning, the gradient of the conditional log-likelihood  $\nabla_w \log P_w(\mathbf{H} = \mathbf{h} | \mathbf{O} = \mathbf{o})$  is given by (Lowd & Domingos, 2007)

$$\frac{\partial}{\partial w_k} \log P_w(\mathbf{H} = \mathbf{h} | \mathbf{O} = \mathbf{o}) = n_k(\mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o}) - E_w[n_k(\mathbf{H} | \mathbf{O} = \mathbf{o})].$$

Hence,  $n_k(\mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o})$  is the number of true groundings of formula  $F_k$  according to the ground truth in the training data and  $E_w[n_k(\mathbf{H}|\mathbf{O} = \mathbf{o})]$  denotes the expected number of true groundings of  $F_k$  under  $P_w$  given that  $\mathbf{O} = \mathbf{o}$ . The calculation of the expected number of true groundings  $E_w[n_k(\mathbf{H}|\mathbf{O} = \mathbf{o})]$  is intractable (Singla & Domingos, 2005). However, Singla & Domingos (2005) adapt the *voted perceptron algorithm* – originally proposed by Collins (2002) for HMMs – for MLNs by approximating the expectations with the number of true groundings according to the most probable explanation. The most probable explanation is then given by the most probable assignment  $\mathbf{h}'$  to  $\mathbf{H}$  given the observed predicates  $\mathbf{O} = \mathbf{o}$  (Singla & Domingos, 2005) and can be obtained via computing the MAP solution given  $\mathbf{O} = \mathbf{o}$ . Hence, at each iteration  $t$  the weight  $w_k$  for a formula  $F_k$  is updated according to the difference of the true groundings in the training data and in the inferred MAP solution given  $\mathbf{O} = \mathbf{o}$ .

To avoid overfitting, the final weights are obtained by averaging over the weights from all iterations (Singla & Domingos, 2005). It has been shown that this strategy is superior over only taking the weight vector from the final iteration (Collins, 2002). By using an *online learning* approach, where the weights are updated on an instance-by-instance basis, weight learning is more efficient than in batch learning (Riedel, 2009, p. 35–36).

Besides the voted perceptron approach of Singla & Domingos (2005), other discriminative learning algorithms have been proposed. For instance, Lowd & Domingos (2007) propose to use a different learning rate for each weight  $w_k$ , while Huynh & Mooney (2009) and Huynh & Mooney (2011) propose a batch and an online maximum margin algorithm for MLNs. In this thesis, we do not explore these algorithms, but use the described voted perceptron algorithm, which has shown good results in practice (Singla & Domingos, 2005; Zirn et al., 2011) and is straightforward to extend with latent variables (Poon & Domingos, 2008) (Section 4.4.2).

### 4.2.5 Implementation

MLNs come with the advantage that several implementations are currently available. *Alchemy*<sup>3</sup> includes generative and discriminative learning algorithms, a structure learning algorithm and various approximate inference algorithms such as MaxWalkSat or belief propagation (Kok et al., 2005). *Tuffy*<sup>4</sup> implements an improved strategy for the grounding phase and uses a MaxWalkSat algorithm for MAP inference (Niu et al., 2011). It outperforms Alchemy in terms of speed at the same or even higher level of quality (Niu et al., 2011). In contrast to *Alchemy* and *Tuffy*, *thebeast*<sup>5</sup> implements an exact MAP inference technique that combines a cutting plane algorithm with integer linear programming (Riedel, 2008) (Section 4.2.3). It is written in Java and contains discriminative learning algorithms including a voted perceptron

<sup>3</sup><http://alchemy.cs.washington.edu/>, 15.5.2014.

<sup>4</sup><http://hazy.cs.wisc.edu/hazy/tuffy/>, 15.5.2014.

<sup>5</sup><https://code.google.com/p/thebeast/>, 15.5.2014.

algorithm. In *RockIt*<sup>6</sup>, a similar MAP inference strategy as in *thebeast* is implemented. By exploiting parallelization and local symmetries (Noessner et al., 2013), it outperforms *thebeast* and all other implementations in terms of speed. It also supports discriminative training.

All experiments in this thesis are based on the implementation in *thebeast*. At the time the experiments for this thesis were conducted, *thebeast* was the only implementation allowing for exact inference. *RockIt*, which is superior to all other implementations, was still under development at this time and did not yet support discriminative training.

### 4.3 Joint Concept Disambiguation and Clustering in Markov Logic

Our linguistic analysis of concept disambiguation and clustering (Section 2) shows that these tasks require a relational inference technique that allows to model interdependencies between instances and tasks. Markov logic allows to conveniently model such interdependencies and therefore suits our linguistically motivated requirements (Section 4.2).

Based on our linguistic analysis, we derived three implications concerning the modeling of the two tasks (Section 2.4.2): joint modeling of interrelated mentions (Implication 1), joint disambiguation and clustering of concepts (Implication 2), joint identification of the relevant context for a mention and its concept (Implication 3). In this and the next sections, we show for each of them how we implement it in Markov logic. First, we focus on the disambiguation task and discuss how we can disambiguate interrelated mentions dependent on each other (Implication 1) using Markov logic (Section 4.3.1). We then present how we model concept clustering in Markov logic (Section 4.3.2) and describe how the interdependencies between concept disambiguation and clustering (Implication 2) are implemented (Section 4.3.3). The interrelation between the relevant context of a mention and its denoted concepts (Implication 3) is discussed in Section 4.4.

#### 4.3.1 Modeling Concept Disambiguation

Concept disambiguation is the task of predicting for each mention the concept it denotes. The mentions are extracted from the texts to analyze, while the concepts are organized in an inventory which is in our case derived from Wikipedia. To model this task in Markov logic, we define a hidden predicate that takes as arguments a mention  $m$  and a concept  $c$ . This hidden predicate – we call it  $hasConcept(m, c)$  – defines the relation we aim to predict. It is only true, if the mention  $m$  denotes the concept  $c$  and is false in all other cases. For each mention  $m$

---

<sup>6</sup><http://code.google.com/p/rockit/>, 15.5.2014.

there is at most one ground predicate of type  $hasConcept(m, c)$  that it is true. We express this restriction using a hard cardinality constraint, i.e. a formula with infinite weight:

$$\forall m : \mathbf{if} \text{ mention}(m) \mathbf{then} |\{\forall c : \text{concept}(c) \wedge \text{hasConcept}(m, c)\}| \leq 1$$

The two observed predicates  $\text{mention}(m)$  and  $\text{concept}(c)$  ensure that the respective variable is of type mention and concept respectively. The assumption behind this cardinality constraint is that all ambiguities can be resolved and a mention only denotes one concept. This is a simplification, as mentions sometimes remain ambiguous even given their context. For instance, linguistic humor can be originated in such ambiguities. At the same time, it is also possible that no ground  $hasConcept$ -predicate is true for a mention. This does not indicate that the mention lacks a denoted concept, but that the mention is a *NIL*, i.e. its denoted concept is not part of the inventory and therefore unknown. By using this single formula, concept disambiguation and recognition of NILs is modeled jointly. In this formula and also in the following formulas, we use the notation **if...then**. We use this notation to account for a specificity of Markov logic how it is implemented in *thebeast*. First, the condition (part before **then**) is evaluated and only if this part is true, the second part (part after **then**) is added to the network. If the condition is false, the second part is not added to the network. As this cannot be expressed in first-order logic, we use the **if...then** notation.

Together with the hard cardinality constraint, the hidden predicate  $hasConcept(m, c)$  builds the core of the disambiguation system. To this core, formulas with learned weights can be added. We distinguish between *local* and *global* disambiguation formulas.

A local disambiguation formula only involves one instance, i.e. one mention. It consists of several observed predicates and the hidden predicate  $hasConcept$ . Typically it has the form

$$w \cdot q \quad \forall m, c : \mathbf{if} \text{ mention}(m) \wedge \text{concept}(c) \wedge \text{hasCandidateConcept}(m, c) \\ \wedge \text{localDisambigInformation}(m, c, q) \mathbf{then} \text{hasConcept}(m, c).$$

Note that this is only a template formula that illustrates the general form of a local formula. All predicates, except the  $hasConcept$  predicate are observed. The predicate  $hasCandidateConcept(m, c)$  is true if a concept  $c$  is a candidate concept of a mention  $m$  according to the lexicon (Section 3.2.1 and 6.2), while  $localDisambigInformation(m, c, q)$  is a placeholder for a single or several observed predicates that are specific for a formula. For instance, in a formula that injects prior knowledge about mention-concept frequencies into the model, this part is replaced by the observed predicate  $hasPriorProbability(m, c, q)$  where  $q$  is a score that expresses frequency information of the mention-concept pair. All local formulas – obtained by instantiating the template formula above with specific information – are explained in Chapter 5. Each local formula  $F_k$  is associated with a learned weight  $w$  and – if the formula represents a numeric feature – a score  $q$  which may depend on the formula, the mention and the concept

(Section 4.2.2).

A global formula involves more than one instance, i.e. at least two mentions, and allows to model interrelations between mentions. Whenever two mentions are part of the same ground formula they are disambiguated jointly, as the disambiguation decision for one mention influences the disambiguation decision for the other mention. How strong this influence is depends on the weight of the respective formula and on the score if applicable. From a linguistic point of view, mentions that are tied on the concept-level should be modeled jointly (Implication 1). Ideally, all mentions that are connected by concept-level cohesive ties are disambiguated jointly. However, to determine concept-level cohesive ties between mentions, the disambiguation decision of these mentions need to be known due to the reciprocal relationship between concept-level cohesive ties and disambiguation decisions (Implication 3). Modeling all these interactions jointly leads to a highly connected ground Markov network making the inference intractable or at least very slow.

To accommodate scalability, we propose a hybrid strategy to implement the notion of joint disambiguation of mentions. We start from the observation that concept-level cohesive ties partially correlate with cohesive ties on other linguistic levels such as on string or entity level (Section 2.2, Observations 2.2.3–2.2.4). To further analyze these correlations, we distinguished two types of concept-level cohesive ties: concept-level cohesive ties of type identity involve mentions that denote the same concept and concept-level cohesive ties of type relatedness occur between mentions that denote related or similar concepts (Section 2.2). Investigating correlations between concept-level cohesive ties and ties on other linguistic levels revealed that cohesive ties of type identity often correlate across linguistic levels (Observations 2.2.3–2.2.4). For instance, if two mentions are tied by a string-level tie of type identity, it is likely that they are also tied by an identity relation on the concept level. In contrast, correlations between cohesive types of type relatedness are weaker. If two mentions are related on the entity level it does not necessarily imply that they are also tied on the concept level (Observations 2.2.3–2.2.4). With respect to prediction of concept-level cohesive ties, these observations imply that concept-level cohesive ties of type identity are predictable with higher accuracy than concept-level cohesive ties of type relatedness. By exploiting for instance correlations with string-level ties, we can obtain strong features to determine if two mentions are tied on the concept level by an identity relation. However, it is more difficult to extract indicative features for mentions that exhibit concept-level cohesive ties of type relatedness as these ties are only weakly correlated with ties on other linguistic levels.

Our hybrid strategy takes into account these observations. The idea is to disambiguate mentions jointly if some features strongly indicate concept-level cohesive ties between them. If such strong features are missing we refrain from modeling them jointly. However, instead of completely disregarding such weak concept-level cohesive ties, we consider them using an iterative approach (Section 4.1.1). Given a mention to disambiguate, we use aggregated

relatedness scores between the candidate concept of this mention and some selected candidate concepts of other mentions (Section 5). This aggregation strategy has also led to good result in previous work which exclusively relies on it (e.g. Milne & Witten (2008b) and Ratinov & Roth (2011)). Although such an aggregation strategy is not genuinely joint, it at least approximates the notion of a joint decision. Hence, such a hybrid strategy that involves joint decisions and aggregative features reduces the connectivity in the ground Markov network compared to the fully joint approach, making the inference tractable.

Which mentions are modeled genuinely jointly depends on the features used. Given the observation that cohesive ties of type identity are easier to predict, we jointly disambiguate mentions that are tied by an identity relation. Identifying such identity relations between mentions corresponds to the task of concept clustering and is therefore modeled via concept clustering (Sections 4.3.2 and 4.3.3). However, in case we aim to propagate candidate concepts from one mention to another one, it is beneficial to directly model them jointly using the predicate *hasConcept* and state

$$(w \cdot q) \quad \forall m, n, c : \quad \mathbf{if} \quad \text{mention}(m) \wedge \text{mention}(n) \wedge \neg(m = n) \wedge \text{concept}(c) \\ \wedge \text{hasCandidateConcept}(m, c) \wedge \text{featureIdentityTie}(m, n, q) \\ \wedge \text{easierToDisambiguate}(m, n) \\ \mathbf{then} \quad \text{hasConcept}(m, c) \wedge \text{hasConcept}(n, c).$$

The placeholder *featureIdentityTie*( $m, n, q$ ) stands for a feature indicating that mention  $m$  and  $n$  are tied on the concept level by an identity relation. The score  $q$  measures the strength of this tie. The observed predicate *easierToDisambiguate*( $m, n$ ) expresses that  $m$  is easier to disambiguate than  $n$ . For instance, if the two mentions are two person names and  $m$  is the full name, while  $n$  is only the surname, it is easier to disambiguate  $m$  than  $n$ . At the same time, it often happens that the correct concept is not even under the candidate concepts for the mention  $n$ . In this case, we would like to move the candidates during inference from one mention to the other mention. This can be achieved by using a formula derived from the template formula above. It is only required that the concept  $c$  is a candidate concept of mention  $m$ . Even if it is not a candidate concept of mention  $n$  according to the lexicon, it can become one during the inference process. This approach shows that global formulas can also be used to propagate candidates from one instance to another one during inference. Such a propagation strategy is not only interesting in the context of concept disambiguation, but can be adapted for other tasks. For instance, it has been successfully applied to propagate candidate antecedents in the bridging resolution task (Hou et al., 2013).

**Backbone for Concept Disambiguation.** Table 4.4 summarizes the backbone of our disambiguation approach in Markov logic. As in the description above, a template notation is

---

**Predicates**


---

## Hidden Predicate

 $p_{d1} \quad \text{hasConcept}(m, c)$ 


---

## Observed Predicates

 $p_{d2} \quad \text{mention}(m)$ 
 $p_{d3} \quad \text{concept}(c)$ 
 $p_{d4} \quad \text{hasCandidateConcept}(m, c)$ 
 $p_{d5} \quad \text{easierToDisambiguate}(m, n)$ 


---

## Predicate Templates for Disambiguation

 $t_{pd1} \quad \text{featureDisambiguation}(m, c, q)$ 
 $t_{pd2} \quad \text{featureIdentityTie}(m, n, q)$ 


---

**Formulas**


---

## Hard Constraints

 $f_{d1} \quad \forall m : \mathbf{if} \text{ mention}(m) \mathbf{then} |\{\forall c : \text{concept}(c) \wedge \text{hasConcept}(m, c)\}| \leq 1$ 


---

## Template Formulas with Learned Weights

 $t_{fd1} \quad (w \cdot q) \quad \forall m, c : \quad \mathbf{if} \text{ mention}(m) \wedge \text{concept}(c) \wedge \text{hasCandidateConcept}(m, c) \\ \wedge \text{localDisambigInformation}(m, c, q) \mathbf{then} \text{hasConcept}(m, c)$ 
 $t_{fd2} \quad (w \cdot q) \quad \forall m, n, c : \quad \mathbf{if} \text{ mention}(m) \wedge \text{mention}(n) \wedge \neg(m = n) \wedge \text{concept}(c) \\ \wedge \text{hasCandidateConcept}(m, c) \wedge \text{featureIdentityTie}(m, n, q) \\ \wedge \text{easierToDisambiguate}(m, n) \\ \mathbf{then} \text{hasConcept}(m, c) \wedge \text{hasConcept}(n, c)$ 

Table 4.4: Concept disambiguation in Markov logic.  $m, n$  are mentions,  $c$  is a concept,  $q$  is a score and  $w$  is a weight.

used for the formulas with learned weights. These templates indicate the general structure of the formulas and predicates and are implemented by the features discussed in Chapter 5.

Compared to a standard disambiguation method using a binary classifier (Milne & Witten, 2008b) or a ranker (Dredze et al., 2010), our approach has several advantages.

First, it allows us to jointly disambiguate the mentions and recognize the NILs. In the standard approach that formulates the disambiguation as a ranking problem, NILs are recognized based on a separately tuned threshold or by using a separate binary classifier. In both cases, an additional optimization step is required leading to a two-step approach and hence to potential error propagation.

A second advantage of our approach is that candidate concepts can be propagated from one mention to another mention during the inference using global formulas. It is often difficult to balance between *coverage* (the correct concept is among the candidate concepts for a mention) and *average ambiguity* (average number candidate concepts per mentions). If the average ambiguity is higher (e.g. 70 candidate concepts per mentions), the disambiguation becomes more difficult. By propagating candidate concepts from one mention to another mention during the inference, we can afford a lower average ambiguity at the same level of coverage. The assumption is that a highly ambiguous mention might be tied to a less ambiguous one by a concept-level identity relation.

Third, Markov logic allows us to incorporate the notion of joint disambiguation not only by aggregation, but also by directly model interdependencies between mentions. Our hybrid approach combines both strategies and models these mentions jointly for which reliable features indicate a concept-level cohesive tie. The restriction to only disambiguate a few selected mentions jointly hints that joint inference comes at the price of efficiency. If a model contains too many interdependencies, inference becomes intractable or slow. The hybrid strategy helps us to balance between expressiveness and time efficiency. However, compared to a standard classifier or ranker, the proposed approach is still less time efficient.

### 4.3.2 Modeling Concept Clustering in Markov Logic

The aim of concept disambiguation is to resolve lexical ambiguities of mentions. Concept clustering offers another perspective on the same task. Instead of linking mention to an inventory as in concept disambiguation, mentions are clustered so that all mentions in one cluster denote the same concept. Compared to concept clustering, concept disambiguation has the advantage that – by linking mentions to concepts in an inventory – all information that is associated with the linked concepts in the inventory becomes accessible for downstream tasks. In contrast, concept clustering is not limited to resolve ambiguities of mentions that denote a concept in a predefined inventory, but is applicable to all mentions. In previous work, concept disambiguation and clustering have been investigated as two cascaded steps. Some work first clusters words to obtain concept clusters that are used in a second step as an automatically generated inventory for concept disambiguation (Schütze, 1998). Other approaches – mainly in the context of the shared task TAC – first disambiguate the mentions using an inventory and then only cluster all NILs (Ji & Grishman, 2011). In this case, concept clustering is a fall-back strategy for NILs.

In this thesis, we argue that concept disambiguation and clustering are two perspectives on the same problem and can mutually help each other (Implication II). To exploit their connection, both tasks need to be modeled in the same framework, which is Markov logic in our case. While we show in this section how we can model concept clustering in Markov logic, we

focus in the next section how the two perspectives, i.e. concept disambiguation and clustering, can be combined.

To model concept clustering in Markov logic, we introduce the hidden predicate *hasSameConcept*( $m, n$ ) which defines a relation between two mentions. It is true if the two mentions denote the same concept and it is false in all other cases. This predicate exerts a pairwise view on the clustering task. In the context of coreference resolution, i.e. clustering on the level of entities, such pairwise approaches have been successful and serve as a strong baseline (Soon et al., 2001). However, Markov logic allows to exploit both symmetric and transitive relations. We can define these properties using hard constraints. By defining the *hasSameConcept*-predicate as symmetric via

$$\forall m, n : \text{ if } \text{mention}(m) \wedge \text{mention}(n) \wedge \neg(m = n) \text{ then } \text{hasSameConcept}(m, n) \\ \rightarrow \text{hasSameConcept}(n, m)$$

we enforce that if *hasSameConcept*( $m, n$ ) is true also *hasSameConcept*( $n, m$ ) is true. To establish transitivity, we introduce the hard constraint

$$\forall m, n, l : \text{ if } \text{mention}(m) \wedge \text{mention}(n) \wedge \text{mention}(l) \\ \wedge \neg(m = n) \wedge \neg(m = l) \wedge \neg(n = l) \\ \text{ then } \text{hasSameConcept}(m, n) \wedge \text{hasSameConcept}(n, l) \\ \rightarrow \text{hasSameConcept}(m, l).$$

By defining the *hasSameConcept*-predicate as transitive, we ensure that, given that the mention pairs ( $m, n$ ) and ( $n, l$ ) denote the same concept respectively, also  $m$  and  $l$  are in a *hasSameConcept*-relation. In a pairwise model, transitivity is enforced in a post-processing step and exerts no influence on the pairwise clustering decisions. In such a two-step approach the information flow is suboptimal, as transitive clustering relations that affect the pairwise decisions are not considered during inference. This contrasts with how transitivity is modeled in the proposed approach. Markov logic allows to inject transitive information into the inference process. Hence, with the hard constraint above transitive information is available and leveraged during the inference.

Based on these two constraints, formulas with learned weights are added. These formulas are all of the form

$$(w \cdot q) \quad \forall m, n, c : \text{ if } \text{mention}(m) \wedge \text{mention}(n) \wedge \neg(m = n) \\ \wedge \text{concept}(c) \wedge \text{hasCandidateConcept}(m, c) \\ \wedge \text{featureClustering}(m, n, q) \\ \text{ then } \text{hasSameConcept}(m, n),$$

where *featureClustering* is a placeholder for an observed predicate. This observed predi-

**Predicates**

Hidden Predicate

 $p_{c1} \quad hasSameConcept(m, n)$ 

Observed Predicates

 $p_{c2} \quad mention(m)$ 

Predicate Templates for Clustering

 $t_{pc1} \quad featureClustering(m, c, q)$ **Formulas**

Hard Constraints

 $f_{c1} \quad \forall m, n : \text{if } mention(m) \wedge mention(n) \wedge \neg(m = n) \text{ then } hasSameConcept(m, n) \rightarrow hasSameConcept(n, m)$  $f_{c2} \quad \forall m, n, l : \text{if } mention(m) \wedge mention(n) \wedge mention(l) \wedge \neg(m = n) \wedge \neg(m = l) \wedge \neg(n = l) \text{ then } hasSameConcept(m, n) \wedge hasSameConcept(n, l) \rightarrow hasSameConcept(m, l)$ 

Template Formulas with Learned Weights

 $t_{fc1} \quad (w \cdot q) \quad \forall m, n, c : \text{if } mention(m) \wedge mention(n) \wedge \neg(m = n) \wedge concept(c) \wedge hasCandidateConcept(m, c) \wedge featureClustering(m, n, q) \text{ then } hasSameConcept(m, n)$ Table 4.5: Concept clustering in Markov logic.  $m, n, l$  are mentions,  $c$  is a concept,  $q$  is a score and  $w$  is a weight.

cate encodes a feature that indicates if two mentions are in the same concept cluster. Such a feature can be binary or numeric.

**Backbone for Concept Clustering.** In Table 4.5, the predicates and formulas building the core of the proposed concept clustering approach are listed. The template formula  $t_{fc1}$  with the template predicate  $t_{pc1}$  shows how formulas with learned weights are structured. We discuss them in Chapter 5.

To leverage the inter-dependencies between concept disambiguation and clustering (Section 4.3.3), we model both tasks in the same framework, i.e. Markov logic. Markov logic is suitable to model concept clustering, as it allows to account for symmetry and transitivity with hard constraints. In this way, transitivity information is available and leveraged at inference

time.

Introducing transitivity raises the connectivity in the ground Markov network making the inference more complex. Within-document concept clustering and cross-document concept clustering are feasible as long as the data set consists of a few hundreds of texts. However, if the corpus to be processed consists of thousands or even millions of text documents scaling is an issue. In this case, the corpus may need to be split up into different parts and the inference needs to be distributed. Issues related to scaling up the approach to thousands or millions of texts are not addressed in the context of this thesis. For more information on this topic, the reader is referred to Rao et al. (2010) and Singh et al. (2011).

### 4.3.3 Modeling Joint Disambiguation and Clustering in Markov Logic

One of the key contributions of this thesis is to define and model concept disambiguation and clustering as two interrelated tasks that can mutually help each other (Implication II). While the linguistic relationship between these two tasks is discussed in Section 2, this section shows how we model these two tasks jointly with Markov logic. Our joint approach is based on our concept disambiguation and clustering model proposed in Section 4.3.1 and 4.3.2.

Instead of modeling disambiguation and clustering as two cascaded tasks as in previous work (see Ji & Grishman (2011) for an overview), we approach them jointly to leverage the clustering decisions during disambiguation and the disambiguation decisions during clustering. This work is in line with research that assumes that a concept inventory is available. The selection of the inventory (Chapter 3) is irrelevant for the core of our approach (discussed in this section) and mainly affects some features (Chapter 5). With respect to the core of our approach, every inventory can be chosen, even an inventory that has been automatically built.

In previous cascaded disambiguation and clustering approaches requiring a pre-defined inventory, the clustering is restricted to the NILs, i.e. mentions with no corresponding concept in the respective inventory. By clustering NILs, novel concepts are identified that might be integrated in the inventory. In contrast, we cluster all the mentions and argue that clustering is not only useful to obtain new previously unknown concepts (NIL clustering), but also to support the disambiguation decision of mentions that denote known concepts. Following this argumentation, we further state that concept clustering is relevant, even if only the disambiguation decisions are of interest. At the same time, we also claim that disambiguation is relevant, even if only the outcome of the NIL clustering is of interest. Clustering NILs presupposes the recognition of NILs which in turn is closely tied to concept disambiguation.

In order to jointly model two arbitrary tasks in Markov logic, two pieces of information are required: (1) the exact relationships between the two tasks and (2) the validity of these relationships. While the relationships define the formulas that link the two tasks, the validity determines the weight of these formulas. If the formulas are always valid, we can combine the

tasks with hard constraints that have to be fulfilled. Otherwise if the formulas that tie the tasks are only sometimes true, soft constraints with manually set or learned weights are required accounting for the uncertainty.

In the case of concept disambiguation and clustering, the relationship can be specified as a mutual dependency: if two mentions are clustered, both of them should either be disambiguated to the same concept or recognized as NILs; if two mentions are disambiguated to the same concept, they should be resolved to the same cluster. Both relations are always valid. We therefore can link the two tasks via hard constraints. The following hard constraint defines that if two mentions  $m$  and  $n$  are in the same cluster and mention  $m$  denotes concept  $c$ , mention  $n$  also denotes concept  $c$ :

$$\begin{aligned} \forall m, n, c : \quad & \mathbf{if} \text{ mention}(m) \wedge \text{mention}(n) \wedge \text{concept}(c) \wedge \neg(m = n) \\ & \mathbf{then} \text{ hasSameConcept}(m, n) \wedge \text{hasConcept}(m, c) \\ & \rightarrow \text{hasConcept}(n, c) \end{aligned}$$

Due to the fact that all possible combinations are instantiated, the formula further guarantees that if mention  $m$  is a NIL, mention  $n$  is also a NIL, given they are part of the same cluster.

The following hard constraint states that two mentions that denote the same concept, are also in the same cluster:

$$\begin{aligned} \forall m, n, c : \quad & \mathbf{if} \text{ mention}(m) \wedge \text{mention}(n) \wedge \text{concept}(c) \wedge \neg(m = n) \\ & \mathbf{then} \text{ hasConcept}(m, c) \wedge \text{hasConcept}(n, c) \\ & \rightarrow \text{hasSameConcept}(m, n) \end{aligned}$$

These two formulas are sufficient to link the two tasks assuming that they are defined as in Section 4.3.1 and 4.3.2.

**Backbone for Joint Concept Disambiguation and Clustering.** All formulas that build together the core of our joint concept disambiguation and clustering approach are listed in Table 4.6. This includes (1) the formulas used for disambiguation, (2) the formulas used for clustering and (3) the formulas that link the two tasks.

#### 4.3.4 Discussion

The proposed concept disambiguation and clustering approach is joint in four ways: (1) the disambiguation and the recognition of NILs is done jointly (Formula  $f_{d1}$ ); (2) mentions are at least partially disambiguated jointly (features of the form  $t_{fd2}$  and through clustering features); (3) the clustering decisions for different mentions are made jointly due to the symmetry and transitivity (Formula  $f_{c1}$ ,  $f_{c2}$ ); (4) the disambiguation and the clustering are modeled jointly

---

**Predicates**


---

## Hidden Predicate

 $p_{d1}$   $hasConcept(m, c)$  $p_{c1}$   $hasSameConcept(m, n)$ 


---

Observed Predicates

---

 $p_{d2}$   $mention(m)$  $p_{d3}$   $concept(c)$  $p_{d4}$   $hasCandidateConcept(m, c)$  $p_{d5}$   $easierToDisambiguate(m, n)$ 


---

Predicate Templates for Joint Disambiguation and Clustering

---

 $t_{pd1}$   $featureDisambiguation(m, c, q)$  $t_{pd2}$   $featureIdentityTie(m, n, q)$  $t_{pc1}$   $featureClustering(m, c, q)$ 


---

**Formulas**


---

## Hard Constraints

 $f_{d1}$   $\forall m : \mathbf{if} \text{ mention}(m) \mathbf{then} |\{\forall c : \text{concept}(c) \wedge \text{hasConcept}(m, c)\}| \leq 1$  $f_{c1}$   $\forall m, n : \mathbf{if} \text{ mention}(m) \wedge \text{mention}(n) \wedge \neg(m = n) \mathbf{then} \text{hasSameConcept}(m, n) \rightarrow \text{hasSameConcept}(n, m)$  $f_{c2}$   $\forall m, n, l : \mathbf{if} \text{ mention}(m) \wedge \text{mention}(n) \wedge \text{mention}(l) \wedge \neg(m = n) \wedge \neg(m = l) \wedge \neg(n = l) \mathbf{then} \text{hasSameConcept}(m, n) \wedge \text{hasSameConcept}(n, l) \rightarrow \text{hasSameConcept}(m, l)$  $f_{j1}$   $\forall m, n, c : \mathbf{if} \text{ mention}(m) \wedge \text{mention}(n) \wedge \text{concept}(c) \wedge \neg(m = n) \mathbf{then} \text{hasSameConcept}(m, n) \wedge \text{hasConcept}(m, c) \rightarrow \text{hasConcept}(n, c)$  $f_{j2}$   $\forall m, n, c : \mathbf{if} \text{ mention}(m) \wedge \text{mention}(n) \wedge \text{concept}(c) \wedge \neg(m = n) \mathbf{then} \text{hasConcept}(m, c) \wedge \text{hasConcept}(n, c) \rightarrow \text{hasSameConcept}(m, n)$ 


---

Template Formulas with Learned Weights

---

 $t_{fd1}$   $(w \cdot q)$   $\forall m, c : \mathbf{if} \text{ mention}(m) \wedge \text{concept}(c) \wedge \text{hasCandidateConcept}(m, c) \wedge \text{localDisambigInformation}(m, c, q) \mathbf{then} \text{hasConcept}(m, c)$  $t_{fd2}$   $(w \cdot q)$   $\forall m, n, c : \mathbf{if} \text{ mention}(m) \wedge \text{mention}(n) \wedge \neg(m = n) \wedge \text{concept}(c) \wedge \text{hasCandidateConcept}(m, c) \wedge \text{featureIdentityTie}(m, n, q) \wedge \text{easierToDisambiguate}(m, n) \mathbf{then} \text{hasConcept}(m, c) \wedge \text{hasConcept}(n, c)$  $t_{fc1}$   $(w \cdot q)$   $\forall m, n, c : \mathbf{if} \text{ mention}(m) \wedge \text{mention}(n) \wedge \neg(m = n) \wedge \text{concept}(c) \wedge \text{hasCandidateConcept}(m, c) \wedge \text{featureClustering}(m, n, q) \mathbf{then} \text{hasSameConcept}(m, n)$ 

Table 4.6: Joint concept disambiguation and clustering in Markov logic.  $m, n, l$  are mentions,  $c$  is a concept,  $q$  is a score and  $w$  is a weight.

(Formula  $f_{j1}, f_{j2}$ ).

This core accounts for the first two linguistically motivated implications. Some selected mentions, for which strong features indicate that they are cohesively tied, are modeled jointly and disambiguation and clustering decisions are simultaneously taken. From a discourse perspective, the clustering models concept-level cohesive ties of type identity. By linking the two tasks, all mention pairs that form candidate arguments for such an identity relation are disambiguated jointly. If two mentions are candidate arguments for an identity relation depends on the clustering ( $t_{fc1}$ ), the disambiguation features ( $t_{fd2}$ ) and the transitivity.

The following section shows how we can introduce more discourse information into the model and account for the third linguistically motivated principle, i.e. joint context selection and disambiguation.

## 4.4 Integrating Discourse Information

Concept-level cohesive ties and the concepts denoted by mentions are mutually dependent on each other. As concept-level cohesive ties determine which context is relevant to disambiguate a mention, they are highly relevant to define the disambiguation context of a mention (Section 2.2.2). However, as the identification of cohesive ties and the disambiguation of mentions are mutually dependent on each other, it is difficult to integrate context selection into a disambiguation model without complicating it so much that inference becomes infeasible. In this section, we propose a feasible approach that models context selection and disambiguation jointly (Implication III).

Instead of determining the context that is relevant to disambiguate a mention for each mention individually, we propose a binwise context selection approach with three different context definitions (Section 2.3.1). For each mention, we determine its *cohesive scope*. The *cohesive scope* of a mention is the text span within which a concept denoted by a mention shows concept-level cohesive ties. We distinguish three broad categories of cohesive scopes: (1) Mentions with *local cohesive scope* exhibit cohesive ties with lexical units in the same sentence; (2) mentions with *intermediate cohesive scope* show cohesive ties both within the sentence and beyond; (3) mentions with *global cohesive scope* form cohesive ties with mentions across sentence boundaries. The notion of scope is a means to define the appropriate context to disambiguate a mention. A mention of local scope does not exhibit relations with lexical units outside its sentence. The global context therefore does not help to disambiguate it or can even lead to the wrong disambiguation. For a mention with global scope, the global context is crucial, while the local context is not discriminative or even misleading. For a mention with intermediate scope both the local and the global context are relevant. Hence, while the scope influences the appropriate disambiguation context, the disambiguation of a mention

influences its scope.

Given a set of features for disambiguation, we aim to weight them differently depending on the scope. To model the reciprocal relationship between scope assignment and disambiguation, we propose a latent variable model in the framework of Markov logic. It allows us to learn the parameters for the scope assignment and the disambiguation tasks jointly and enables us to perform joint inference.

Our approach is *joint* as we simultaneously predict the scope  $s$  and the concept  $c$  of mention  $m$ . As during learning training data is available for the disambiguation and clustering tasks but not for the scope assignment task, we face a problem with *latent variables*. Latent variables represent missing information in the input or a part of the output which is not relevant except for supporting the prediction of the target (Smith, 2011). In our approach, the different cohesive scopes are modeled by latent variables. Each mention to be disambiguated is assigned a scope  $s$ . All feature weights are parameterized by scope  $s$ . The parameters for the disambiguation and scope assignment tasks are learned jointly and are guided by the annotations available for the disambiguation task.

To implement this scope-aware approach two parts of the model proposed in Section 4.3.3 need to be adapted: the model is extended with additional predicates and formulas to integrate scope information (Section 4.4.1) and the learning process is adapted to account for latent variables (Section 4.4.2).

#### 4.4.1 A Scope-aware Approach

The purpose of assigning a scope  $s$  to each mention  $m$  is to learn scope-specific weights for disambiguation to account for heterogeneous scopes of mentions. The learned weights are parameterized by scopes. We indicate this parameterization of learned weights by  $w(s)$ .

To extend our approach for scopes, we define a hidden predicate  $hasScope(m, s)$  that describes a relation between a mention  $m$  and a scope  $s$ . Each mention is assigned exactly one scope enforced via the hard cardinality constraint

$$\forall m : \text{if } mention(m) \text{ then } |\{\forall s : scope(s) : hasScope(m, s)\}| = 1.$$

The observed predicate  $scope(s)$  ensures that the variable  $s$  is of type scope. To make the disambiguation dependent on the mentions' scopes, we could now simply adapt all formulas with learned weights by adding the  $hasScope$  predicates by conjunction and make the weight dependent on it, e.g.

$$\begin{aligned}
w(s) \cdot q \quad \forall m, c, s : & \text{ if } \textit{mention}(m) \wedge \textit{concept}(c) \wedge \textit{hasCandidateConcept}(m, c) \\
& \wedge \textit{scope}(s) \\
& \wedge \textit{localDisambigInformation}(m, c, q) \\
& \text{ then } \textit{hasConcept}(m, c) \wedge \\
& \wedge \textit{hasScope}(m, s).
\end{aligned}$$

All parts that are new are highlighted in blue. As this enriched template formula shows, all formulas with learned weights would become more complicated. In this example, we would have an additional hidden predicate for each formula. This decreases the efficiency of the inference massively. We therefore propose a more efficient approach by introducing an additional hidden predicate *relatesScopeToConcept*(*m, c, s*). The purpose of this predicate is to relate the scope assignment task to the disambiguation task. The hidden predicate *relatesScopeToConcept*(*m, c, s*) expresses a relation between a mention *m*, a concept *c* and a scope *s*. It is true if the mention *m* is assigned the concept *c* and the scope *s* and false in all other cases. Instead of using an inefficient formula of the form above – requiring two hidden predicates –, we can now express the same relations more efficiently with only one hidden predicate, i.e.

$$\begin{aligned}
w(s) \cdot q \quad \forall m, c, s : & \text{ if } \textit{mention}(m) \wedge \textit{concept}(c) \wedge \textit{hasCandidateConcept}(m, c) \\
& \wedge \textit{scope}(s) \\
& \wedge \textit{localDisambigInformation}(m, c, q) \\
& \text{ then } \textit{relatesScopeToConcept}(m, c, s).
\end{aligned}$$

Hence, the formulas for concept disambiguation that are associated with a learned weight are now defined with respect to the hidden predicates *relatesScopeToConcept*. Instead of learning one single weight for each formula, each formula is associated with several learned weights: for each scope or each scope combination a separate weight is learned making the weight of a formula scope-dependent.

To guarantee that the model is meaningful and consistent, the *relatesScopeToConcept* predicate needs to be further specified and linked to the other hidden predicates.

As each mention is assigned at most one concept and exactly one scope, we restrict that for each mention at most one ground predicate of type *relatesScopeToConcept* can be true via the hard cardinality constraint

$$\forall m : \text{ if } \textit{mention}(m) \text{ then } |\{\forall c, s : \textit{concept}(c) \wedge \textit{scope}(s) : \textit{relatesScopeToConcept}(m, c, s)\}| \leq 1.$$

We further ensure accordance between the ground predicate *hasConcept* and the ground

predicate *relatesConceptToScope* via some hard constraints. By stating

$$\forall m, c, s : \text{ if } \text{mention}(m) \wedge \text{concept}(c) \wedge \text{scope}(s) \\ \text{ then } \text{relatesScopeToConcept}(m, c, s) \rightarrow \text{hasConcept}(m, c)$$

we enforce that whenever a mention is assigned a concept according to the ground predicate *relatesConceptToScope*, the corresponding ground predicate *hasConcept* – i.e. the ground predicate *hasConcept* that links this concept to the mention – must be true. The opposite is established via

$$\forall m, c : \text{ if } \text{mention}(m) \wedge \text{concept}(c) \text{ then } \text{hasConcept}(m, c) \\ \rightarrow (|\{\forall s : \text{scope}(s) \wedge \text{relatesScopeToConcept}(m, c, s)\}| = 1)$$

and

$$\forall m, c, s : \text{ if } \text{mention}(m) \wedge \text{concept}(c) \wedge \text{scope}(s) \text{ then } \text{hasConcept}(m, c) \\ \wedge \text{hasScope}(m, s) \rightarrow \text{relatesScopeToConcept}(m, c, s).$$

If no concept is assigned to a mention according to one of the hidden predicates, the mention must also be a NIL according to the other hidden predicate. Together with

$$\forall m, c, s : \text{ if } \text{mention}(m) \wedge \text{concept}(c) \wedge \text{scope}(s) \\ \text{ then } \text{relatesScopeToConcept}(m, c, s) \rightarrow \text{hasScope}(m, s),$$

the last hard constraint guarantees the agreement between the ground predicates *relatesScopeToConcept* and *hasScope*.

The features for scope assignment task (e.g. string-level text structure features) take the form

$$w(s) \cdot q \quad \forall m, s : \text{ if } \text{mention}(m) \wedge \text{scope}(s) \wedge \text{scopeFeature}(m, s, q) \\ \text{ then } \text{hasScope}(m, s),$$

The score  $q$  can also be binary.

**Backbone of the Scope-aware Concept Disambiguation Approach.** Table 4.7 shows the core of our scope-aware concept disambiguation approach. It contains all hidden predicates, some basic observed predicates and all hard constraints. In addition, it lists some template predicates and formulas with learned scope-specific weights. Together with the formulas and templates for the clustering (Table 4.6), these formulas form the core of our joint scope-aware concept disambiguation and clustering approach.

**Predicates**

## Hidden Predicates

- $p_{d1}$   $hasConcept(m, c)$   
 $p_{s1}$   $hasScope(m, s)$   
 $p_{sd1}$   $relatesScopeToConcept(m, c, s)$

## Observed Predicates

- $p_{d2}$   $mention(m)$   
 $p_{d3}$   $concept(c)$   
 $p_{d4}$   $hasCandidateConcept(m, c)$

## Predicate Templates

- $t_{pd1}$   $featureDisambiguation(m, c, q)$   
 $t_{ps1}$   $featureScope(m, s, q)$

**Formulas**

## Hard Cardinality Constraints

- $f_{d1}$   $\forall m : \mathbf{if} \text{ mention}(m) \mathbf{then} |\{\forall c : \text{concept}(c) \wedge \text{hasConcept}(m, c)\}| \leq 1$   
 $f_{s1}$   $\forall m : \mathbf{if} \text{ mention}(m) \mathbf{then} |\{\forall s : \text{scope}(s) \wedge \text{hasScope}(m, s)\}| = 1$   
 $f_{sd1}$   $\forall m : \mathbf{if} \text{ mention}(m) \mathbf{then} |\{\forall c, s : \text{concept}(c) \wedge \text{scope}(s) : \text{relatesScopeToConcept}(m, c, s)\}| \leq 1$

## Hard Constraints

- $f_{sd2}$   $\forall m, c, s : \mathbf{if} \text{ mention}(m) \wedge \text{concept}(c) \wedge \text{scope}(s)$   
 $\mathbf{then} \text{relatesScopeToConcept}(m, c, s) \rightarrow \text{hasConcept}(m, c)$   
 $f_{sd3}$   $\forall m, n : \mathbf{if} \text{ mention}(m) \wedge \text{concept}(c) \mathbf{then} \text{hasConcept}(m, c)$   
 $\rightarrow (|\{\forall s : \text{scope}(s) \wedge \text{relatesScopeToConcept}(m, c, s)\}| = 1)$   
 $f_{sd4}$   $\forall m, c, s : \mathbf{if} \text{ mention}(m) \wedge \text{concept}(c) \wedge \text{scope}(s) \mathbf{then} \text{hasConcept}(m, c)$   
 $\wedge \text{hasScope}(m, s) \rightarrow \text{relatesScopeToConcept}(m, c, s)$   
 $f_{sd5}$   $\forall m, c, s : \mathbf{if} \text{ mention}(m) \wedge \text{concept}(c) \wedge \text{scope}(s)$   
 $\mathbf{then} \text{relatesScopeToConcept}(m, c, s) \rightarrow \text{hasScope}(m, s)$

## Template Formulas with Learned Weights

- $t_{fd1}$   $w(s) \cdot q \quad \forall m, c, s : \mathbf{if} \text{ mention}(m) \wedge \text{concept}(c) \wedge \text{scope}(s)$   
 $\wedge \text{hasCandidateConcept}(m, c)$   
 $\wedge \text{featureDisambiguation}(m, c, q)$   
 $\mathbf{then} \text{relatesScopeToConcept}(m, c, s)$   
 $t_{fs1}$   $w(s) \cdot q \quad \forall m, s : \mathbf{if} \text{ mention}(m) \wedge \text{scope}(s) \wedge \text{scopeFeature}(m, s, q)$   
 $\mathbf{then} \text{hasScope}(m, s)$

Table 4.7: Scope-aware concept disambiguation in Markov logic.  $m$  is a mention,  $c$  is a concept,  $q$  is a score,  $s$  is a scope.

### 4.4.2 Learning with Latent Variables

The voted perceptron algorithm described in Section 4.2.4 requires annotated training data. Since no annotated training data is available for the scope assignment task, but only for the concept disambiguation and clustering task, the learning algorithm needs to be adapted accordingly.

If the assignment of scopes was completely decoupled from the disambiguation and clustering of concepts, the model parameters for the scope assignment could only be learned in an unsupervised way. However, as the scope assignment task and the disambiguation task are highly interrelated, we can benefit from these interdependencies and guide the weight learning for the scope assignment task by the training data available for the concept disambiguation and clustering task. Our actual target prediction tasks are concept disambiguation and clustering, whereas assigning scopes is a supportive task. We model the scopes as latent variables and optimize the model parameters – i.e. the weights – of the scope assignment to maximize its benefit for the disambiguation. From a machine learning perspective, the scopes can be considered as parts of the input that need to be predicted because they are missing or latent (see Smith (2011), p. 140). Introducing them into our model means to augment the model with additional latent variables to increase its expressiveness.

**Related Work.** The idea of augmenting a model with additional latent variables is known as *hidden* or *latent variable learning* (Smith, 2011) and is a promising research direction with successful applications in for example syntactic parsing (Petrov et al., 2006), statistical machine translation (Blunsom et al., 2008), sentiment analysis (Yessenalina et al., 2010; Trivedi & Eisenstein, 2013) and question answering (Wang et al., 2007). For latent variable learning, generative approaches (Chiang & Bikel, 2002; Matsuzaki et al., 2005; Prescher, 2005; Petrov et al., 2006), large margin methods (Smith, 2011) and conditional log-linear models (Wang et al., 2007; Sutton et al., 2007, *inter alia*) have been proposed. As we use a conditional log-linear model, we focus here on such approaches. Blunsom et al. (2008) for instance use latent variables in the context of discriminative machine translation. Koo & Collins (2005) propose to use a conditional log-linear model for reranking with latent variables and evaluate it in the context of parse reranking. Each word in a sentence is assigned to a latent sense cluster driven by the reranking task. Quattoni et al. (2007) augment CRFs with latent states and evaluate their method on the task of object detection and gesture recognition. The closest work to ours is Poon & Domingos (2008). They use latent variables for coreference resolution in the framework of Markov logic. Their aim is to predict the mentions' head, type, number and gender jointly together with the coreference relations. While they assume that some training data are available for the head and partially the type, number and gender prediction tasks, they refrain from using any training data for the coreference resolution task and model the

coreference relations as latent variables. Hence, they benefit from supportive tasks to model the target prediction task latent, whereas we model the supportive task with latent variables.

**Learning Algorithm.** Log-linear models are straightforward to be extended with latent variables. Following previous work (e.g. Poon & Domingos (2008)), we split our hidden predicates into two parts:  $\mathbf{H}$  are the ones for which the ground truth is known during training (concepts and clusters) and  $\mathbf{L}$  are the ones for which no annotated training data is available (scopes). Let  $\mathbf{O}$  be the observed predicates. Let  $\mathbf{o}$  and  $\mathbf{h}$  be the values of  $\mathbf{O}$  and  $\mathbf{H}$  in the training data.  $\mathbf{l}$  denotes values assigned to  $\mathbf{L}$ . In our case, weight learning finds a  $w$  that maximizes the conditional log-likelihood. If we only consider  $\mathbf{O}$  and  $\mathbf{H}$  as in Section 4.2.4, the conditional log-likelihood is given by

$$Q_w = \log P_w(\mathbf{H} = \mathbf{h} | \mathbf{O} = \mathbf{o}).$$

By definition, we can incorporate additional variables into the conditional log-likelihood by marginalizing over them. To incorporate  $\mathbf{L}$  we therefore sum over all possible values of  $\mathbf{L}$  leading to

$$\begin{aligned} Q_w &= \log P_w(\mathbf{H} = \mathbf{h} | \mathbf{O} = \mathbf{o}) \\ &= \log \sum_{\mathbf{l}} P_w(\mathbf{H} = \mathbf{h}, \mathbf{L} = \mathbf{l} | \mathbf{O} = \mathbf{o}). \end{aligned}$$

The gradient  $\nabla Q_w$  – required by the gradient ascent method we applied – is then given by (Poon & Domingos, 2008)

$$\begin{aligned} \nabla Q_w &= \frac{\partial}{\partial w_k} \log \sum_u P_w(\mathbf{H} = \mathbf{h}, \mathbf{L} = \mathbf{l} | \mathbf{O} = \mathbf{o}) \\ &= E_w[n_k(\mathbf{L} | \mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o})] - E_w[n_k(\mathbf{L}, \mathbf{H} | \mathbf{O} = \mathbf{o})]. \end{aligned}$$

$E_w$  denotes the expectation according to  $P_w$  given the current weight vector  $w$ .  $E_w[n_k(\mathbf{L} | \mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o})]$  and  $E_w[n_k(\mathbf{L}, \mathbf{H} | \mathbf{O} = \mathbf{o})]$  are the expected number of true groundings of formula  $f_k$  specified by  $(\mathbf{L} | \mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o})$  and  $(\mathbf{L}, \mathbf{H} | \mathbf{O} = \mathbf{o})$  respectively. As before, we use a voted perceptron (Lowd & Domingos, 2007; Poon & Domingos, 2008) which approximates the expectations via computing the MAP solutions (Section 4.2.3). To compute  $E_w[n_k(\mathbf{L} | \mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o})]$ , we calculate the MAP solution given  $\mathbf{H}$  and  $\mathbf{O}$  fixed to the true groundings according the gold data.  $E_w[n_k(\mathbf{L}, \mathbf{H} | \mathbf{O} = \mathbf{o})]$  is approximated by the MAP solution with  $\mathbf{O}$  fixed to  $\mathbf{o}$ .

In contrast to the supervised case without any latent variables (Section 4.2.4), we need now to approximate two expectations instead of one. While in the fully supervised case, the conditional log-likelihood ( $\log P_w(\mathbf{H} = \mathbf{h} | \mathbf{O} = \mathbf{o})$ ) is a concave function, the conditional log-

likelihood augmented with latent variables ( $\log \sum_{\mathbf{L}} P_w(\mathbf{H} = \mathbf{h}, \mathbf{L} = \mathbf{l} | \mathbf{O} = \mathbf{o})$ ) is not concave anymore. For the conditional log-likelihood augmented with latent variables, it is therefore not guaranteed that the global optimum of the function can be found via gradient ascent, but only a local optimum. Hence, a good initialization of the weights is crucial when learning with latent variables (Section 9.1.3).

**Implementation.** We implemented the discussed approach for weight learning with latent variables in *thebeast*. The algorithm requires two main adaptations. First, the inference part needs to be adapted. In the original version of *thebeast*, the MAP solution is calculated given the values of the observed predicates, i.e.  $P(\mathbf{H} | \mathbf{O} = \mathbf{o})$ . Learning weights with latent variables also requires to calculate the MAP solution given the values of the observed predicates and the hidden predicates for which the ground truth is known during training, i.e.  $P(\mathbf{L} | \mathbf{H} = \mathbf{h}, \mathbf{O} = \mathbf{o})$ . We extended the inference part of *thebeast* accordingly to account for the both options. Second, we implemented the discussed algorithm itself in *thebeast* to support weight learning with latent variables.

### 4.4.3 Discussion

In this section, we propose a discourse-aware approach for concept disambiguation and clustering. By incorporating cohesive scopes into the model, we account for the reciprocal relationship between the context that is relevant to disambiguate a mention and the concept of a mention (Implication III). As we use a predefined finite number of scopes and thus context definitions, the proposed approach uses a binwise context definition strategy. From a purely linguistic point of view, an individual context definition for each single mention (Section 2.3.1) is preferable over a binwise approach. However, it is hardly tractable. The scope-aware binwise approach balances between linguistic appropriateness and tractability.

Our proposed approach for scope-aware modeling is fairly general and can also be adapted for other tasks. We provide a formulation in Markov logic for binwise modeling of a target prediction task. Given that the following conditions are met for another task, the formulas can be easily adapted for it: (1) the target prediction task is dependent on a supportive classification task and vice versa; (2) training data are at least available for the target prediction task. Such candidate tasks are for instance polarity detection where the supportive task could be subjectivity detection as in Yessenalina et al. (2010) or bridging resolution where the target prediction task is the antecedent selection and the supportive task is for example the information status classification (Markert et al., 2012). In addition, the notion of cohesive scope could be analogously defined on other linguistic levels. For instance entity-level cohesive scopes might be relevant in the context of coreference resolution.

# Chapter 5

## Features

In this chapter, we describe the features used in our approach. This chapter builds upon Chapter 4 where we focus on the core of our approach, i.e. on how we model the interdependencies between the tasks. These interdependencies account for the implications we derived from our linguistic analysis (Chapter 2). While Chapter 4 only provides the backbone and some template predicates and formulas showing the general form of different formula types, we now fill these templates with content. In Chapter 6, we then explain our end-to-end concept disambiguation and clustering system.

Each feature roughly corresponds to one formula in Markov logic associated with a learned weight. Because of this one-to-one correspondence, we use the terms *feature* and *formula* interchangeably in the following.

In this chapter, we assume that the language to analyze is English. While the core provided in Chapter 4 is largely language-independent, the features discussed in this chapter are more sensitive to differences across languages and are partially dependent on the availability of language-specific resources. We will discuss the portability of the proposed approach in Chapter 7.

In the following, we first discuss the features used for concept disambiguation (Section 5.1) and concept clustering (Section 5.2), before we focus on the features for the scope assignment task (Section 5.3). Tables 5.1, 5.2, 5.3, 5.4 summarize all features. The tables show the feature identifier (*ID*) – we use these identifiers to refer to the features in the experiments (Chapter 9) –, the predicate name (*Predicate*) and the templates the corresponding formula is derived from (*Templ.*). Moreover, the tables contain brief feature descriptions and list the information sources we used to derive the features. In general, we use few, but strong features.

ID	Predicate	Templ.	Description & Information Sources
<b>Concept Disambiguation</b>			
<b>Prominence of a Concept</b>			
1	$hasPriorProbability(m, c, q)$	$t_{pd1},$ $t_{fd1}$	Prior probability of a concept given a mention according to corpus statistics <i>Information sources:</i> internal hyperlinks in Wikipedia
2	$hasStringEditDistance(m, c, q)$	$t_{pd1},$ $t_{fd1}$	String edit distance between a mention and the canonical names of a candidate concept <i>Information sources:</i> article and redirect titles from Wikipedia
<b>Co-occurrence Information</b>			
3,4	$hasRelatedness(m, c, q)$	$t_{pd1},$ $t_{fd1}$	Aggregated document-level relatedness score based on concept-level co-occurrence information <i>Information sources:</i> concept-level co-occurrences obtained from the internal hyperlinks in Wikipedia
5	$hasCoocProbability(m, c, q)$	$t_{pd1},$ $t_{fd1}$	Aggregated document-level relatedness score based on concept- and string-level co-occurrence information <i>Information sources:</i> concept- and string-level co-occurrences obtained from the internal hyperlinks in Wikipedia
6	$hasContextSimilarity(m, c, q)$	$t_{pd1},$ $t_{fd1}$	String-level sentence-based context compatibility score <i>Information sources:</i> surrounding context of internal hyperlinks in Wikipedia
<b>Type of Concept</b>			
7	$hasDescriptorSentence(m, c, q)$	$t_{pd1},$ $t_{fd1}$	Relative proportion of type information (e.g. ACTOR) in the same sentence, the neighbouring sentences and the whole document respectively
8	$hasDescriptorNeighbours(m, c, q)$	$t_{pd1},$ $t_{fd1}$	
9	$hasDescriptorDocument(m, c, q)$	$t_{pd1},$ $t_{fd1}$	<i>Information sources:</i> article and redirect titles from Wikipedia

Table 5.1: Features for concept disambiguation. The identifiers for the template formulas correspond to the templates in Table 4.4.  $m$  stands for a mention,  $c$  for a concept and  $q$  for the feature value (e.g. prior probability or aggregated relatedness). The features derived from template formula  $t_{fd2}$  are listed in Table 5.2, as they account for concept-level cohesive ties of type identity.

ID	Predicate	Templ.	Description & Information Sources
<b>Concept Clustering</b>			
10	$haveSameLemma(m, n, q)$	$t_{pc1},$ $t_{fc1}$	Pairs of mentions that occur in the same document and share the same lemma <i>Information sources:</i> lemma information obtained from the <i>Stanford CoreNLP</i> pipeline (Toutanova et al., 2003)
11	$isSubStringHeadMatch(m, n, q)$	$t_{pd2},$ $t_{fd2}$	Pairs of mentions that occur in the same document, share the same head lemma and are substrings of each other <i>Information sources:</i> head and lemma information obtained from the <i>Stanford CoreNLP</i> pipeline (Toutanova et al., 2003; de Marneffe et al., 2006)
11	$isAcronym(m, n, q)$	$t_{pd2},$ $t_{fd2}$	Acronym and a candidate full form obtained from the respective text; this feature has the same ID as the head match feature, as they share the same weight
12,13	$hasSubStringPersonName(m, n, q)$	$t_{pd2},$ $t_{fd2}$	Two mentions that are likely to be person names and one is a substring of the other <i>Information sources:</i> gender information from Bergsma & Lin (2006) to determine if a name is a potential person name
14	$hasCrossConceptSimilarity(m, n, q)$	$t_{pc1},$ $t_{fc1}$	Cross-document similarity based on a concept representation <i>Information sources:</i> statistics obtained from the internal hyperlinks in Wikipedia

Table 5.2: Features for concept clustering. The identifiers for the template formulas correspond to the templates in Table 4.4 and Table 4.5.  $m$  and  $n$  stand for mentions,  $c$  for a concept and  $q$  for the inverse distance in sentences between two mentions. In the *ML Clustering* approach (Section 9.1.3), all features are derived from template formula  $t_{fc1}$ .

## 5.1 Features for Concept Disambiguation

We use three types of information for concept disambiguation: the *prominence of a concept*, *co-occurrence information* and information regarding the *type of a concept*. In the following,

ID	Predicate	Templ. Description & Information Sources	
		Scope Assignment	
<b>Mention-based Features</b>			
15	$hasIdfHead(m, q)$	$t_{ps1},$ $t_{fs1}$	<i>Idf</i> score of the head of a mention <i>Information sources:</i> English Gigaword Corpus (Parker et al., 2011) for <i>idf</i> scores
16	$isPropername(m)$	$t_{ps1},$ $t_{fs1}$	A mention that is a proper name <i>Information sources:</i> proper name information obtained from the <i>Stanford CoreNLP</i> pipeline (Toutanova et al., 2003; de Marneffe et al., 2006)
17	$isSinglewordNoun(m)$	$t_{ps1},$ $t_{fs1}$	A mention that is a single word noun
18	$isAbbreviation(m)$	$t_{ps1},$ $t_{fs1}$	An abbreviation with a terminal dot
<b>Modification-based Features</b>			
19	$isPreModified(m)$	$t_{ps1},$ $t_{fs1}$	A mention that is pre-modified <i>Information sources:</i> syntactic information obtained from the <i>Stanford CoreNLP</i> pipeline
20	$isHeadOfRelClause(m)$	$t_{ps1},$ $t_{fs1}$	A mention that is the head of a relative clause <i>Information sources:</i> syntactic information
<b>Features based on the Sentence and Text Structure (Part I)</b>			
21	$isInSubjPosition(m)$	$t_{ps1},$ $t_{fs1}$	A mention in theme position, which is in English often the subject (Daneš, 1974) <i>Information sources:</i> syntactic information
22	$hasPosInSentence(m, q)$	$t_{ps1},$ $t_{fs1}$	Relative position of a mention in a sentence
23	$hasFocusingAdverb(m)$	$t_{ps1},$ $t_{fs1}$	A mention that is preceded by a focusing adverb <i>Information sources:</i> manually compiled list of focusing adverbs
24	$modifiesArgument(m)$	$t_{ps1},$ $t_{fs1}$	A mention that is a premodifier of a verbal argument <i>Information sources:</i> syntactic information

Table 5.3: Features for the scope assignment task. The identifiers for the template formulas correspond to the templates in Table 4.4.  $m$  is a mention,  $q$  is a feature value (e.g. relative position in sentence).

we discuss each category and explain the corresponding features. Some features are derived from previous work such as Milne & Witten (2008b). We discuss the features that model cohesive ties of type identity (Section 4.3.1) in Section 5.2.

ID	Predicate	Templ.	Description & Information Sources
<b>Features based on the Sentence and Text Structure (Part II)</b>			
25	$isPassiveBy(m)$	$t_{ps1},$ $t_{fs1}$	A mention that is the agent in a passive construction <i>Information sources:</i> syntactic information
26	$inConjunction(m)$	$t_{ps1},$ $t_{fs1}$	A mention that is part of a conjunction <i>Information sources:</i> syntactic information
27	$hasMorphoTiesHead(m, q)$	$t_{ps1},$ $t_{fs1}$	Frequency of the head of a mention in the text (including derivations of it) <i>Information sources:</i> derivational information obtained from <i>CatVar</i> (Habash & Dorr, 2003)
28	$hasPositionInText(m, q)$	$t_{ps1},$ $t_{fs1}$	Relative position of a mention in the text

Table 5.4: Features for the scope assignment task. The identifiers for the template formulas correspond to the templates in Table 4.4.  $m$  is a mention,  $q$  is a feature value (e.g. relative position in text).

### 5.1.1 Prominence of a Concept

Some concepts are per se more probable than other concepts as they are more frequent. Hence, features that model the prominence of a concept indicate how likely a concept is. We use two features that fall into this category: the prior probability and the string edit distance. While the former leverages corpus statistics to estimate the probability of a concept given a mention, the latter considers string distance information. We also experimented with other prominence scores such as the overall prominence of a concept independent of a specific surface form. We approximated this mention-independent prominence score by a normalized occurrence score and by the fraction of times an article has been viewed by a user in the past six months.<sup>1</sup> However, such information turned out to be less indicative and is not considered in the following.

Our two prominence scores are context-independent and can be considered as prior knowledge. In contrast to context-dependent features, their values only depend on the surface form and the candidate concepts and are determined a priori, i.e. before disambiguation.

**Prior probability (Table 5.1, 1).** The prior probability is defined as the conditional probability of a concept  $c$  given a mention  $m$ . If a corpus annotated with concepts is available the conditional probabilities can be estimated via

<sup>1</sup>We derived these numbers by downloading the hit counts (from January to June 2011) from <http://dammit.lt/wikistats/>.

$$p(c|m) = \frac{\text{counts}(m, c)}{\sum_{c' \in C_m} \text{counts}(m, c')}.$$

The estimator  $\text{counts}(m, c)$  is obtained by counting how many times the mention  $m$  denotes a concept  $c$  in the corpus according to the annotations.  $C_m$  comprises all candidate concepts of a mention according to a lexicon.

For mentions that show a skewed distribution over their candidate concepts this is an indicative feature. Gale et al. (1992a) therefore propose to exploit this feature as a lower bound for word sense or concept disambiguation. To obtain this lower bound each mention or word is assigned the concept or sense with the highest prior probability, i.e.

$$c^* = \arg \max_{c \in C_m} \frac{\text{counts}(m, c)}{\sum_{c' \in C_m} \text{counts}(m, c')}.$$

This context-independent lower bound, also known as the *most frequent sense or concept baseline* or the *first sense or concept baseline*, has turned out to be a strong baseline that, in particular, unsupervised approaches often fail to beat (Navigli, 2009).

The corpus from which the counts are obtained to estimate the prior probabilities highly influences the quality of this feature. As prior probabilities are domain-dependent (Martínez & Agirre, 2000; Koeling et al., 2005), the best results can be obtained if the corpus from which the statistics are obtained matches the domains of the test data. However, our test data covers multiple domains. Thus, we used a corpus that covers multiple domains and extracted the prior probabilities from all internal hyperlinks in the whole English Wikipedia dump (Section 3.2.1).

Mihalcea (2007) studied the quality of the concept distributions obtained from Wikipedia. According to these investigations, concept distributions derived from Wikipedia only show medium correlation with the concept distributions obtained from *SemCor*. This might indicate a lower quality of the concept distributions extracted from Wikipedia. However, our most frequent concept baseline obtains an F-measure above 60 on most test data sets. This shows that the concept distributions obtained from Wikipedia are of sufficient quality and that the prior probability is a strong feature.

**String edit distance (Table 5.1, 2).** This feature accounts for the string difference between the surface form of a mention  $m$  and the canonical names for a candidate concept  $c$ . The assumption is that the more distant the mention’s surface form is from the canonical names of a candidate concept, the less likely it is that the mention denotes this concept. Consequently, this feature indicates a negative relation between a candidate concept and a mention.

We assume that the Wikipedia article title and the titles of its redirects are canonical names ( $T_c$ ) for a concept  $c$ . To measure the distance  $\text{dist}(m, t_c)$  between a canonical name  $t_c \in T_c$  and

the mention  $m$ , we calculate the edit distance<sup>2</sup>  $edits(m, t_c)$  and normalize it by the length of the longer string. If more than one canonical name exists for a concept, we take the minimum distance given by

$$\arg \min_{t_c \in T_c} dist(m, t_c) = \frac{edits(m, t_c)}{\max(|m|, |t_c|)}.$$

Similar to the prior probability, the string edit distance is independent of the context and only considers the surface form and the candidate concepts of a mention.

### 5.1.2 Co-occurrence Information

Mentions that are tied on the concept level tend to disambiguate each other (Section 2.2). As the concept-level cohesive ties depend on the concepts denoted by the mentions, it is difficult to identify them before disambiguation. While we model concept-level cohesive ties of type identity via global formulas (Section 5.2), we use an iterative strategy to account for concept-level cohesive ties of type relatedness. Additionally, we also exploit correlations on the string level (Section 4.3.1).

We use three different co-occurrence scores. The first score has been proposed by Milne & Witten (2008b) and only takes into account concept-level co-occurrence information. The second co-occurrence score has been proposed by us (Fahrni et al., 2011b) and considers both string-level and concept-level co-occurrences. The third score exploits co-occurrences on the string level. Different variants of it have been used by e.g. Kulkarni et al. (2009) and Ratinov et al. (2011).

These co-occurrence scores also account for domain information. For instance, concepts that often appear together are likely to be of the same domain. Domain information is presumed to be effective for sense or concept disambiguation, as it often constrains the possible senses or concepts (Madhu & Lytel, 1965; Buitelaar et al., 2006; Agirre & Stevenson, 2006; Guiliano et al., 2009). In the context of word sense disambiguation, domains are conventionally characterized as “common areas of human discussion, such as economics, politics, law, science” (Gliozzo et al., 2004) or more practically as sets of “words between which there are strong semantic relations.” (Magnini et al., 2002). We also experimented with category and portal information extracted from Wikipedia (Fahrni et al., 2011b), but this information did not improve the results.

**Concept-level Co-occurrences (Table 5.1, 3, 4).** In the context of concept disambiguation, concept-level co-occurrence features have been extensively used (e.g. Agirre & Stevenson (2006, p. 242), Milne & Witten (2008b), Kulkarni et al. (2009) and Ratinov et al. (2011)). We

<sup>2</sup>We use the Lingpipe implementation (<http://alias-i.com/lingpipe/>).

follow Milne & Witten (2008b) and exploit concept-level co-occurrences that can be extracted from the internal hyperlinks in Wikipedia (Section 3.2.1). It is an aggregated co-occurrence measure. Calculating an aggregated score for a candidate concept requires (1) a relatedness or similarity measure, (2) a context definition and (3) an aggregation strategy.

The aggregated relatedness score proposed by Milne & Witten (2008b) is based on a pairwise relatedness measure (Milne & Witten, 2008a). Given two concepts  $c_m$  and  $c_n$ , the Normalized Google Distance ( $NGD\_concept$ ) (Cilibrasi & Vitányi, 2007) is calculated via

$$NGD\_concept(c_m, c_n) = \frac{\log(\max(|I(c_m)|, |I(c_n)|)) - \log(|I(c_m) \cap I(c_n)|)}{\log(|W|) - \log(\min(|I(c_m)|, |I(c_n)|))}.$$

$W$  is the total number of concepts in the inventory, while  $I(c_m)$  and  $I(c_n)$  denote the concepts tied to the concepts  $c_m$  and  $c_n$  respectively by incoming links (Section 3.2.1). If two concepts share an incoming link, they co-occur in a document. Hence, the more concepts are shared between  $I(c_m)$  and  $I(c_n)$ , i.e. the more often the two concept  $c_m$  and  $c_n$  co-occur, the higher the relatedness score is. This relatedness measure has been successfully applied for disambiguation (Milne & Witten, 2008b; Kulkarni et al., 2009; Ratnov et al., 2011) and leads to higher results than for instance measures based on the cosine similarity (Ratnov et al., 2011). We also experimented with concept-level co-occurrences extracted from the same sentence or the same paragraph in Wikipedia. However, these scores lead to lower results due to sparsity.

Given a mention  $m$  to be disambiguated, we follow Milne & Witten (2008b) and first identify all mentions that have a candidate concept with a prior probability higher than 0.95. The idea of this context selection strategy is to exploit that some mentions are unambiguous or are at least highly skewed towards a certain candidate concept. These concepts serve as disambiguation context  $C_{Dis}$  for the mention  $m$ . For each candidate concept  $c_m$  of the mention to be disambiguated, we calculate the pairwise relatedness score  $NGD\_concept(c_m, c_n)$  to each concept  $c_n \in C_{Dis}$ . The final aggregated relatedness score for the candidate concept  $c_m$  is then given as in Milne & Witten (2008b) by

$$NGD\_concept\_agg(c_m) = \frac{\sum_{c_n \in C_{Dis}} NGD\_concept(c_m, c_n) \cdot NGD\_concept\_avg(c_n)}{\sum_{c_d \in C_{Dis}} NGD\_concept\_avg(c_d)}$$

with  $NGD\_concept\_avg(c_n)$  being the average relatedness of a concept  $c_n \in C_{Dis}$  to all other concepts in the disambiguation context weighted by the probability of the corresponding mention to be linked. The higher the  $NGD\_concept\_avg$  score for a concept  $c_n \in C_{Dis}$  is, the more it is considered as indicative for this context. Hence, the final score  $NGD\_concept\_agg$  for a candidate concept  $c_m$  depends on its relatedness to concepts in the disambiguation context, but also on how indicative the related concepts in the disambiguation context are for the context.

The higher the  $NGD\_concept_{agg}$  score for a candidate concept  $c_m$  is, the more likely it is that a mention denotes that candidate concept  $c_m$ . We experimented with sentence-, paragraph- and document-level context definitions. As this aggregated score relies on the availability of unambiguous mentions or at least mentions with a highly skewed concept distribution, it turned out to be most suitable for a document-level context definition. We thus use it to account for the document-level context and consider it as a positive or a negative feature for a candidate concept depending on its value: if the score for a candidate concept is higher than 0 it means that this candidate concept is at least partially related to its context and we consider it as a positive indicator (Table 5.1, 3); if its score is 0 it means that the candidate concept is not related to its context and we use it as negative indicator for the candidate concept (Table 5.1, 4). Hence, we use two different formulas for which we learn different weights.

**Concept- and String-level Co-occurrences (Table 5.1, 5).** In addition to the concept-level co-occurrence score introduced above, we used the same aggregation strategy and context definition with a different pairwise score that accounts for concept- and string-level co-occurrences.

This pairwise score  $cooc(c_m, c_n|m, n)$  builds upon the conditional co-occurrence probability of a concept pair given a mention pair (Fahrni et al., 2011b) and is defined as

$$cooc(c_m, c_n|m, n) = p(c_m, c_n|m, n) - chance(C_m, C_n)$$

with  $c_m$  being a single candidate concept and  $C_m$  encompassing all candidate concepts of the mention  $m$ .  $c_n$  and  $C_n$  are defined analogously. The conditional co-occurrence probability is approximated via

$$p(c_m, c_n|m, n) = \frac{counts(c_m, c_n)}{\sum_{c'_m \in C_m, c'_n \in C_n} counts(c'_m, c'_n, m, n)}$$

where  $counts(c_m, c_n, m, n)$  is the number of times the mentions  $m$  and  $n$  denote  $c_m$  and  $c_n$  respectively. We subtract the chance from the conditional co-occurrence probability given by

$$chance(C_m, C_n) = \frac{1}{|C_m| \cdot |C_n|}$$

to account for the case that  $c_m$  and  $c_n$  given  $m$  and  $n$  could also randomly occur together. To ease further computations such as calculating aggregated scores, we ensure that the final score is between 0 and 1 via a linear transformation.

**String-level Co-occurrences (Table 5.1, 6).** While concept-level co-occurrences can only be identified for tokens with a corresponding concept in the inventory, string-level co-occurrences can also be extracted with context tokens that lack a corresponding concept in the inventory. As in Wikipedia, from where we extract the concept-level co-occurrences, only some

words are annotated – mainly noun phrases –, string-level co-occurrences may at least partially compensate for unannotated context tokens. In particular, context tokens of parts of speech that are hardly annotated in Wikipedia such as verbs or prepositions can be considered via string-level co-occurrences.

To exploit co-occurring string-level information, we follow previous work (e.g. Ratinov et al. (2011) and Kulkarni et al. (2009)) and extract string-level co-occurrence information from the English Wikipedia. For each concept, we retrieve all the internal hyperlinks that point to it and extract all co-occurring lemmas  $L_c$  in a certain context window. Given a mention  $m$  to be disambiguated, we use the same context definition and extract all lemmas  $L_m$  within this context window from its text. For each candidate concept  $c_m$  of mention  $m$  we then calculate the string-level context similarity via

$$\text{sim}(c_m, L_m, L_c) = \frac{1}{|L_m|} \sum_{l \in L_m} s(l, L_c).$$

The first term is used for normalization and  $s(l, L_c)$  denotes the frequency of  $l$  in  $L_c$  divided by the number of times  $l$  appears in the context of all concepts in Wikipedia.

We experimented with different context definitions and different similarity metrics. It turned out that this feature is not effective on the document level, but on the sentence level. We thus use it to measure the local context with a sentence-level context definition.

### 5.1.3 Concept Type Information

Type information, e.g. that *Watson* in a specific sentence must denote a concept of type COMPUTER and not an ACTOR can be useful for disambiguation, in particular if the candidate concepts of a mention are of different types. Our sentence-level score (Table 5.1, 6) partially accounts for this information.

However, in particular for proper names this information is insufficient, especially if many candidate concepts of a mention are of the same type. For instance, *Hyderabad* denotes many different places. We thus use some additional features that account for type information in the experiments that focus on proper names (TAC, CoNLL).

**Concept Type Information (Table 5.1, 7, 8, 9).** Article titles in Wikipedia often contain type information in the article name either in parentheses or after a comma. We call these descriptions *descriptors*. For each candidate concept of a mention, all canonical names are obtained (redirects and article titles) and all descriptors that either appear in parentheses or after a comma are extracted from these names. It is then checked how often these descriptors occur in the neighboring sentences of a mention, in the same sentence or in the same document. The score is the portion of domain descriptors for a candidate concept given the domain

descriptors for all candidate concepts for a mention.

## 5.2 Features for Concept Clustering

Concept clustering features model cohesive concept-level relations of the type identity. While cohesive ties of type relatedness are difficult to predict, cohesive ties of type identity are easier to identify (Section 4.3.1). We exploit such identity relations and jointly disambiguate mentions that are tied by such an identity relation. We distinguish between two different types of features: features for within-document concept clustering (Section 5.2.1) and features for cross-document concept clustering (Section 5.2.2).

### 5.2.1 Features for Within-document Concept Clustering

For within-document concept clustering, we use the features described in the following.

**Shared lemma (Table 5.2, 10).** The one sense per discourse hypothesis states that within one discourse one mention string denotes one sense, i.e. in our case one concept (Gale et al., 1992c). This assumption is well tested (Krovetz, 1998; McCarthy et al., 2007) and has been successfully exploited by e.g. Yarowsky (1995). McCarthy et al. (2007) found that the one sense per discourse hypothesis better applies to nouns than other part-of-speech tags. Only in 27.8% of the cases where a polysemous noun appears more than once in a document it is used ambiguously (McCarthy et al., 2007, p. 559).

For each document we extract all mentions with the same lemma and calculate for each such pair the inverse distance in sentences. The bigger the inverse distance is, the closer the two mentions are to each other and the more likely it is that they denote the same concept. This feature leverages that cohesive ties of type identity on the string-level often correlate with identity relations on the concept-level.

**Substring with the Same Head (Table 5.2, 11).** If one mention  $n$  is a substring of another mention  $m$  and they share the same head, it is likely that they also denote the same concept. This observed tendency relaxes the one sense – or concept – per discourse hypothesis by not requiring fully identical strings or lemmas (Gale et al., 1992c). For instance, if the mention *peer review* and the mention *review* appear in the same text, it is likely that both denote the concept PEER REVIEW. As this example illustrates, longer strings tend to contain more information and are therefore often less ambiguous than shorter strings (Csomai & Mihalcea, 2008). We exploit this disparity in ambiguity and state for mentions that are in a substring relation and share the same head (in terms of lemmas): if the longer mention  $m$  denotes a specific concept, the shorter mention  $n$  denotes the same concept. This formula allows us

to propagate concepts from one mention to another. Even if for instance PEER REVIEW is initially not a candidate concept of *review*, it can still be disambiguated to this concept via this formula (Section 4.3.1). Distance is important in this context. As stated by Halliday & Hasan (1976, p. 289-290), the strength of a cohesive tie not only depends on the relatedness, but also on the closeness in the text. The closer the two mentions appear together in the text, the more likely such a concept-level identity relation between them is. We therefore use the inverse distance in sentences to score the substring relations.

**Acronyms (Table 5.2, 11).** Acronyms are often more ambiguous and more difficult to disambiguate than the fully expanded version. Analogously to the substring feature, we also define an acronym feature. We first identify acronym mentions considering all mentions that consist of one token and contain uppercase letters within the token string as acronyms. In the next step, we search for the full version of the acronyms in the text. A mention is considered as the full version of an acronym if the first character of each token of the mention forms the acronym. In addition, we exploit the pattern *mention (acronym mention)* with the first mention being the full version and the acronym mention being the acronym. This pattern is quite common in news paper texts. The score is given by the inverse distance in sentence between the acronym and the full version.

As in our Wikipedia training data, acronyms are relatively rare, it is difficult to learn a weight for the acronym feature. As it is similar to the substring feature, we use the same predicates, formulas and weight for these two features.

**Person Name Substrings (Table 5.2, 12, 13).** For person names, we use an additional relaxed version of the substring and head match feature and only require that the mention  $n$  is a substring of mention  $m$ . By dropping the head match requirement we account for the common practice of denoting people by their first name which is not the head of the full person name. To also account for short forms of first names, e.g. *Dan* for *Daniel Craig*, the substring can also only comprise the beginning of a token of mention  $m$ . We derive two formulas for this feature. The first formula says that if the longer name version denotes a certain concept, the shorter version is likely to denote the same concept. The second formula says that if the longer name version does not denote a certain concept, the short name version is also likely to not denote this concept. To decide if a mention is person name we rely on the gender information in Bergsma & Lin (2006). As for the substring features, we score the person name substring relation by the inverse distance in sentences.

We also tried to exploit coreference information, i.e. correlations between the entity and the concept level. However, given all other features, this feature did not add any additional information.

### 5.2.2 Features for Cross-document Concept Clustering

Although our approach also works for single documents, we generally assume that we work with a text corpus consisting of a finite number of texts. Features for cross-document concept clustering model concept-level identity relations across documents.

**Cross-document Concept-level Similarity (Table 5.2, 14).** Vector space models have been successfully applied to calculate the similarity between texts (Salton & Lesk, 1968) and are highly used in cross-document concept clustering (Schütze, 1998; Bagga & Baldwin, 1998b; Gooi & Allen, 2004) or similarity calculations between words (Church & Hanks, 1990; Turney, 2006). Given two mentions from different documents, we calculate the string similarity based on the edit distance<sup>3</sup> between them. If the string similarity between the mentions is higher than a certain threshold<sup>4</sup> or they are substrings according to one the substring features (Section 5.2.1), we calculate similarity between the two contexts  $t_1$  and  $t_2$  as follows. We represent each of them by a vector  $v_1$  and  $v_2$  respectively. The coordinates of these vectors are *tf idf* values of concepts with the *tf* given by the frequency of the concept in the current document. The *idf* score of a concept is calculated based on the internal hyperlinks in Wikipedia. To obtain the concept frequencies for each context representation, we identify all mentions in the respective text that are unambiguous or have a highly skewed concept distribution (Section 5.1.2).

The similarity between these two vectors is then given by their cosine similarity, i.e.

$$\text{sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \cdot \|v_2\|}.$$

We also experimented with a string-based and a category-based representation. However, while these features slowed down the inference, they did not lead to substantial improvements.

## 5.3 Features for Scope Assignment

In the following, we describe all formulas that we use for the scope assignment task. We assume that three different scopes are distinguished: local scope, intermediate scope and global scope (Section 2.3.2). As we have discussed in Section 2.2, the scope of a mention depends on its embedding into discourse. We use three different types of features: mention-based features (Section 5.3.1), modification-based features (Section 5.3.2) and features based on the sentence and text structure (Section 5.3.3).

<sup>3</sup>We use the Lingpipe implementation (<http://alias-i.com/lingpipe/>).

<sup>4</sup>We empirically set the threshold to 0.9.

### 5.3.1 Mention-based Features

Mention-based features incorporate information related to the respective mention and mainly help to tell apart local scopes from intermediate and global scopes.

***Idf of the Head* (Table 5.3, 15).** Halliday & Hasan (1976, p. 290) state that the cohesive force of a lexical item on other items is influenced by its overall frequency. Highly frequent lexical items tend to be less important for the textual cohesion than less frequent or even rare lexical items. Hence, mentions that are highly frequent tend to be of local scope, whereas rare mentions are more likely to be of intermediate or global scope. We estimate the frequency of a mention by the *idf* score of its head obtained from the Gigaword Corpus (Parker et al., 2011). We only consider the heads of the mentions to extract better statistics. This feature is also inspired by work on indexing for information retrieval (Spärck Jones, 1972).

**Proper names (Table 5.3, 16).** Proper names tend to be more prominent than common nouns and are more likely to have an intermediate or global scope than common nouns. This is also supported by studies in the field of summarization. Proper names tend to appear in human summaries which indicates their importance for the whole text (Hong & Nenkova, 2014). To identify proper names, we rely on the named entity recognizer used in the preprocessing step.

**Single word nouns (Table 5.3, 17).** Single word mentions are often less prominent than multi-word mentions, and are therefore more likely to be of local scope.

**Abbreviations (Table 5.3, 18).** Abbreviations with a terminal dot such as *Mr.* or *Ms.* tend to have a local scope as they are usually local modifiers or specifications.

### 5.3.2 Features Based on Modification

Modification-based features take into account aspects related to the modification of mentions. Modifiers can be useful for the disambiguation of a mention if they describe the mention more precisely. Hence, the local context is likely to be informative for mentions that are modified. At the same time, mentions that are prominent are more likely to be modified than mentions that are less important. Modified mentions therefore tend to be of intermediate scope, as both the local and the global context are relevant for them.

**Premodification (Table 5.3, 19).** For each mention, it is checked if it is premodified by some adjectives. Premodifiers are identified based on the syntactic dependencies. All sentences are parsed with a dependency parser during the preprocessing. Mentions that are premodified tend

to be more prominent than other mentions, as they are described in more detail. They are thus less likely to be of local scope.

**Head of Relative Clause (Table 5.3, 20).** From the dependency tree, we derive if a mention is the head of a relative clause. This is a good indicator to detect mentions of intermediate scope, as for these mentions both the local and the global context tend to be relevant.

### 5.3.3 Features Based on the Sentence and Text Structure

Features based on the text structure incorporate information how the mentions are embedded into the larger context.

**Theme position (Table 5.3, 21, 22).** Mentions in theme position tend to pick up what has already been mentioned before (Daneš, 1974). Although this has been originally stated with respect to discourse entities, we assume that it also applies to concept-level cohesive ties. It implies that mentions in theme position are more likely to establish cohesive ties with the larger context and accordingly tend to be of intermediate or global scope.

We approximate the theme by syntactic dependency and positional information. The subject feature exploits that the theme is often the subject in English. The second feature considers the relative position of a mention in a sentence. The earlier a mention appears in the sentence in English, the higher the score and the more thematic it is.

**Focusing Adverbs (Table 5.3, 23).** Focusing adverbs such as e.g. *particularly* allow to stress a mention. This happens for instance in the text pattern *<focusing adverb> <mention>* – e.g. “*particularly Jack*”. Mentions that are stressed are more prominent and therefore tend to be of intermediate or global scope. We manually built a list of focusing adverbs.

**Modifier of a Verbal Argument (Table 5.3, 24).** A premodifier of a verbal argument is more likely to be of local scope, as it often only describes the verbal argument.

**Agent in Passive Sentence (Table 5.4, 25).** In a passive construction (e.g. *the thief was caught by the police*), the agent (*police*) – if it is realized at all – tends to be less in the focus than in the corresponding active formulation and is prone to be of local scope.

**Mentions in Conjunctions (Table 5.4, 26).** Since conjunctions are often used for exemplifications, mentions in conjunctions tend to be of local scope. At least the local context is usually relevant to disambiguate them.

**Morphological Ties (Table 5.4, 27).** The more often the head of a mention appears in the text – also as a derivation – the more prominent it is. Using this feature, we exploit correlations of cohesive ties on the string- and concept-level. Milne & Witten (2008b) also use a frequency feature for keyword identification assuming that more frequent concepts are more likely to be important and therefore good keyword candidates. Derivational information is obtained from CatVar (Habash & Dorr, 2003).

**Position in the Text (Table 5.4, 28).** The earlier a mention appears in a text, the more likely it is to exhibit intermediate or global cohesive scope. This is in line with the hard-to-be-beat lead baseline in summarization (Radev et al., 2003).

## Chapter 6

# Monolingual English Architecture

In previous chapters, we examined concept disambiguation and clustering in isolation. However, an end-to-end concept disambiguation and clustering system additionally involves text preprocessing, mention recognition and candidate concept identification – all aspects that we mainly disregarded so far or implicitly considered as given.

In this chapter, we discuss these hitherto omitted aspects and show how the proposed disambiguation and clustering approach is integrated in an end-to-end system. While we assume that the texts to process are in English in this chapter, we show in the next chapter how we can extend the proposed system for multi- and cross-lingual concept disambiguation and clustering.

Our proposed approach for concept disambiguation and clustering builds upon the findings of a linguistic analysis (Section 2.4) and explicitly models the interrelations between concept disambiguation, NIL recognition, concept clustering and context selection. These modeled interrelations shape the overall architecture of our approach and lead to a workflow that is different from the workflow of previous approaches. Figure 6.1 contrasts a typical cascaded approach for concept disambiguation and clustering with our joint approach. As Figure 6.1 illustrates, our joint approach is distinguished from the cascaded approach in two ways. First, in our proposed joint approach all texts to process are first preprocessed and all mentions and the candidate concepts are identified. Then the disambiguation, the recognition of NILs and the clustering is performed jointly across documents. In contrast, in a standard cascaded approach, each document is disambiguated in isolation and NILs are recognized in an additional step either before or after the disambiguation. Only the clustering is done across documents in an additional step, but in contrast to our joint approach only NILs are clustered. Such a cascaded architecture prevents the explicit modeling of interrelations between the different tasks and suffers from error propagation. Second, while our approach uses different models depending on the jointly identified cohesive scopes, other approaches generally apply one single model to all mentions.

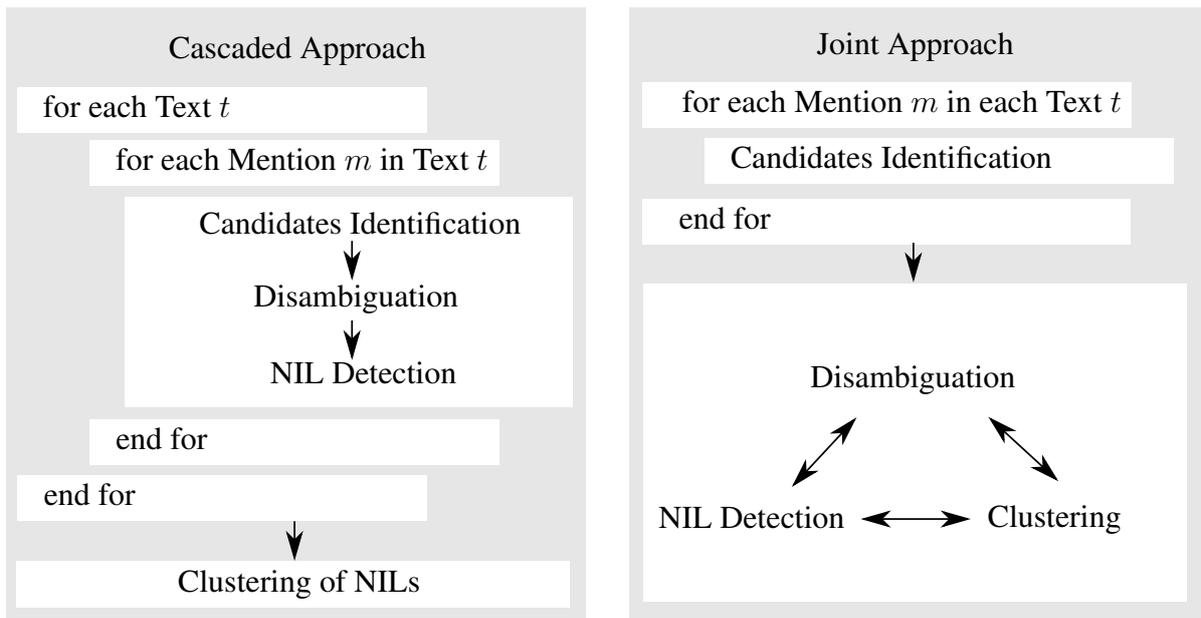


Figure 6.1: Cascaded approach vs. joint approach.

Figure 6.2 illustrates the workflow of our end-to-end concept disambiguation and clustering system. Given that we have a number of documents - in the example only two -, we first preprocess them (Section 6.1) and identify the mentions in these documents and their candidate concepts (Section 6.2). Then we extract all the features (Section 6.3). As features can cross the boundaries of documents, we regroup the mentions into new pseudo documents, which are then given to the inference module that outputs the results.

## 6.1 Preprocessing

The preprocessing includes text cleaning and linguistic preprocessing.

The importance of the text cleaning depends on the source of the input texts. While it is for instance essential for web documents, it is less crucial for news paper texts. During the text cleaning phase, we remove all markup and sequences of special characters such as e.g. sequences of stars or diamonds.

The linguistic preprocessing comprises tokenization, sentence splitting, lemmatization, named entity recognition and dependency parsing. For all these steps, we use the *Stanford CoreNLP* pipeline (Toutanova et al., 2003; Finkel et al., 2005; de Marneffe et al., 2006; Lee et al., 2011)<sup>1</sup>. The maximum sentence length for parsing is set to 70. To ensure a fair evaluation, we exclude the NER models that are partially trained on our testing data (e.g. ACE data) while processing them.

<sup>1</sup><http://nlp.stanford.edu/software/corenlp.shtml>, version from May 22nd 2012.

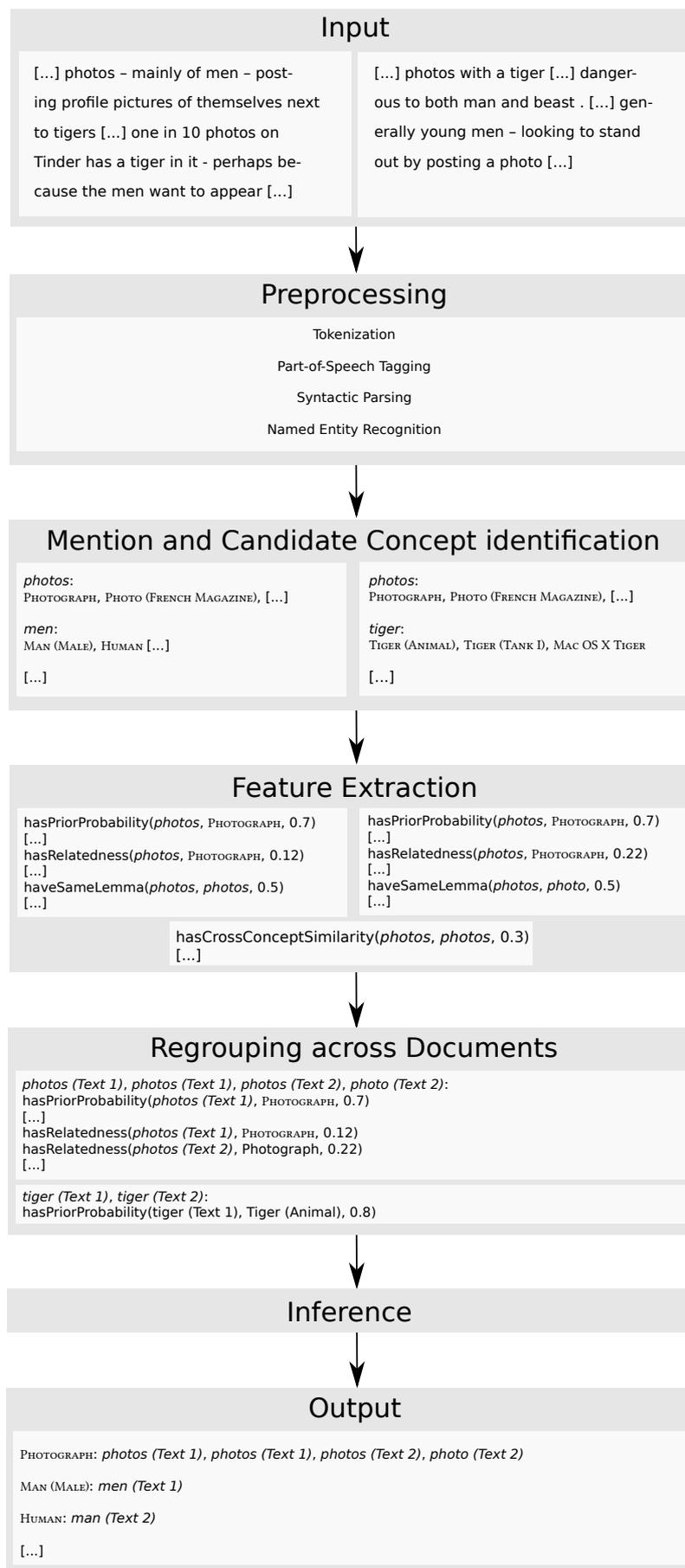


Figure 6.2: Workflow for monolingual English concept disambiguation and clustering.

## 6.2 Mention Recognition and Candidate Concept Identification

The purpose of the mention recognition is to identify continuous chunks of tokens that build one unit for concept disambiguation and clustering. Mention recognition is closely related to work on multiword expressions (see Sag et al. (2002) or Gelbukh & Kolesnikova (2013) for a survey) – which has been studied in the context of word sense disambiguation (Finlayson & Kulkarni, 2011) and Wikipedia (Vincze et al., 2011) – and named entity recognition (see Nadeau & Sekine (2007) for a survey).

In this thesis, we use a rule-based mention recognizer that relies on our lexical resources extracted from Wikipedia (Section 3.2.1), syntactic information and named entity recognition (Section 6.1). We first obtain all noun phrases (excluding discontinuous phrases and determiners) from the syntactic dependency trees. To compensate for errors made by the syntactic parser, we additionally check for each  $n$ -gram in the text of length one up to six tokens if it exists in our lexicon. As our lexicon also contains noise, we only consider such  $n$ -grams as mentions if their keyphraseness exceeds a certain threshold. Mihalcea & Csomai (2007) define keyphraseness as the probability of an  $n$ -gram to be linked in Wikipedia and use it to rank keywords. We assume that  $n$ -grams that tend to be keywords are of higher quality. As Mihalcea & Csomai (2007) we only consider  $n$ -grams that appear at least 5 times in Wikipedia – for others no reliable statistics can be obtained – and use an empirically tuned threshold of 0.1. Moreover, we add all nonnumeric named entities that have been recognized by the named entity extractor to our mention pool. For each mention in this pool, we determine its syntactic head based on the syntactic dependency information. If the head cannot be identified in this way, we consider the last token of the mention as a head. This is a heuristic that is commonly used in English natural language processing (e.g. Poon & Domingos (2008)).

For each mention, we retrieve candidate concepts from our lexicon. The main challenge of the candidate identification step is to balance between high recall and average ambiguity. A high average ambiguity makes the task more difficult and reduces the efficiency (Ratinov et al., 2011; Hachey et al., 2013). Given a mention, we use different string variations for the lexicon lookup, i.e. its token string, its lemma string and the two strings without any white space. As all our lexicon entries are in lower-case, we also use lower-cased versions for the lexicon lookup. Even if no candidate concept can be obtained for a mention, it still can be assigned a concept during the inference due to the clustering information (Section 4.3.1).

## 6.3 Feature Extraction and Inference

After the mentions and their corresponding candidate concepts have been identified in all input documents, we extract all intra- and cross-document features (Chapter 5) and run the MAP inference (Section 4.2.3). To speed up the inference, we split the whole input into different tranches and perform the inference for each tranche separately. All mentions that are tied by an intra- or cross-document global formula are supposed to be part of the same tranche, as their corresponding assignments influence each other. If the tranches become too big, the inference slows down. However, we only deal with relatively small input corpora consisting of a few hundreds of texts (Chapter 8).

For the MAP inference we use *thebeast*<sup>2</sup> (Riedel, 2008) with *Gurobi* (Gurobi Optimization, 2014) as ILP solver. We run the inference on a Linux server with 48 cores and 500 GB of memory.

The final results are then extracted from the output of the inference and stored as stand-off XML annotations. We use a format inspired by MMAX (Müller & Strube, 2001) that has been used in the context of the EU project CoSyne and that is publicly available.<sup>3</sup>

## 6.4 Discussion

The proposed concept disambiguation and clustering system is subject to a few assumptions that all concern the input texts. First of all, some features and the modeled interrelations are based on the assumption that each input text is coherent and involves a few main topics. In particular, the aggregated relatedness features and the scopes are obsolete if a sequence of unrelated sentences is processed. This assumption is inherent to our approach. In contrast, all other assumptions are not inherent to the approach, but rather a result of the current implementation.

Generally, our linguistic preprocessing works better for news texts than for noisy texts with lots of errors and less standard formulations. The quality of the preprocessing and irregularities in e.g. orthography mainly affect the mention recognition and the candidate concept identification, which in turn influence the outcome of the disambiguation and clustering. Noisy texts can still be processed, but the quality of the disambiguation and clustering might be lower. By enhancing the robustness – e.g. our mention recognition already partially accommodates for irregularities by combining a syntax- with a lexicon-based strategy –, the system can be adapted for noisier input.

Another constraint of our current implementation relates to the number of documents that can be processed at the same time. The system can deal with a few hundreds of texts, but

---

<sup>2</sup><https://code.google.com/p/thebeast/>, 15.5.2014.

<sup>3</sup><http://sourceforge.net/projects/cosyne-eu/>, 5.6.2014.

not thousands of them. The bottleneck is the MAP inference. If the tranches become big, the inference slows down. Splitting the tranches is problematic, as only mentions within one tranche can be clustered. In case many tranches need to be split, an additional clustering step across tranches might be required. However, this thesis does not address issues related to distributed inference (Singh et al., 2011).

## Chapter 7

# Multi- and Cross-lingual Concept Disambiguation and Clustering

The previous chapters focused on monolingual concept disambiguation and clustering. Using English as a development language, we analyzed the tasks and proposed a model for monolingual concept disambiguation and clustering. However, given the vast amount of texts written in different languages, an approach that can easily be adapted to languages other than English is preferable over an approach that only works for English. In this chapter, we therefore investigate how our concept disambiguation and clustering approach can be adapted for other languages than English.

We identified two dimensions along which a monolingual concept disambiguation and clustering system can be extended, i.e. the multi- and the cross-lingual dimension (Figure 1.6). A concept disambiguation and clustering system that is extended along the multilingual dimension is able to disambiguate and cluster concepts in different languages. A concept disambiguation and clustering system that is extended along the cross-lingual dimension is capable to disambiguate and cluster concepts across languages. In the following, we visit two scenarios: the cross-lingual scenario in which texts are still in English, but the inventory is in another language and the multi- and the cross-lingual scenario in which the inventory is in English, but texts are in different languages.

The success of a multi- or cross-lingual extension can be measured by the *performance* of the system and the *costs* that the adaptation requires (e.g. in terms of time or money). Ideally, no or only little effort is necessary, while the performance is stable across languages. Hence, the aim is to minimize the costs and maximize the performance (Khapra et al., 2010).

In the following, we first discuss different strategies for multi- and cross-lingual concept disambiguation and clustering (Section 7.1). We then analyze, which parts of our approach are language-specific and which scale well across languages (Section 7.2). This analysis builds the basis for our multi- and cross-lingual concept disambiguation and clustering approach which

is described in Section 7.3. Finally, we discuss the languages we selected to evaluate our approach (Section 7.4). This chapter provides more detail about the linguistic assumptions incorporated in our approach. Only if the linguistic assumptions underlying a system are revealed, it can be analyzed how a system may cope with other languages (Bender, 2011, p. 6).

## 7.1 Strategies for Multi- and Cross-lingual Concept Disambiguation and Clustering

To design a multi- or cross-lingual concept disambiguation and clustering system, different strategies can be applied. In the following, we describe different multi- and cross-lingual strategies (Section 7.1.1, 7.1.2) and discuss their advantages and disadvantages. We then describe the strategy used in this thesis (Section 7.1.3).

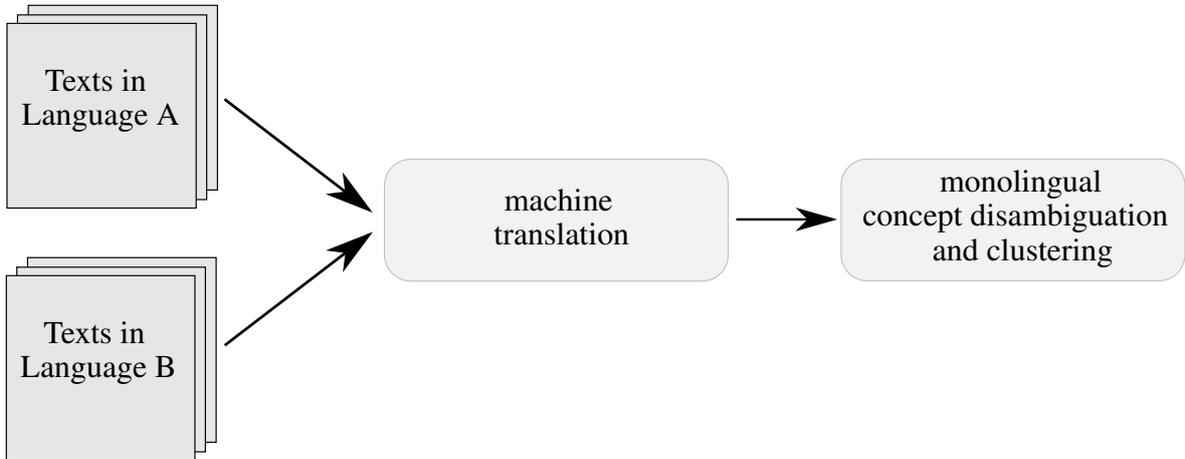
### 7.1.1 Multilingual Concept Disambiguation and Clustering

In contrast to a monolingual system, a multilingual system can cope with input texts in multiple languages. We distinguish between three general strategies to design a multilingual concept disambiguation and clustering system: a *translation-based strategy* that exploits machine translation techniques to translate the multilingual problem into a monolingual problem, an *adaptation-based strategy* that involves language-specific adaptations and a *language-independent strategy* that requires a language-independent design of the system. Figure 7.1 illustrates these three strategies that we will discuss in the following in more detail. In the whole discussion, we assume that an inventory is available for the languages under considerations.

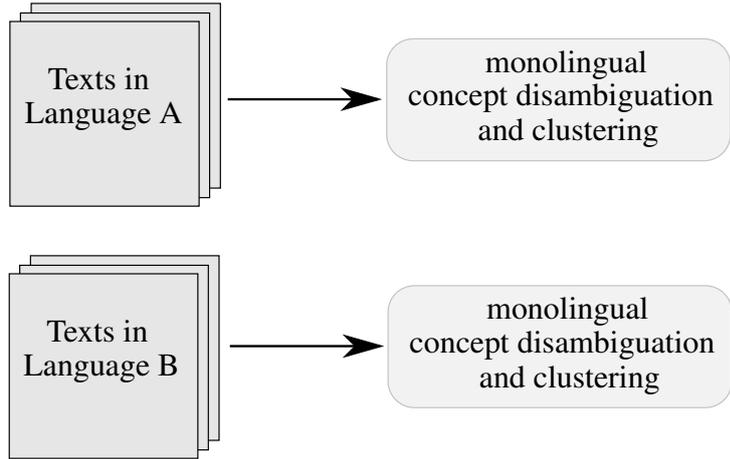
**Translation-based Strategy.** To bypass a possibly tedious system adaptation for each additional language to cover, translation-based approaches exploit translation techniques. The idea is to use one single monolingual system – that operates for instance in English – and to transform the multilingual input so that this system is able to process it. The applied translation-based techniques range from mention translation based on dictionaries or phrase tables to full text translation using a machine translation system.

The main advantage of such an approach is that the disambiguation and clustering system requires no adaptations. In addition, resource-poor languages may benefit from better resources that are available for the system language. For instance, the string-level co-occurrences for English might be of higher quality than the ones for e.g. Bulgarian, as more annotated data is available for English. For the same reason, the concept-level co-occurrences

**Translation-based Strategy**



**Adaptation-based Strategy**



**Language-independent Strategy**

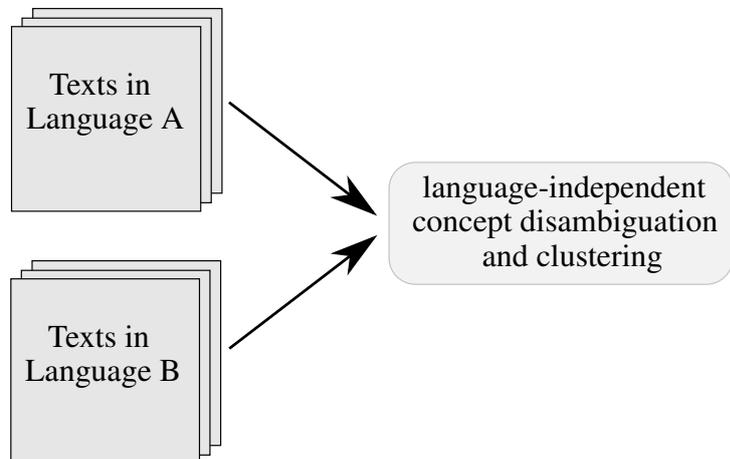


Figure 7.1: Strategies for multilingual concept disambiguation and clustering

are richer and more reliable for a resource-rich language such as English than for a resource-poor language. On the downside, translation is a difficult task that may introduce many errors. These errors are propagated and may hamper the performance of the disambiguation and clustering. In particular, resource-poor languages for which the phrase-tables and dictionaries may have low coverage, suffer from such error propagation. In case only the mentions are translated, the machinery is less heavy than if the full text is translated. However, if only the mentions are translated all other context information is completely lost. From a broader perspective, a translation-based approach that uses a complete translation system is questionable, as the purpose of concept disambiguation – which is more a supportive task than an end application – is to resolve ambiguities to improve downstream applications such as machine translation. We therefore do not consider a translation-based strategy as an option for our multilingual approach any further.

**Adaptation-based Strategy.** In adaptation-based approaches, a separate concept disambiguation and clustering system is used for each language.

Depending on how language-dependent a system is, the adaptation might be tedious and expensive. These costs are one of the main disadvantages of such an approach. However, by adapting a system for a particular language it can be tuned to this language by leveraging language specificities. Such a fine tuning is not possible in the translation-based approach. Moreover, as the system directly analyzes the input texts, it is free from error propagation originating from translation. In contrast to translation-based approaches that only translate the mentions, the whole context can be exploited. Besides the high cost, the other main disadvantage is that for resource-poor languages only few annotated data might be available which leads to worse statistics, e.g. to estimate prior probabilities, co-occurrences or to tune the model parameters.

**Language-independent Strategy.** According to Bender (2011, p. 6) a “truly language-independent system works equally (or nearly equally) well across languages.” Given that such a language-independent concept disambiguation and clustering can be designed, it could be applied to any language without requiring any tedious adaptations or a heavy translation machinery. As Bender (2011, p. 4–5) points out, language-independent should not be confused with linguistic-lean. A system that uses hardly any linguistic knowledge (e.g. n-gram model) can nevertheless hide some linguistic assumptions (e.g. assumptions about the word order) (Bender, 2011, p. 4–5).

Hence, while a language-independent strategy is cost-effective, it also does not exploit language-specific knowledge that may boost the performance for a specific language. In addition, it is extremely challenging to design a strictly language-independent approach for concept

disambiguation and clustering. It is more realistic to design an approach that is language-independent given certain prerequisites (e.g. the same writing system).

### 7.1.2 Cross-lingual Concept Disambiguation and Clustering

Cross-lingual concept disambiguation involves some input texts in a source language and an inventory that has been developed for another language, which we henceforth call the target language. The mentions in the input texts are then disambiguated with respect to the inventory designed for the target language. In cross-lingual concept clustering, the input texts are in different languages and mentions are clustered across languages. While cross-lingual concept disambiguation does not necessarily imply that the system is multilingual – e.g. the source language can for instance be English, while the target language is Chinese –, cross-lingual concept clustering requires input texts in different languages. In the following, we assume that the input texts are in different languages. Hence, such a system is not only cross-lingual but also multilingual, as it processes texts in multiple languages.

Figure 7.2 shows three strategies for cross-lingual concept disambiguation and clustering: a *target-sided*, a *source-sided* and a *cross-lingual strategy*. These strategies are also distinguished by Ji et al. (2011).

**Target-sided Strategy.** If a target-sided strategy for cross-lingual concept disambiguation and clustering is pursued, the input texts are first translated into the target language, before a monolingual concept disambiguation and clustering system is applied to them. Either the whole texts are translated or only the mentions. In the context of cross-lingual textual entailment, this approach is known as the pivoting approach (Mehdad et al., 2011).

The advantages and disadvantages are comparable to the ones for the multilingual translation-based strategy discussed above.

**Source-sided Strategy.** If a source-sided strategy is pursued, all mentions are first disambiguated and clustered – separately for each source language – by using e.g. an adaptation-based or language-independent concept disambiguation and clustering system (Section 7.1.1). Then the clusters are merged across languages and the concepts are mapped to the concepts in the target inventory. This mapping and merging steps – which can be understood as a translation step – require some cross-lingual resources such as a mapping between inventories, some dictionaries or a translation system.

The advantages and disadvantages correspond to the ones of the multilingual adaptation-based and language-independent strategy respectively. In contrast to the target-sided strategy, the main advantage is that the original context that is free of potential machine translation errors can be exploited for disambiguation and clustering. On the downside, interdependencies

between mentions can only be leveraged within the same source language, but not across source languages. Consequently, the clustering is divided into two steps – within-language clustering and across-language clustering – which may lead to error propagation.

**Cross-lingual Strategy.** A cross-lingual strategy combines the advantages of the target-sided and the source-sided strategy. Instead of using an additional translation step before or after the disambiguation and clustering, the mapping between languages is incorporated into the disambiguation and clustering system. In the context of cross-lingual textual entailment, this approach is therefore also known as the integrated approach (Mehdad et al., 2012). To bridge between the languages, cross-lingual resources (e.g. mapped inventories, dictionaries, phrase tables, machine translation systems) and techniques (e.g. transliteration) can be applied on the fly during disambiguation and clustering. The cross-lingual strategy also allows to exploit cross-lingual information and to use information that has been extracted from the target language to process the source language.

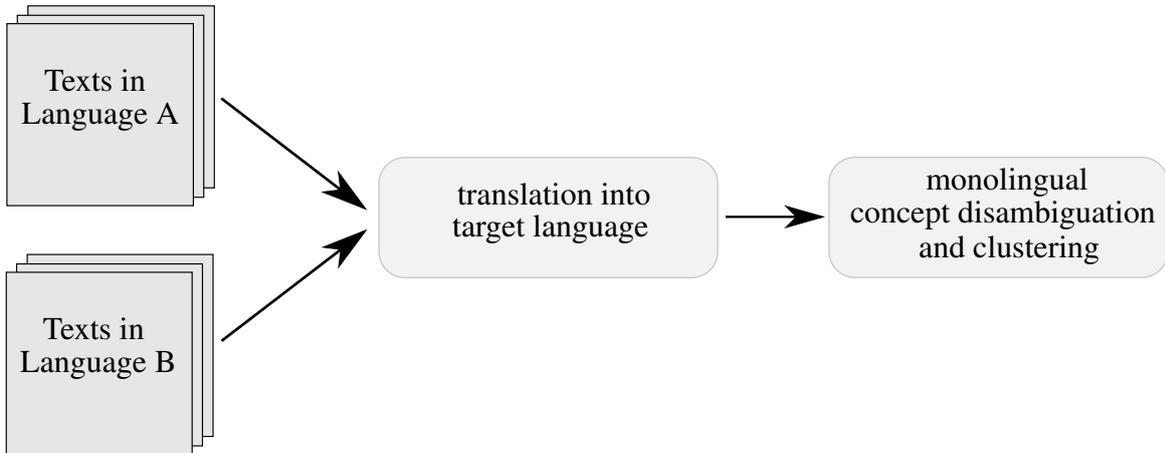
### 7.1.3 Discussion

Each strategy discussed in this section has advantages and disadvantages. Concerning the multilingual design of a system, a language-independent design is cost-effective, but may not lead to the best performance, as no language-specific knowledge is exploited. An approach that requires many adaptations is less cost-effective, but may lead to a higher performance, as it can exploit language-specific resources and information. In this thesis, we analyze to which degree our proposed approach is language-independent and discuss the adaptations that are necessary to port it to another language. In the experiment section, we will evaluate the influence of some language-specific adaptations to obtain a better idea about which language-specific adaptations tend to be more effective. Hence, our approach is between language-independent and adaptation-based. Concerning the cross-lingual strategy, we use a cross-lingual approach, as it allows us to share information across languages.

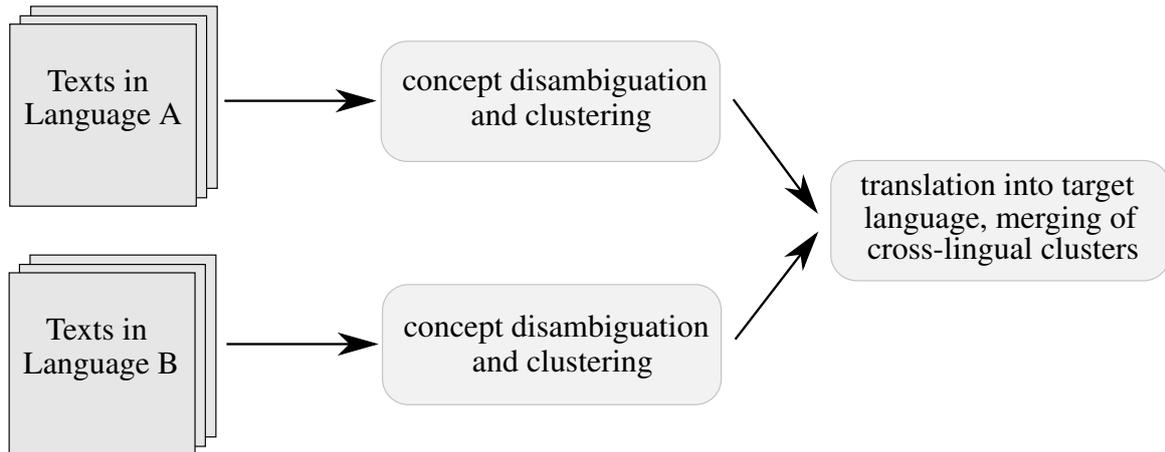
## 7.2 Scalability of our Concept Disambiguation and Clustering System across Languages

A concept disambiguation and clustering system that is scalable along the multilingual dimension can easily be applied to another language. A concept disambiguation and clustering system that is scalable along the cross-lingual dimension can easily be adapted for an inventory in another target language (disambiguation) and cross-lingual clustering. In the following, we discuss to which degree the different components of our system are language-independent and

**Target-sided Strategy**



**Source-sided Strategy**



**Cross-lingual Strategy**

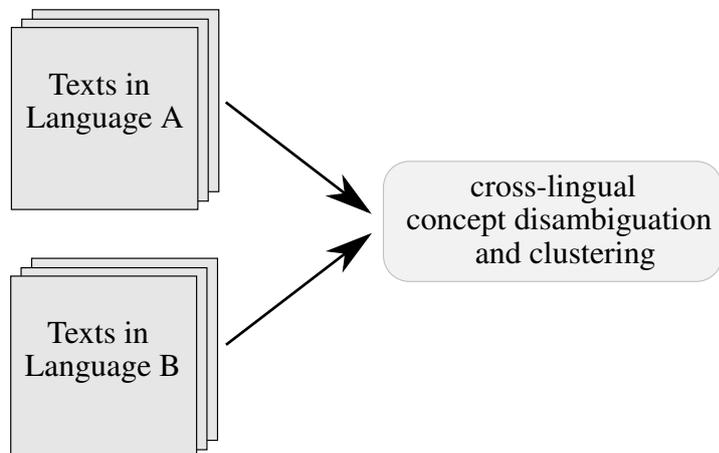


Figure 7.2: Strategies for cross-lingual concept disambiguation and clustering.

which adaptation might be required when it is ported to a new language. This analysis builds the basis for our multi- and cross-lingual concept disambiguation and clustering approach that is described in the next section.

### 7.2.1 Inventory

In this thesis, we assume that we can automatically derive an inventory for a language from Wikipedia. We further assume that the concepts are shared across languages and that we can map the inventories derived for different languages via the inter-language links in Wikipedia. If these assumptions are met, our approach is applicable to a language.

Although Wikipedia covers many languages – in March 2014, 286 languages were covered<sup>1</sup> –, the number of articles highly varies across languages. The number of concepts and lexicalizations that can be obtained from a language version heavily affects the performance our approach in terms of coverage. In addition, the number of internal hyperlinks in a language version influences the coverage and quality of the language-specific information that can be extracted and with it the quality of the overall approach.

### 7.2.2 Preprocessing

Our English monolingual approach relies on tokenization, lemmatization, sentence splitting, part-of-speech tagging, syntactic parsing and named entity recognition. This information is used during the mention recognition and the feature extraction (in particular, during the extraction of features for the scope assignment task).

This preprocessing makes our approach dependent on the availability of language-specific preprocessing tools. In this thesis, we do not discuss how to port the system to languages for which these preprocessing components are not available, because for the languages we investigate in this thesis all necessary preprocessing components are available. However, in the context of the EU project CoSyne, we ported a variant of our approach to other languages – including Bulgarian and Turkish – by only relying on tokenization and part-of-speech tagging.

### 7.2.3 Mention Recognition and Candidate Concepts Identification

Given that the input texts are preprocessed as described above, our mention recognition can be applied to any language. If in a language multi-word expressions are discontinuous, it would be beneficial to adapt the  $n$ -gram based rules of our mention recognition (Section 6.2), as they assume that mentions are continuous.

---

<sup>1</sup>Statistics from [http://meta.wikimedia.org/wiki/List\\_of\\_Wikipedias](http://meta.wikimedia.org/wiki/List_of_Wikipedias), 5.3.2014.

Our candidate concept identification is applicable to any Wikipedia language, as it mainly looks up the string in the lexicon derived from Wikipedia. As for most components, a language-specific tuning might improve its quality. For instance, in Chinese, strings in simplified Chinese could also be transformed to traditional Chinese and be looked up in both versions in the lexicon. In other languages, it might be helpful to remove the diacritics for the lexicon lookup.

## 7.2.4 Backbone of Our Approach

The formulas that define the tasks and described the interrelations between them build the backbone of our approach (Chapter 4). We assume that these interrelations (e.g. the interrelations between the disambiguation and the clustering task) are applicable to any language. This means the core of our approach is language-independent and does not need any adaptations.

## 7.2.5 Features

In contrast to our backbone, the features may incorporate some language-specific assumptions that might not be true for all languages or require some language-specific information. In Table 5.1, 5.2, 5.3 and 5.4, the resources we use to obtain the features in English are indicated. Most of them require information extracted from Wikipedia.

**Feature for Concept Disambiguation.** For the disambiguation, we can distinguish between features that are language-independent and features that require language-specific resources.

*Language-independent Features.* The features based on concept-level co-occurrence information (Table 5.1, 3, 4) can be considered as language-independent if we assume that concepts are shared across languages and the same concepts tend to co-occur across different languages. These features not even require any language-specific resources, as the concept-level co-occurrences extracted from English can be used.

*Language-dependent Features.* All features that are based on string-level information require some language-specific resources. The string edit distance that measures the distance between the concept name and the mention requires that the concept name is available in the respective language. Although it better suits languages that use the Latin alphabet, it may be still applicable to Chinese. While the string-level co-occurrence features and the prior probability (5.1, 1, 5, 6) require annotated data in the respective language, a high quality of the descriptors is necessary for the other string-level features (5.1, 7-9).

Overall, the features for concept disambiguation can be ported across languages. Although only the concept-level co-occurrence features are language-independent, all language-specific information that is required for the other features can be extracted from the Wikipedia dump in the respective language version. Thus, both our backbone and our features for concept

disambiguation are largely portable, although the size of Wikipedia in the respective language version affects the quality of the language-dependent features. For instance, if only a few hyperlinks are available, the string-level co-occurrence features become sparse.

**Features for Concept Clustering.** Our concept clustering features are largely scalable across languages, in particular across languages that use a Latin alphabet. However, for some features adaptations are required.

*Language-independent Features.* The string and head match features (Table 5.2, 10, 11) are portable to all languages for which the one concept per discourse hypothesis applies. The head match feature requires that the head of a mention can be identified, while the string match features may require lemmatization. However, in languages with no or only little morphology it can directly be applied to the tokens. The cross-document clustering feature uses a concept-based representation for the similarity calculation and is thus language-independent.

*Language-dependent Features.* The substring features for person names (Table 5.2, 12, 13) require that person names are identified. Dependent on the resources available for a certain language (e.g. person name lists), some adaptations might be required. The acronym feature is suitable for languages that use a Latin alphabet, but does for instance not scale to Chinese.

The features for the concept clustering are portable across languages, requiring only few adaptations.

**Features for Scope Assignment.** The features for the scope assignment task are much more language-dependent than the features for the concept disambiguation and concept clustering. They have been designed with English in mind and incorporate some language-specific knowledge such as the tendency of the subject in English sentences to be topical. To port these features to another language, some language-specific linguistic knowledge is required. Even if this language-specific knowledge is available, it is not clear if it is sufficient to adapt the English features or if some different or additional features are required. In terms of resources, a large – although unannotated – corpus is required to extract the *idf* scores (Table 5.3, 15). In addition, a resource for derivational families (Table 5.4, 27) and a resource containing focusing adverbs (Table 5.3, 23) is necessary. To obtain the syntactic information, the input texts need to be parsed with a syntactic parser.

We conclude that although the core of our scope-aware approach is language-independent, the features for the task require some language-specific knowledge to be ported.

## 7.2.6 Model Parameters

Our features for the concept disambiguation and clustering task can be ported from one language to another language and only require a few adaptations. As we do not use the string-level

features directly, but derive numeric scores, all our features are binary or numeric (between 0 and 1). We thus assume that the model parameters, i.e. the feature weights, are portable across languages and that no retraining is necessary. For the scope-aware approach, however, the model parameters need to be adapted, as the features may considerably change or have a different meaning depending on the respective language.

## 7.3 Multi- and Cross-lingual Concept Disambiguation and Clustering

The analysis in the last section reveals that the core of our approach is language-independent. Some features for the concept disambiguation and clustering are language-independent, while others require some minor adaptations or some language-specific resources. However, these language-specific resources can be harvested from Wikipedia in the respective languages, although the quality may vary across language versions. Our joint concept disambiguation and clustering approach is thus easily portable to other languages than English and does not require any adaptations of the feature weights. In contrast, its scope-aware extension is more difficult to port from English to another language, as it makes use of language-specific knowledge. In this thesis, we therefore only port the joint concept disambiguation and clustering approach to other languages and leave the adaptation of the scope-aware approach for future work. Hence, our multilingual approach for concept disambiguation and clustering corresponds to the English monolingual joint concept disambiguation and clustering approach with no scope information. In the following, we summarize the steps that are required for porting our approach to a new language (Section 7.3.1), before we describe the steps that are necessary for cross-lingual extensions (Section 7.3.2).

### 7.3.1 Multilingual Adaptation

To port our joint concept disambiguation and clustering system to a new language, four steps are required.

**Extension of the Inventory.** To obtain the inventory for a new language, the Wikipedia dump in this language needs to be downloaded<sup>2</sup> and all articles need to be extracted (Section 3.2.1). Each article corresponds to one concept. These concepts can be mapped to the English version by exploiting the cross-language links (Section 3.2.2). The lexicon can also be extracted from the Wikipedia dump (Section 3.2.1).

---

<sup>2</sup>The Wikipedia dumps can be downloaded from:  
<http://dumps.wikimedia.org/backup-index.html>, 25.3.2014.

**Extraction of Language-specific Information.** Concept-level co-occurrence information can be shared across languages. We use the concept-level co-occurrence information we extracted from the English Wikipedia for all languages (Section 3.2.1 and 3.3). As the English Wikipedia is bigger than all other language versions, we assume that the co-occurrence information we extract from it is more reliable than the one extracted from other language versions. In contrast to the concept-level co-occurrence information, the string-level co-occurrence information and the prior probabilities are language-specific and need to be extracted from the Wikipedia dump in the respective language version.

**Adaptation of the Preprocessing.** As our mention recognition currently relies on syntactic information and named entity recognition, the preprocessing components need to be adapted for the new language.

**Additional Language-specific Adaptations.** A few features may require some language-specific adaptations. In addition, the candidate concept identification step can be tuned for a certain language.

### 7.3.2 Cross-lingual Adaptation

To link to a Wikipedia-based inventory in another language than English only requires to extend the inventory with for the new target language (first step for the multilingual adaptation). No other steps are necessary. If texts in multiple languages need to be processed, the same steps as for a multilingual adaptation are required (Section 7.3.1).

## 7.4 Selection of Languages for the Evaluation

To evaluate the performance of the proposed approach, other languages than English need to be selected. Bender (2011, p. 10–19) discusses different criteria that should be considered when selecting languages to evaluate multilingual approaches. An important aspect is the *language family*. Highly related languages tend to behave similarly with respect to many linguistic factors. Thus, it is important to consider different language families. Other aspects that might affect the performance of our approach is the *availability of resources* (e.g. size of respective Wikipedia version from where we obtain the statistics) and the *writing system* (e.g. with respect to string match features). It is therefore preferable to select languages that show some variations with respect to these aspects. As we want to compare our approach to other multi- and cross-lingual approaches, our choice is also affected by the languages that have been used in previous work on concept disambiguation and clustering with Wikipedia. Given these consideration, we selected two languages to evaluate our multi- and cross-lingual

approach: *Spanish* and *Chinese*. While Spanish is a Romance language, Chinese belongs to the Sino-Tibetan language family. At the same time, the Spanish Wikipedia version is rather big, while the Chinese one is small. Finally, Chinese also uses another writing system than English and Spanish. By covering *English*, *Spanish* and *Chinese*, we also account the top three languages with respect to number of native speakers.<sup>3</sup> In addition, testing data is available for both Spanish and Chinese from the entity linking task at TAC that also allows us to compare our approach to related work.

To port our system to Spanish and Chinese we preceded as described above (Section 7.3.1). For the Spanish preprocessing, we used *FreeLing* (Padr3 & Stanilovsky, 2012), while we used the Chinese models from *Stanford’s CoreNLP* (Manning et al., 2014) for the Chinese preprocessing. Table 7.1 shows the few language-specific adaptations we made for the two languages. We reported some statics for the Spanish and Chinese inventory in Table 3.5 (Section 3.3).

Adapted Part	Description of Adaptation
<b>Spanish</b>	
Candidate concept identification	Besides the string variants we use for English, we additionally look up the string without any diacritics
Person name substring (Table 5.2, 12, 13)	We use the same resource as for English (Bergsma & Lin, 2006)
<b>Chinese</b>	
Candidate concept identification	We use both the traditional and the simplified version of the string. The mapping is done based on a table lookup <sup>4</sup>
Person name substring (Table 5.2, 12, 13)	To decide if a mention is a person name, we use the two resources: a name lexicon <sup>5</sup> and a name list extracted from <i>Baidu Baike</i> <sup>6</sup>
Acronym feature (Table 5.2, 11)	The acronym feature is not used for Chinese

Table 7.1: Language-specific adaptations for Spanish and Chinese.

In addition, we also evaluate the capability of our approach to link from English to an inventory in another language. We link from English to *Japanese*, English to *Korean* and English to *Chinese*. For these three languages, a shared task has been organized at NTCIR 9, which allows us to compare our results to the state of the art for this setting. We processed the

<sup>3</sup><http://www.ethnologue.com/statistics/size>, July 2015.

<sup>4</sup>The table is obtained from here: <http://ishare.iask.sina.com.cn/f/6514604.html?retcode=6102>, July 2011.

<sup>5</sup>The lexicon is obtained from here: <http://ishare.iask.sina.com.cn/f/15763907.html>, July 2011.

<sup>6</sup><http://baike.baidu.com>, July 2012.

corresponding dumps as described above (Section 7.3.2). The statistics for the mappings are provided in Table 3.4 (Section 3.3).

# Chapter 8

## Data

The proposed approach for multi- and cross-lingual concept disambiguation and clustering is supervised. To train the model and to develop and tune the features some annotated data is required. At the same time, annotated data is essential to evaluate the proposed approach and to compare it to the state of the art (Chapter 9). In this chapter, we present the data we use for training, development and testing.

Over the years, a lot of research has been conducted and many shared tasks have been organized in the field of concept disambiguation and clustering focusing on different aspects and targeting different applications (Chapter 10). These research activities have led to a wide range of data sets that are annotated for different aspects reflecting the respective research objectives. For instance, some data sets are annotated to benchmark concept disambiguation systems, others to evaluate concept clustering approaches; some data sets are monolingual, others have been constructed to assess cross- or multi-lingual methods; some data sets only contain annotations for proper names, in others common nouns and proper names are annotated.

To obtain a complete picture of the performance of our general-purpose, multi- and cross-lingual concept disambiguation and clustering system and to show its application range, we consider multiple data sets. However, to keep the evaluation concise, we refrain from evaluating on all available data sets and restrict ourselves to a selection.

In the following, we first discuss the general data situation in concept disambiguation and clustering and present our criteria to select the data sets (Section 8.1.2). We then describe our training (Section 8.2), development (Section 8.3) and testing data (Section 8.4) in more detail.

### 8.1 Blazing a Trail Through the Data Jungle

Since quantitative evaluation strategies have been established in computational linguistics in the nineties, many data sets have been released in the field of concept disambiguation and

clustering. Shared tasks, but also research activities in the information retrieval community, have further boosted the creation of data sets for these tasks.

A first step to blaze a trail through this data jungle is to restrict ourselves to data sets that are annotated with Wikipedia concepts. Wikipedia is not only our selected inventory for concept disambiguation, but Wikipedia-like concepts are also the target of our concept clustering approach. As several resources (e.g. *Yago* or *FreeBase*) have been built based on Wikipedia or are interlinked with it, we include corpora annotated with links to these resources.

Table 8.4 summarizes some key aspects of data sets that are annotated with Wikipedia or at least Wikipedia-like concepts.<sup>1</sup> As this pool is still quite big, we have to further restrict ourselves. In the following, we present the criteria we used to select a representative sample of data sets (Section 8.1.1) and present the final selection (Section 8.1.2). All data sets that we selected are marked by an asterisk (★) in Table 8.4 and are described in more detail in Section 8.4.

### 8.1.1 Selection Criteria

Besides the inventory, we consider several other criteria to restrict our pool of data sets used to evaluate our approach (Chapter 9). A careful selection is necessary to provide a concise, but still extensive evaluation of our approach. In the following, we discuss our criteria to select data sets.

**Target.** Our proposed approach is a disambiguation and clustering approach. We should evaluate both the disambiguation and the clustering. As we consider both common nouns and proper names, data sets that are annotated for both are preferable.

**Language Coverage.** As our approach is multi- and cross-lingual, it also needs to be evaluated on different languages. Hence, we require some multi- and cross-lingual data sets that allow us to compare our approach to the state of the art.

**Comparison to the State of the Art.** To compare to the state of the art and related work, data sets that have been established as a benchmark are important. We focus on data sets that originate from shared tasks (e.g. TAC, NTCIR data) or that have been used to evaluate closely related approaches (e.g. ACE 2004).

**Annotation Strategy.** While some data sets are annotated independently of any systems (e.g. ACE 2005), others result from a post-hoc evaluation of a specific system output (e.g.

---

<sup>1</sup>A few data sets for other inventories are listed in Table 3.2.

MSNBC, Aquaint). In the latter case, the annotations might be biased towards the respective systems. We give preference to data sets that are manually annotated based on guidelines.

**Quality of the Annotations.** Only if the quality of the annotations is high the evaluation is meaningful. One quality measure is the inter-annotator agreement. We consider both the inter-annotator agreement and the clearness of the guidelines if available for selecting the data sets. For instance, the mention definitions that have been used to annotate the ITB2 and the NewsSc data sets are fuzzy. We thus do not use these two data sets for evaluation in this thesis.

**Size.** While some data sets only consist of a few hundred mentions (e.g. MSNBC, Aquaint), others comprise thousands of mentions (e.g. ACE 2005). We exclude data sets that only consist of a few hundred mentions except they have been used to evaluate other closely related approaches (ACE 2004). At the same time, we exclude data sets with millions of mentions, as the focus of this thesis does not lie on aspects related to scalability.

**Text Sort.** As we aim for a general purpose disambiguation and clustering approach it is also important to cover different text sorts including news paper texts and web pages. We do not consider artificially modified data sets such as the KORE data sets where parts of person names are removed to increase the ambiguity. Our aim is to evaluate our systems on texts that have been created without having any evaluation task in mind. In addition, we exclude Twitter data, as tweets are extremely short and might require a tailored approach (Abel et al., 2011; Cassidy et al., 2012; Meij et al., 2012; Guo et al., 2013a; 2013b; 2013c; Habib et al., 2014; Chang et al., 2014).

### 8.1.2 Data Set Selection

Based on the criteria introduced above, we selected the following data sets for evaluation: ACE 2005, ACE 2004, CoNLL 2003, the TAC data sets and the NTCIR 9 data sets. Each of these data sets allows us to evaluate different aspects of our system and to compare our approach to related work. The data we used for training and development is derived from the internal hyperlinks in Wikipedia. Hence, we only require data that has been deliberately annotated for concept disambiguation and clustering during testing, but not during training and development.

## 8.2 Training Data

We derived our training data from the internal hyperlinks in Wikipedia. Harvesting training data from internal hyperlinks is a cheap strategy to obtain training data. The internal hyper-

links are manually created by Wikipedia editors to enhance the navigation in Wikipedia and not to provide concept annotations to train or test a system. Hence, while the annotations obtained from the internal hyperlinks come for free, they may contain some noise and biases. For instance, in Wikipedia, mainly keywords or terms that need further explanations are linked to other articles. Typically, not each occurrence of a term or a word is linked in an article, but only a few occurrences (often only the first occurrence). According to Csomai & Mihalcea (2008) only 6% of the tokens in a Wikipedia article are linked. Another bias is introduced by the text sort. Wikipedia articles are encyclopedic descriptions and not news or web texts which are more likely to be the target text sort of a disambiguation and clustering system.

Despite these biases, internal hyperlinks have been successfully used to train concept disambiguation systems (Bunescu & Paşca, 2006; Csomai & Mihalcea, 2008; Milne & Witten, 2008b) and even to evaluate disambiguation systems (Cucerzan, 2007; Ratinov et al., 2011; Ferragina & Scaiella, 2012). In this thesis, we use Wikipedia articles for the training and development of our system, while we use mainly other sources than Wikipedia articles for the evaluation.

Following Milne & Witten (2008b), we use 500 Wikipedia articles for training. In contrast to them, we only consider articles that are marked as *featured*<sup>2</sup>. *Featured Wikipedia articles* are supposed to be of high quality and have been successfully used in other natural language processing applications (Strötgen et al., 2010). We assume that the internal hyperlinks in these articles are of high quality. In the English Wikipedia dump used in this thesis,<sup>3</sup> we identified 3,403 featured articles. Out of them, we randomly selected 500 articles for training. After the text cleaning – we stripped away all markup except the internal hyperlinks –, we preprocessed the texts using the same preprocessing pipeline as for the testing data (Section 6.1). We identified the mentions using our mention recognition and their candidate concepts. We then extracted all mentions that match an internal hyperlink. The article the internal hyperlink points to serves as our gold concept (Figure 3.1, Section 3.1.1). From these mentions we derived our training instances. Further, we added all mentions that have the same lemma as exactly one mention in the mention pool (or multiple if they all point to the same Wikipedia article) and annotated them with the concepts of the corresponding mentions. Using this technique, we compensate for the fact that usually only a few occurrences of a term are hyperlinked in a Wikipedia article.

To obtain some NILs, we experimented with different techniques. For instance, we tried to randomly remove some concepts from Wikipedia and consider all the mentions that point to them as NILs. In our final version, we refrained from removing concepts and considered a mention as a NIL if its corresponding concept is not among the candidates of any mention in the document. We also experimented with different numbers of training instances. However,

<sup>2</sup>[http://en.wikipedia.org/wiki/Wikipedia:Featured\\_articles](http://en.wikipedia.org/wiki/Wikipedia:Featured_articles), 3.7.2014.

<sup>3</sup>We used the English Wikipedia dump from the 4th of January 2012.

Data set	Number of Mentions	Non-NILs	NILs	Average Ambiguity
Training Data	56,372	53,097	3,275	2.31
Development Data	9,992	9,375	617	2.28

Table 8.1: Statistics for training and development data derived from Wikipedia.

500 Wikipedia articles turned out to be a reliable sample size. If less training data is used (e.g. only 200 Wikipedia articles), the learned weights can substantially vary across different article samples.

Regarding our scope-aware approach, we were concerned that mentions with local scope are underrepresented in our training data, so that no model can be learned for them. This concern turned out to be wrong as also less prominent mentions can be linked in a Wikipedia article. However, mentions with intermediate and global scope are more frequent in our training and development data<sup>4</sup> than e.g. in the ACE 2005 data. While on the development data 72% mentions were assigned intermediate or global scope by our system, this is the case for less than 40% of the mentions in ACE 2005 (Figure 9.3, Section 9.4.2).

Table 8.1 summarizes the number of mentions, the number of Non-NILs and the number of NILs in our training set. In addition, we report the average number of candidate concepts per mention (*average ambiguity*) which is determined by our candidate identification strategy.

### 8.3 Development Data

Our development data is obtained in the same way as our training data. We randomly selected 100 articles out of the 3,403 featured articles (excluding all articles that were chosen for training). Then we processed these articles in the same way as our training articles (Section 8.2). The statistics for the development data are given in Table 8.1.

In addition, we downloaded from time to time a few articles from BBC news.<sup>5</sup> Although these news articles do not contain any annotations, we could still use them to develop features and to perform a linguistic analysis.

### 8.4 Testing Data

In this section, we describe the data sets our system is evaluated on (Chapter 9). We first focus on the English monolingual testing sets (Section 8.4.1), before we describe the multi-

<sup>4</sup>Our development data is also derived from Wikipedia (Section 8.3).

<sup>5</sup><http://www.bbc.com/news/>, 3.7.2014.

and cross-lingual data sets (Section 8.4.2). Table 8.4 summaries key aspects of the data sets (such as size, mention definition and inter-annotator agreement).

### 8.4.1 Monolingual Data Sets

We evaluate our English monolingual system on the ACE 2005, the ACE 2004, the TAC and the CoNLL data sets. We describe all of them in this section.

**ACE 2005.** The English part of the ACE 2005 data set<sup>6</sup> has been manually annotated with links to the English Wikipedia by Bentivogli et al. (2010).<sup>7</sup> Intending to add an additional annotation layer to an existing annotated corpus, Bentivogli et al. (2010) manually disambiguated the syntactic head of non-pronominal ACE mentions. Hence, this data set contains annotations for both common nouns (*NOM*) and proper names (*NAM*). The ACE mentions comprise mentions that refer to a facility (e.g. an airport or a plant), a geo-political entity (e.g. a nation or a province), a location (e.g. an address or a boundary), an organization (e.g. a commercial organization or a government), a person (e.g. a group or an individual), a vehicle (e.g. an air plane or a car) or a weapon (e.g. a biological weapon).<sup>8</sup> As an inventory, the on-line version of the English Wikipedia from 2010 was used (Bentivogli et al., 2010). NILs are marked as NILs, but they are not clustered. Thus, the ACE 2005 data set allows us to evaluate the capability of our system to disambiguate common nouns and proper names, but it is not suitable to evaluate the performance of the clustering.

In total, the ACE 2005 data set comprises 597 articles from different sources including newswire reports, broadcast news, internet sources and transcribed audio data. Overall, 29,300 mentions (15,242 common nouns and 14,058 proper names) are annotated out of which 2,116 are NILs. Some mentions are annotated with more than one concept. For instance, *president* may be annotated with the concept PRESIDENT OF THE UNITED STATES and PRESIDENT. We consider a mention as correctly disambiguated if any of the annotated concepts has been selected by the system. The agreement between two annotators is estimated based on a subset of the data set and is analyzed in Bentivogli et al. (2010). The Dice coefficient for the concepts is 0.85 and 0.94 before and after reconciliation respectively.

**ACE 2004.** Ratnov et al. (2011) annotated a subset of the ACE 2004 data sets with links to Wikipedia using the *Amazon Mechanical Turk*. The annotations obtained from the Amazon

---

<sup>6</sup>The ACE 2005 data set can be obtained from LDC: <https://catalog.ldc.upenn.edu/LDC2006T06>, 4.7.2014.

<sup>7</sup>The annotations can be obtained from here: [http://www.celct.it/pageReader.php?id\\_page=34](http://www.celct.it/pageReader.php?id_page=34), 4.7.2014.

<sup>8</sup><http://www.itl.nist.gov/iad/mig/tests/ace/2005/doc/ace05-evalplan.v2a.pdf>, 4.7.2014.

Mechanical Turk were then manually corrected (Ratinov et al., 2011). The data set contains annotations for common nouns and proper names.

In total, 306 mentions in 36 texts are annotated out of which 49 are NILs. NILs are marked as such, but they are not clustered. Thus, as the ACE 2005 data set, the ACE 2004 data set allows us to evaluate the performance of our disambiguation system, but not the performance of our clustering approach.

Although this data set is small, it still allows us to compare our approach to the approach of Cheng & Roth (2013) which is closely related to ours. We mainly use the ACE 2004 data set for this comparison and to a lesser extent to analyze the contributions of different parts of our system.

**English TAC Data Sets (TAC EN Test 2009-2013).** The entity linking task at TAC is the most popular shared task for proper name disambiguation with Wikipedia (McNamee & Dang, 2009; Ji et al., 2010; 2011). In 2013, 27 systems participated in this shared task. Given a proper name mention in a text, the task is to identify its corresponding entry in an inventory derived from the English Wikipedia.<sup>9</sup> In case the mention does not correspond to any entry in the inventory, it must be assigned a NIL tag. The English monolingual entity linking task has been organized since 2009 every year by the *National Institute for Standards and Technology* at the *Text Analysis Conference (TAC)*. Since 2011, the participating systems are not only required to recognize the NILs, but also to cluster them accordingly. Hence, the testing data from the English entity linking task at TAC 2011, 2012 and 2013 allow us to evaluate both our disambiguation and our clustering component and to compare them to the state of the art. We mainly focus on the data sets from TAC 2011-2013, but also report our results for the TAC 2009 and 2010 data sets.

In contrast to the ACE 2005 and ACE 2004 data sets, the TAC data sets – provided by LDC – are restricted to proper names (persons, organizations and geopolitical entities) (Li et al., 2011; Ellis et al., 2012; 2013). The participating systems are provided with the mentions and do not need to recognize them. Another major difference between the ACE 2005 and 2004 data sets and the TAC data sets is that in the TAC data sets only a few challenging mentions are annotated per text. The annotators for the TAC data sets are asked to select challenging proper names so that both the variability and the ambiguity in the annotated mention pool are high. In addition, the distribution over entity types (persons, organizations, geopolitical entities) is balanced. Approximately 50% of the mentions in the TAC data sets have a corresponding entry in the inventory (Non-NILs), while the other 50% are NILs. The inter-annotator agreement among three annotators is 90.56% for the TAC 2010 data sets (Ji et al., 2010). For the TAC 2011 data set, humans achieve an accuracy of 90.2% (Ji et al., 2011).

<sup>9</sup>The inventory is based on the English Wikipedia dump from October 2008.

Data Set	Number of Men- tions	Number of Spanish Mentions	Number of English Mentions
TAC ES 2012	2,066	1,991	75
TAC ES 2013	2,117	1,530	587

Table 8.2: Number of mentions per language for the Spanish cross-lingual data sets.

The texts for the TAC data sets are drawn from newswire reports, web documents and – in the 2013 data set – from discussion forums.

**CoNLL 2003.** Recently, the English part of the CoNLL 2003 data sets has become popular to evaluate proper name disambiguation systems. While this data set has been originally created to benchmark named entity recognition systems (Tjong Kim Sang & De Meulder, 2003), Hoffart et al. (2011b) extended the existing annotations with links to *YAGO2* (Hoffart et al., 2011a). As *YAGO2* is derived from the English Wikipedia<sup>10</sup>, also links to the English Wikipedia are available.

The CoNLL 2003 data set consists of 1,393 Reuters newswire articles and is split into a training (946 files), development (216 files) and testing part (231 files). It is annotated for proper names (including persons, locations, organizations and proper names of type miscellaneous). NILs are marked as such, but they are not clustered. Hence, the CoNLL 2003 data set allows us to evaluate the capability of our system to disambiguate proper name mentions. In this thesis, we mainly focus on the TAC data sets to evaluate this capability of our system, as these data sets also contain clustering information. However, we also report our results on the CoNLL 2003 data set, as it contains annotations for all proper names and not only for a few challenging mentions as the TAC data sets. We use all three parts (training, development and testing part) for testing.

## 8.4.2 Multi- and Cross-lingual Data Sets

To evaluate the multi- and cross-lingual performance of our system, we use the data sets from the Spanish and Chinese cross-lingual entity linking task at TAC and the data sets from the cross-lingual link discovery task at NTCIR 9.

**Spanish Cross-lingual TAC Data Sets (TAC ES Test 2012-2013).** In 2012 and 2013, a Spanish cross-lingual entity linking task has been organized at TAC. The setup is similar as in the monolingual English entity linking task (Section 8.4.1) and comprises the disambiguation and clustering of proper names including persons, organizations and geopolitical entities.

<sup>10</sup>It is derived from the English Wikipedia dump from the 17th of August 2010.

In contrast to the English monolingual entity linking task, most texts are in Spanish and only a few texts are in English. Both Spanish and English mentions are required to be linked to the same inventory derived from the English Wikipedia<sup>11</sup> or – in case they are NILs – are required to be clustered across languages. As in the English monolingual task, only a few challenging mentions are annotated (Ellis et al., 2012; 2013). The texts used in the Spanish cross-lingual entity linking task are mainly newswire reports. Only a few English mentions are from web texts.

The two data sets from the Spanish cross-lingual entity linking task allow us to evaluate both the multi- and cross-lingual performance of our system. It is necessary to process both Spanish and English input texts and to link and cluster mentions across languages. In addition, we can compare our results to the results of the systems that participated in the Spanish cross-lingual entity linking task at TAC 2012 and TAC 2013. Table 8.2 shows the number of mentions per language for each of the two data sets. All other statistics are reported in Table 8.4.

**Chinese Cross-lingual TAC Data Sets (TAC ZH Test 2011-2013).** Besides the Spanish cross-lingual entity linking task, a Chinese cross-lingual entity linking task has been organized at TAC 2011 (Ji et al., 2011), TAC 2012 (Ellis et al., 2012) and TAC 2013 (Ellis et al., 2013). The setup is exactly the same as in the Spanish cross-lingual entity linking task with the only exception that the texts and mentions are in Chinese and English instead of in Spanish and English. Hence, the Chinese cross-lingual entity linking task comprises the disambiguation and clustering of proper names (persons, organizations and geopolitical entities). The inventory is the same as in all other entity linking tasks at TAC and derived from the English Wikipedia.<sup>12</sup> While all Chinese texts in the *TAC ZH Test 2011* data set are newswire reports, the Chinese texts in the *TAC ZH Test 2012* and the *TAC ZH Test 2013* data sets also include web texts. The English texts comprise newswire reports and web texts in all three Chinese cross-lingual data sets.

Together with the Spanish cross-lingual data sets, the Chinese cross-lingual data sets allow us to evaluate the portability of our system to two other languages that are not related to each other. Table 8.3 shows the number of mentions per language for all three data sets. All other statistics are reported in Table 8.4.

**NTCIR 9 Data Sets (NTCIR 9 EN-JA Test, NTCIR 9 EN-KO Test, NTCIR EN-ZH Test).** The cross-lingual link discovery task at NTCIR 9 aims to evaluate systems that insert cross-lingual hyperlinks in texts (Tang et al., 2011). Given some English texts, systems are required to identify keywords in these texts and link them to their corresponding articles in the Japanese,

<sup>11</sup>The inventory is based on the English Wikipedia dump from October 2008.

<sup>12</sup>The inventory is based on the English Wikipedia dump from October 2008.

<b>Data Set</b>	<b>Number of Men- tions</b>	<b>Number of Chinese Mentions</b>	<b>Number of English Mentions</b>
<b>TAC ZH 2011</b>	2,176	1,481	695
<b>TAC ZH 2012</b>	2,122	1,429	693
<b>TAC ZH 2013</b>	2,155	1,640	515

Table 8.3: Number of mentions per language for the Chinese cross-lingual data sets.

Korean and Chinese Wikipedia. As we derive our inventory from Wikipedia, we can formulate this task as a cross-lingual disambiguation problem from English to an inventory in another language than English. Hence, the NTCIR data sets allow us to evaluate the performance of our system to link English mentions to an inventory in another language than English, while the TAC data sets allow us to evaluate the performance of our system to link from another language than English to an English inventory. In contrast to the TAC tasks, the NTCIR tasks require the systems to identify keywords. For each text, at most 250 keywords are allowed.

In total, the NTCIR 9 data sets comprise an English to Japanese (*NTCIR 9 EN-JA Test*), an English to Korean (*NTCIR 9 EN-KO Test*) and English to Chinese (*NTCIR 9 EN-ZH Test*) data set. All three data sets use the same 25 English Wikipedia articles as source texts. However, the gold annotations are in Japanese, Korean and Chinese respectively. The inventories are obtained from the Japanese<sup>13</sup>, Korean<sup>14</sup> and Chinese Wikipedia<sup>15</sup>. The gold annotations for the NTCIR data sets are derived from the internal hyperlinks in Wikipedia. For each English article out of the 25 English articles, all internal hyperlinks are obtained. For each of these hyperlinks, it is checked if the article it points to also exists in the target language. If this is the case, the article in the target language is added to the ground truth for this target language. In addition, the counter parts of the 25 English articles are obtained in the respective target language and all articles they linked to are added to the ground truth for the respective target language. As these annotations are on the document and not on the mention level, we can only report document-level scores for the NTCIR 9 data sets. At the shared task, the system outputs were also manually evaluated. However, the annotations from the post-hoc evaluation are of limited usage as they are strongly biased to the system outputs. We do not use them in this thesis.

At NTCIR 10, a second edition of the cross-lingual link discovery shared task took place. While we participated in the shared task at NTCIR 9, we did not participate in its second edition and did not evaluate our system on these data sets.

<sup>13</sup>The Japanese inventory is derived from the Japanese Wikipedia dump from the 24th of June 2010.

<sup>14</sup>The Korean inventory is derived from the Korean Wikipedia dump from the 28th of June 2010.

<sup>15</sup>The Chinese inventory is derived from the Chinese Wikipedia dump from the 27th of June 2010.

## 8.5 Summary

In this section, we discussed the training, development and testing data used in the context of this thesis. For training and development, we exclusively rely on the internal hyperlinks that have been created by Wikipedia editors for navigation purposes. Hence, although our system is supervised, no additional manual annotation efforts are thus required to tune our system. If an application requires a high performance for one specific domain, we could even extract a domain-specific corpus from Wikipedia (e.g. sports or chemistry) and tune the parameters on this domain-specific part of Wikipedia.

To evaluate our system, we selected different data sets. Many data sets are annotated for the tasks addressed in this thesis, allowing us to evaluate different facets of our approach. As they are annotated with concepts from different Wikipedia versions, we need to map between the concepts of different Wikipedia versions. Such a mapping is not always possible, as Wikipedia articles can be deleted or merged over time. However, these effects are rather small ( $< 1\%$ ). We consider mentions that are linked to concepts that do not exist in our Wikipedia version as wrongly disambiguated in our experiments independently of how we disambiguate them.

Table 8.4 summarizes the most prominent data sets annotated with Wikipedia concepts including some prominent data sets we do not consider in this thesis.

Data Set	Task			Language Information			Lang.	Mention Information		Corpus Information		Inventory	Annotation Information		Usage
	Name	Dis-ambig.	NIL Recogn.	Clust.	Mono-ling.	Multi-ling.		Cross-ling.	Definition	Tokens	Source		Texts	Version	
ACE 2005* (Bentivogli et al., 2010)	×	×		×			EN	ACE mentions (common nouns and proper nouns)	29,300 92.8% in KB 7.2% NILs	Broadcast news, newspapers, newswire reports, internet sources, transcribed audio data	597	Online version of Wikipedia 2010 (February - April; August)	Annotated by humans, partly by two annotators	0.85 (Dice coefficient with respect to annotated concepts and entities; before reconciliation)	No
ACE 2004* (Ratinov et al., 2011)	×	×		×			EN	ACE mentions (common and proper nouns)	306 (84.0% in KB, 16% NILs)	Newswire, broadcast news	36	Wikipedia 2011	Mechanical turk, only first mention in coreference chain is annotated	0.85 (agreement, then corrected)	No
IITB (Kulkarni et al., 2009)	×	×		×			EN	As much as possible, identified by people (including common and proper nouns)	17,200 (60% in KB; 40% NILs)	Collection of web pages (sports, entertainment, science and technology, health)	107	Wikipedia dump from August 2008	annotated by humans, partly by two annotators; candidate mentions and tokens were suggested by the system	0.80 (agreement)	No
NewsSc (Turdakov & Lizorkin, 2009)	×	×		×			EN	Identified by humans (as many as possible, including common and proper nouns)	8,236 (80.6% in KB, 19.4% NILs)	News articles, scientific papers	131	Wikipedia dump from October 2008	Annotated by humans	n.a.	No
MSNBC (Cucerzan, 2007)	×	×		×			EN	Proper names recognized by a system	756 (83.2% in KB, 16.8% NILs)	MSNBC news (Business, US politics, Entertainment, Health, Sports, Tech & Science, Travel TV news, U.S. News, World News)	20	Wikipedia version from the 11.9.2006	Post-hoc evaluation of system output	n.a.	No

Data Set	Task			Language Information			Lang.	Mention Information		Corpus Information		Inventory	Annotation Information		Usage
	Name	Dis-ambig.	NIL Recogn.	Clust.	Mono-ling.	Multi-ling.		Cross-ling.	Definition	Tokens	Source	Texts	Version	Strategy	Agreement
<b>CoNLL 2003*</b> (Hoffart et al., 2011b)	×	×					EN	Proper names (CoNLL shared task)	34,956 (79.6% in KB, 20.4% NILs)	Reuters newswire data	1,393	Yago2, derived from Wikipedia dump from 17.8.2010	Manually annotated by two students	78.9% (agreement before reconciliation)	No
<b>KORE-50</b> (Hoffart et al., 2012)	×	×					EN	Proper names	148 (97.3% in KB, 2.7% NILs)	Sentences, no documents (politics, sports, celebrities, music, and business); handcrafted to be difficult for disambiguation	50	Wikipedia and YAGO2	Manually annotated	n.a.	No
<b>AIDA-EE</b> (Hoffart et al., 2014)	×	×					EN	Proper nouns	9,976 (94.4% in KB, 5.6% NILs)	Gigaword5	300	Wikipedia (17.8.2010)	Manually annotated	n.a.	No
<b>AQUAINT</b> (Milne & Witten, 2008b)	×						en	Keywords (including common and proper nouns)	727 (no NILs)	Newswire stories (part of AQUAINT corpus, Associated Press)	50 (250-300 tokens)	Wikipedia version from 20.11.2007	Post-hoc evaluation of system output (Milne & Witten, 2008b) using Mechanical Turk; extended with missing important mentions	97% (2 out of 3 annotators agreed)	No
<b>WikiLinks</b> (Singh et al., 2012)	×						EN	Keywords	40M	Web pages	10M	Wikipedia (different versions)	Annotated by web page editors: links on web pages to Wikipedia	n.a.	No

Data Set	Task			Language Information			Lang.	Mention Information		Corpus Information		Inventory	Annotation Information		Usage
	Dis-ambig.	NIL Recogn.	Clust.	Mono-ling.	Multi-ling.	Cross-ling.		Definition	Tokens	Source	Texts	Version	Strategy	Agreement	Shared Task
<b>CELC</b> (Mayfield et al., 2011)	×	×				×	Target: EN Source: AR, ZH, DA, NL, FI, IT, PT, SV, CS, FR, DE, ES, SQ, BG, HR, EL, MK, RO, SR, TR, UR	Person names	55,157 for all lan- guages (53.5% in KB, 46.5% NILs; avg. 2,627 per lan- guage)	Parallel data (including Europarl, ProjSynd, SETimes)		TAC KB (Wikipedia dump from October 2008)	Semi- automatical annotation in English; projection to other languages; correction using crowd sourcing	Quality control for different annotation phases	No
<b>Task 12, SemEval-13</b> (Navigli et al., 2013)	×				×		EN, FR, DE, IT, ES	Common nouns and proper names	EN: 1,242 FR: 1,039 DE: 1,156 IT: 1,977 ES: 1,103 (only links to Wikiped- ia)	Texts from a workshop on statistical machine translation; domains range from sports to financial news	13 texts for each lan- guage (par- allel cor- pus)	BabelNet 1.1.1 (with some links to Wikipedia)	English texts are manually annotated; then these annotations are projected to other languages; projected annotations are corrected and missing nouns are manually annotated; final quality check	Final quality check lead to corrections for about 5% of the instances	Yes
<b>TAC EN Test 2009*</b> (McNamee & Dang, 2009)	×	×		×			EN	proper names (persons (16.1%), organizations (69.4%), geo-political entities (14.5%))	3,904 (42.9% in KB, 57.1% NILs)	newswire texts	3,688	TAC KB (based on English Wikipedia from October 2008)	Annotated by humans (LDC)	n.a.	Yes

Data Set	Task			Language Information			Lang.	Mention Information		Corpus Information		Inventory	Annotation Information		Usage
	Name	Dis-ambig.	NIL Recogn.	Clust.	Mono-ling.	Multi-ling.		Cross-ling.	Definition	Tokens	Source		Texts	Version	
<b>TAC EN Test 2010*</b> (Ji et al., 2010)	×	×					EN	Proper names (persons (33.3%), organizations (33.3%), geo-political entities (33.3%))	2,250 (45.3% in KB, 54.7% NILs)	newswire data (66.7%) web data (33.3%)	2,231	TAC KB (based on English Wikipedia from October 2008)	Annotated by humans (LDC)	90.56% (inter-annotator agreement among three annotators)	Yes
<b>TAC EN Test 2011*</b> (Ji et al., 2011; Li et al., 2011)	×	×	×				EN	Proper names (persons (33.3%), organizations (33.3%), geo-political entities (33.3%))	2,250 (50.0% in KB, 50.0% NILs)	Newswire data (66.3%) web data (33.7%)	2,231	TAC KB (based on English Wikipedia from October 2008)	annotated by humans (LDC)	0.902 (accuracy of human)	Yes
<b>TAC EN Test 2012*</b> (Ellis et al., 2012)	×	×	×				EN	Proper names (persons (41.2%), organizations (31.8%), geo-political entities (27.0%))	2,226 (52.9% in KB, 47.1% NILs)	Newswire data, web data	2,016	TAC KB (based on English Wikipedia from October 2008)	Annotated by humans (LDC)	n.a.	Yes
<b>TAC EN Test 2013*</b> (Ellis et al., 2013)	×	×	×				EN	Proper names (persons (31.3%), organizations (32.0%), geo-political entities (36.7%))	2,190 (49.8% in KB, 50.2% NILs)	Newswire data, web data, discussion forums	1,820	TAC KB (based on English Wikipedia from October 2008)	Annotated by humans (LDC)	n.a.	Yes
<b>TAC ZH Test 2011*</b> (Ji et al., 2011; Li et al., 2011)	×	×	×			×	Target: EN Source: ZH, EN	Proper names (persons (37.9%), organizations (32.6%), geo-political entities (29.5%))	2,176 (49.9% in KB, 50.1% NILs)	Newswire data	2,167	TAC KB (based on English Wikipedia from October 2008)	annotated by humans (LDC)	n.a.	Yes
<b>TAC ZH Test 2012*</b> (Ellis et al., 2012)	×	×	×			×	Target: EN Source: ZH, EN	Proper names (persons (32.9%), organizations (33.8%), geo-political entities (33.2%))	2,122 (58.4% in KB, 41.6% NILs)	Newswire data, web data	2,117	TAC KB (based on English Wikipedia from October 2008)	Annotated by humans (LDC)	n.a.	Yes

Data Set	Task			Language Information			Lang.	Mention Information		Corpus Information		Inventory	Annotation Information		Usage
	Name	Dis-ambig.	NIL Recogn.	Clust.	Mono-ling.	Multi-ling.		Cross-ling.	Definition	Tokens	Source		Texts	Version	
<b>TAC ZH Test 2013*</b> (Ellis et al., 2013)	×	×	×			×	Target: EN Source: ZH, EN	Proper names (persons (32.8%), organizations (34.1%), geo-political entities (33.1%))	2,155 (56.0% in KB, 44.0% NILs)	Newswire data, web data	2,143	TAC KB (based on English Wikipedia from October 2008)	Annotated by humans (LDC)	n.a.	Yes
<b>TAC ES Test 2012*</b> (Ellis et al., 2012)	×	×	×			×	Target: EN Source: ES, EN	Proper names (persons (32.4%), organizations (26.1%), geo-political entities (41.5%))	2,066 (44.7% in KB, 55.3% NILs)	Newswire data, web data	1,979	TAC KB (based on English Wikipedia from October 2008)	Annotated by humans (LDC)	n.a.	Yes
<b>TAC ES Test 2013*</b> (Ellis et al., 2013)	×	×	×			×	Target: EN Source: ES, EN	Proper names (persons (32.8%), organizations (36.0%), geo-political entities (31.2%))	2,117 (61.6% in KB, 38.4% NILs)	Newswire data, web data	1,832	TAC KB (based on English Wikipedia from October 2008)	Annotated by humans (LDC)	n.a.	Yes
<b>SIGIR ERD 2014</b>	×			×			en	proper nouns				Freebase (snapshot from September 2013)	Post-hoc evaluation		Yes
<b>NTCIR 9 EN-ZH Test*</b> (Tang et al., 2011)	×					×	Source: EN Target: ZH	Keywords (including common nouns and proper names)	2,116  7,052	Wikipedia articles	25	Chinese Wikipedia (27.6.2010)	Wikipedia ground truth evaluation  post-hoc evaluation	n.a.	Yes
<b>NTCIR 10 EN-ZH Test</b> (Tang et al., 2013)	×					×	Source: EN Target: ZH	Keywords (including common and proper nouns)	1,215	Wikipedia articles	25	Chinese Wikipedia (11.1.2012)	Wikipedia ground truth evaluation  post-hoc evaluation	n.a.	Yes

Data Set	Task			Language Information			Lang.	Mention Information		Corpus Information		Inventory	Annotation Information		Usage
	Name	Dis-ambig.	NIL Recogn.	Clust.	Mono-ling.	Multi-ling.		Cross-ling.	Definition	Tokens	Source		Texts	Version	
<b>NTCIR 9 EN-JA Test*</b> (Tang et al., 2011)	×					×	Source: EN Target: JA	Keywords (including common and proper nouns)	2,939  1,196	Wikipedia articles	25	Japanese Wikipedia (24.6.2010)	Wikipedia ground truth evaluation  Post-hoc evaluation	n.a.	Yes
<b>NTCIR 10 EN-JA Test</b> (Tang et al., 2013)	×					×	Source: EN Target: JA	keywords (including common nouns and proper names)	1,744	Wikipedia articles	25	Japanese Wikipedia (4.1.2012)	Wikipedia ground truth evaluation  post-hoc evaluation	n.a.	Yes
<b>NTCIR 9 EN-KO Test*</b> (Tang et al., 2011)	×					×	Source: EN Target: KO	Keywords (including common nouns and proper names)	1,681  5,107	Wikipedia articles	25	Korean Wikipedia (28.6.2010)	Wikipedia ground truth evaluation  post-hoc evaluation	n.a.	Yes
<b>NTCIR 10 EN-KO Test</b> (Tang et al., 2013)	×					×	Source: EN Target: KO	Keywords (including common nouns and proper names)	948	Wikipedia articles	25	Korean Wikipedia (22.1.2012)	Wikipedia ground truth evaluation, post-hoc evaluation	n.a.	Yes
<b>NTCIR 10 ZH-EN Test</b> (Tang et al., 2013)	×					×	Source: ZH Target: EN	Keywords (including common nouns and proper names)	1,358	Wikipedia articles	25	Chinese Wikipedia (4.1.2012)	Wikipedia ground truth evaluation  post-hoc evaluation	n.a.	Yes
<b>NTCIR 10 JA-EN Test</b> (Tang et al., 2013)	×					×	Source: JA Target: EN	Keywords (including common nouns and proper names)	1,719	Wikipedia articles	25	Japanese Wikipedia (4.1.2012)	Wikipedia ground truth evaluation  post-hoc evaluation	n.a.	Yes
<b>NTCIR 10 KO-EN Test</b> (Tang et al., 2013)	×					×	Source: KO Target: EN	Keywords (including common and proper nouns)	1,114	Wikipedia articles	25	Korean Wikipedia (4.1.2012)	Wikipedia ground truth evaluation  post-hoc evaluation	n.a.	Yes

Data Set	Task			Language Information			Lang.	Mention Information		Corpus Information		Inventory	Annotation Information		Usage
	Dis-ambig.	NIL Recogn.	Clust.	Mono-ling.	Multi-ling.	Cross-ling.		Definition	Tokens	Source	Texts		Version	Strategy	
<b>Twitter</b> (Meij et al., 2012)	×			×			EN	No mentions; concepts contained in, meant by or relevant to a tweet	n.a.	Tweets	419	Wikipedia (11.10.2010)	Annotated by humans	n.a.	No
<b>FACCI</b> (Gabrilovich et al., 2013)	×			×			EN	System mentions (unclear what exactly is annotated)	11,240M	ClueWeb09 and ClueWeb12	796M	Freebase (mapping to Wikipedia is available)	Automatically annotated with a high precision system (precision is estimated to 80-85%, recall to 70-85%), also confidence level of the system is reported	n.a.	No
<b>MASC DATA</b> (Moro et al., 2014b; 2014a)	×			×			EN	All words (including common and proper nouns)	286,416 (6.5% proper nouns)	MASC 3.0	392	BabelNet 2.0	Automatically annotated (70.1% accuracy for proper nouns, based on a manual evaluation of 600 proper nouns)	n.a.	No

Table 8.4: Comparison of different data sets that use Wikipedia as an inventory. The data sets that are marked by an asterisk after the data set name are used for evaluation in this thesis.

# Chapter 9

## Experiments

This chapter is devoted to the evaluation of the proposed concept disambiguation and clustering approach. We compare the performance of our system to the state of the art and also analyze the contribution of its different parts to the overall performance. The latter allows us to test our assumptions and their implementation and to draw conclusions for future work.

Since Gale et al. (1992a) proposed a lower and upper bound for word sense disambiguation, the evaluation setup for disambiguation systems has been widely standardized. The numerous shared tasks in this area have played a significant role in this standardization process. Concerning concept clustering, some commonly used baselines and evaluation metrics exist, but the standardization is less advanced than in disambiguation. To enable a comparison to related work, we adopt the commonly used evaluation metrics, baselines and upper bounds in this thesis.

Different data sets are annotated using different mention definitions (Chapter 8). For instance, the TAC data sets are only annotated for proper names, while the ACE 2005 data set contains annotations for both common nouns and proper names that match certain predefined entity types. As we are interested in disambiguating and clustering all mentions in a text, we disambiguate and cluster all of them. The evaluation is then performed on the respective annotated subset of mentions. We assume that this annotated subset is a representative sample for all mentions or at least for a certain mention type and is thus indicative for the overall performance of our system. By using different data sets for the evaluation that are annotated for different mention types, we can obtain a more complete picture of the performance of our system that is not restricted to one mention type. This evaluation strategy has already been proposed by Resnik & Yarowsky (1999) to reduce the annotation costs.

In the following, we first present the evaluation setup including baselines, upper bounds, related approaches and different variants of our system (Section 9.1). We then describe the used evaluation metrics (Section 9.2). The results on the different data sets described in Section 8.4 are shown in Section 9.3. In Section 9.4, we provide a detailed analysis of the results

focusing on the main contributions of this thesis. A brief summary is given in Section 9.5.

## 9.1 Settings

To analyze the performance of our system, we conduct four types of comparisons. First, we estimate how difficult the tasks are by studying the performance of some baselines. We describe these **baselines** in Section 9.1.1. Second, we assess the maximum performance we can reach to obtain a better idea of what is possible at all. These **upper bounds** are explained in Section 9.1.2. Third, we use different variants of our system to isolate the contributions of different aspects of our system. These **system variants** are sketched in Section 9.1.3. Finally, we aim to compare our system to the **state of the art** lined out in Section 9.1.4.

### 9.1.1 Baselines

For both concept disambiguation and concept clustering, some commonly used baselines exist. In both cases, they are strong and are often hard to beat (McCarthy et al., 2007; Ji et al., 2011).

**Baselines for Concept Disambiguation.** For concept disambiguation, we use the *First Concept Baseline*, which is the most commonly used baseline in this area. It is a supervised baseline and requires an annotated corpus.

*First Concept Baseline (First Concept).* The first concept baseline – also known as the *most frequent sense or concept baseline* – has been proposed by Gale et al. (1992a) and has served as a baseline in many shared tasks (Kilgariff & Rosenzweig, 2000; Palmer et al., 2001; Pradhan et al., 2007, inter alia). Given a mention to disambiguate, its distribution over candidate concepts is estimated based on an annotated corpus. The concept with the highest probability is then considered as the correct concept. If a mention has no candidate concepts, we consider it as a NIL. To estimate the concept distributions, we exploit the internal hyperlinks in Wikipedia in the respective language version (Section 3.3). We also use the distribution over concepts as a feature (prior probability, Section 5.1.1). Hence, the first concept baseline corresponds to a system which only uses the prior probability as a feature. For the first concept baseline, the NILs are clustered based on string match.

**Baselines for Concept Clustering.** For concept clustering, we use the following two baselines.

*String Match Baseline (String Match Clustering).* A strong baseline for concept clustering is to cluster all mentions that share the same string. We use the lemma to compare two strings. This baseline has also been used by e.g. Bagga & Baldwin (1998b) and has been proven as a strong baseline in the entity linking task at TAC (Ji et al., 2011). If the recall of the string match

baseline is high, this indicates that the variability among the mentions in the same clusters is low. The precision of this baseline reflects the ambiguity in the corpus: the lower the precision is, the more mentions with the same string denote different concepts.

*Singleton Baseline (Singleton Clustering).* The second baseline for clustering is the singleton baseline, where each mention builds its own cluster. Dependent on the evaluation metric, this baseline leads to a maximal precision. If the recall is high, it indicates that many mentions are singletons.

### 9.1.2 Upper Bounds

The upper bound should indicate the maximal performance that can be achieved on a data set. In the past, the human inter-annotator agreement has been used to measure the upper bound (Gale et al., 1992a; Ji et al., 2011; Palmer et al., 2006, inter alia). We reported the inter-annotator agreement in Table 8.4 if available. However, in this section, we are more interested in the maximal performance our disambiguation and clustering algorithm can achieve given our preprocessing and lexicon. We thus use the following upper bounds.

*Upper Bound Mention.* This upper bound indicates the maximal performance our system can reach given our mention and candidate concept identification without considering clustering relations during disambiguation. For each mention, we check if the correct concept is among its candidate concepts. If this is the case, it is assigned this concept given that we recognized the mention. Otherwise it is randomly assigned a wrong candidate concept given that we recognized the mention. If it is a NIL according to the gold standard, we assign it a NIL tag if we identified the mention. The NILs are then clustered according to the gold standard.

*Upper Bound Document.* As candidate concepts can be shared by mentions via clustering, we calculate a document-based upper bound. We first collect the candidate concepts of all mentions in a document. We then check for each mention if the annotated concept is a candidate concept of any mention in the document. If this is the case, we assign it this concept given that we identified the mention. Otherwise, we assign it one of its wrong candidate concepts given that we identified the mention. If it is a NIL according to the gold standard, we assign it a NIL tag given that we identified the mention. This upper bound roughly approximates the maximal performance given we exploit intra-document clustering relations. The NILs are clustered according to the gold standard.

### 9.1.3 System Variants

In this thesis, we propose a multi- and cross-lingual joint scope-aware disambiguation and clustering approach. Hence, we need to isolate the effects of the joint disambiguation and clustering (Contribution 1 & 3) and the scope awareness (Contribution 2 & 3) and evaluate its

multi- and cross-lingual portability (Contribution 4). To thoroughly analyze all these aspects, multiple variants of our system are required. In Section 9.3, we give a broad overview over our system’s performance and restrict ourselves to a few system variants. We summarize all of them in Table 9.1. More system variants that allow to further fragment the results are then discussed in the analysis section (Section 9.4).

Our first system variant (*ML Clustering*) only performs clustering without disambiguation. It is implemented in Markov logic and uses all our clustering features. It allows us to compare our clustering approach to the *string match* and *singleton baselines*. The second system variant (*ML Dis WM*) is a reimplementaion of the approach of Milne & Witten (2008b) in Markov logic and serves as our basic disambiguation approach. It only uses the features of Milne & Witten (2008b)<sup>1</sup> and allows us to quantify how the approach of Milne & Witten (2008b) performs in Markov logic. Only mentions with no candidate concepts are considered as NILs and are clustered after the disambiguation using *ML Clustering*. It is thus a cascaded approach. The third system variant (*ML Dis+NILs*) extends the basic disambiguation approach (*ML Dis WM*) with more features. The disambiguation and the recognition of NILs are performed jointly. The NILs are clustered after disambiguation using *ML Clustering*. The difference in the performance between *ML Dis WM* and *ML Dis+NILs* is thus due to more features and the joint approach for disambiguation and the recognition of the NILs. We did not separately evaluate the contributions of these two factors in this thesis, as joint disambiguation and recognition of the NILs is not in our main focus here. The readers interested in this aspect are referred to Fahrni & Strube (2012) where we provide more information about this aspect. The next system variant (*ML Dis+NILs+Clust*) performs joint disambiguation and clustering using exactly the same features as *ML Dis+NILs*. It thus allows us to quantify the effect of performing disambiguation and clustering jointly (Contribution 1 & 3). Our last variant (*ML Dis+NILs+Clust Scope*) is the scope-aware extension of *ML Dis+NILs+Clust* and shows the contribution of using more than one context definition (Contribution 2 & 3).

To learn the weights for our *ML Dis+NILs+Clust Scope* approach, the initialization needs to be considered. As the conditional log-likelihood augmented with latent variables is not a concave function anymore, it is not guaranteed that the global optimum is found. In this case, initialization is crucial (Smith, 2011). We initialize the weights with  $\pm 1$  based on linguistic knowledge.

---

<sup>1</sup>We do not use the context quality feature, as it did not influence the results in the Markov logic setup.

System	Disambiguation	Clustering	Context Modeling	Features
<b>ML Clustering</b>	n.a.	Pair-wise clustering approach in Markov logic (Section 4.3.2)	Uniform context modeling	<i>Clustering</i> : 10-14
<b>ML Dis WM</b>	Basic disambiguation system in Markov logic that only uses the features of Milne & Witten (2008b) (Section 4.3.1); a mention is a NIL if it has no corresponding candidate concept	Clustering of the NILs after disambiguation using <i>ML Clustering</i>	Uniform context modeling	<i>Disambiguation</i> : 1,3 <i>Clustering of NILs</i> : 10-14
<b>ML Dis+NILs</b>	Extension of <i>ML Dis WM</i> with more features; joint disambiguation and recognition of NILs (Section 4.3.1); a mention is a NIL if it has no corresponding candidate concept or if none of the candidate concepts has been selected by the system	Clustering of the NILs after disambiguation using <i>ML Clustering</i>	Uniform context modeling	<i>Disambiguation</i> : 1-6 (ACE 2004, ACE 2005); 1-9 (all other data sets) <i>Clustering of NILs</i> : 10-14
<b>ML Dis+NILs+Clust</b>	Joint disambiguation and clustering approach (Section 4.3.3): combination of the system <i>ML Dis+NILs</i> and <i>ML Clustering</i>		Uniform context modeling	<i>Disambiguation</i> : 1-6 (ACE 2004, ACE 2005); 1-9 (all other data sets) <i>Clustering</i> : 10-14
<b>ML Dis+NILs+Clust Scope</b>	Joint disambiguation and clustering approach: combination of the system <i>ML Dis+NILs+Clust</i> and <i>ML Clustering</i>		Scope-aware approach (Section 4.4)	<i>Disambiguation</i> : 1-6 (ACE 2004, ACE 2005); 1-9 (all other data sets) <i>Clustering</i> : 10-14 <i>Scope assignment</i> : 15-28

Table 9.1: Variants of our system.

### 9.1.4 State-of-the-Art Approaches

Many factors can affect the final performance of a system, including the candidate concept identification strategy, the used features or the effective disambiguation method (Chapter 6). It is thus often difficult to say why a particular system performs better than another system. To compare for instance different disambiguation methods, ideally the same candidate concept identification strategies and features should be used. However, the methods are usually developed based on certain candidate concept identification strategies or based on certain features. It is therefore difficult to perform such comparisons. We adapted WikipediaMiner (Milne & Witten, 2008b), which models concept-level interrelations via aggregation, accordingly and plugged in our mentions and our candidate concepts. This allows for a fair comparison with it. In addition, we implemented a basic disambiguation system in Markov logic that only uses the WikipediaMiner features (Section 9.1.3). In all other cases, we compare the systems as a whole.

We selected the state-of-the-art systems to compare to based on two criteria, i.e. their performance and the closeness of their method to ours. They are selected based on the comparison summarized in Table 10.1 and Table 10.2 (Chapter 10). We briefly describe all systems we compare to in Table 9.2. They are discussed in more detail in Chapter 10. A few other related systems (e.g. Turdakov & Lizorkin (2009) and Kulkarni et al. (2009)) are not evaluated on any of the data sets we use and are also not available. We thus do not compare to these systems in this chapter. For the TAC and NTCIR data sets, we additionally report the median performance that is calculated based on the results of all participating systems (*Median*).

Approach	Focus	Lang.	Mentions	Method	Context Definition	Features
<b>WM (Milne &amp; Witten, 2008b)</b>	Keyword linking	EN	Keywords	<p><b>Candidate concept identification:</b> lexicon-based, threshold for link probability (discards all concepts with low prior probability)</p> <p><b>Disambiguation:</b> C4.5 with bagging to balance between three features</p> <p><b>Interrelations between mentions:</b> aggregative approach: weighted by relatedness and link probability</p> <p><b>NILs:</b> n.a. (could be tuned by threshold)</p> <p><b>Clustering:</b> n.a.</p>	All unambiguous mentions weighted by their average relatedness and link probability	Average relatedness, context quality (relatedness of the context by itself), prior probability
<b>Ratinov et al. (2011), Cogcomp (Ratinov &amp; Roth, 2011)</b>	Disambiguation as an optimization problem with a local and global version	EN	Common nouns and proper names	<p><b>Candidate concept identification:</b> lexicon-based</p> <p><b>Disambiguation:</b> SVM-based approach: ranker (obtain best concept for all mentions), linker (is a mention a NIL)</p> <p><b>Interrelations between mentions:</b> aggregation strategy: the disambiguation context is formed by the concepts obtained by applying a local disambiguation approach</p> <p><b>NILs:</b> SVM-based classifier</p> <p><b>Clustering:</b> TAC 2011: string match clustering of NILs after disambiguation</p>	Best concept according to local ranker of all mentions; 100-token window	Prior probability, probability of concept, text-based similarity measures, pairwise relatedness measure based on in- and outlinks
<b>Cheng &amp; Roth (2013), UI CCG (Cheng et al., 2014)</b>	Exploiting relation extraction for disambiguation	EN	Common nouns and proper names	<p><b>Candidate concept identification:</b> lexicon-based, relational queries</p> <p><b>Disambiguation:</b> ILP-based approach</p> <p><b>Interrelations between mentions:</b> combination of aggregative and joint techniques; exploiting of relation extraction techniques and within-document coreference information (hard constraints)</p> <p><b>NILs:</b> SVM-based classifier</p> <p><b>Clustering:</b> TAC 2013: pairwise classification-based approach</p>	Document context, 100-token window, coreferent mentions, local sentence-level mentions obtained via relation extraction	Prior probability, probability of concept, text-based similarity measures, pairwise relatedness measure based on in- and outlinks, coreference information, relatedness score based on DBPedia and Wikipedia
<b>Hoffart et al. (2011b)</b>	Graph-based approach with filters	EN	Proper names	<p><b>Candidate concept identification:</b> lexicon-based</p> <p><b>Disambiguation:</b> graph-based approach: (1) robustness heuristics to filter graph; (2) filtering based on degree; (3) approximation of dense subgraph</p> <p><b>Interrelations between mentions:</b> approximation of dense subgraph; robustness heuristic: only for parts of the mentions</p> <p><b>NILs:</b> n.a.</p> <p><b>Clustering:</b> n.a.</p>	Binwise: concept-level relatedness is only relevant for mentions that cannot be disambiguated based on prior probability and similarity; syntactic context was not helpful	Prior probability, similarity based on keywords, pairwise relatedness measure based on inlinks

Approach	Focus	Lang.	Mentions	Method	Context Definition	Features
<b>Chen &amp; Ji (2011)</b> <b>Blender CUNY (Tamang et al., 2013)</b> <b>RPI Blender (Yu et al., 2014)</b>	Collaborative ranking: clustering first; disambiguation of mention in one cluster at the same time; use of multiple rankers	EN	Proper names	<b>Candidate concept identification:</b> retrieval-based approach with Lucene <b>Disambiguation:</b> disambiguation as a ranking problem <b>Interrelations between mentions:</b> mentions are first clustered; disambiguating mentions in the same cluster together using a ranking-based approach <b>NILs:</b> probably part of the ranking, but not discussed <b>Clustering:</b> before disambiguation using spectral clustering; TAC 2012: collaborative clustering; TAC 2013: string match and singleton clustering	Whole documents	Many features including surface features (e.g. string similarity), document-based features, profiling features that are extracted using information extraction techniques
<b>Dalton &amp; Dietz (2013)</b> <b>UMass (Dietz &amp; Dalton, 2013;</b> <b>Dalton &amp; Dietz, 2014)</b>	Information retrieval based approach with neighbourhood ranking leveraging similar documents; joint candidate identification and ranking	EN	Proper names	<b>Candidate concept identification:</b> jointly done with ranking; defined as IR task <b>Disambiguation:</b> (1) IR-based approach, (2) supervised reranking, (3) threshold to recognize NILs <b>Interrelations between mentions:</b> neighborhood relevance model: interrelations to unambiguous concepts of other proper names <b>NILs:</b> threshold <b>Clustering:</b> n.a.	Weighted context dependent on across-document co-occurrences in similar documents; local sentence (words)	Name similarity, co-reference resolution for query expansion, name overlap, sentences of co-referent mentions; weighted occurrences with concepts of mentions in document (no relations between the mentions)
<b>Cucerzan (2007)</b> <b>MS MLI (Cucerzan, 2012;</b> <b>2013)</b>	Large-scale proper name disambiguation	EN	Proper names (recognized by the system)	<b>Candidate concept identification:</b> lexicon-based; within document coreference information <b>Disambiguation:</b> unsupervised vector space model; for each mention the candidate concept with the highest similarity to the document context is chosen <b>Interrelations between mentions:</b> aggregative approach over all candidate concepts of all mentions in the context <b>NILs:</b> n.a. <b>Clustering:</b> for TAC: heuristics such as acronym expansion	Binwise: the context is iteratively shrunk: from the document, to the paragraph to the sentence of a mention until only one likely candidate concept given the model is left	Lexical features, category overlap between candidate concepts and all candidates in the context (categories are obtained from list pages, categories, lexicosyntactic patterns) TAC 2012 and 2013: many more features including string-level similarities and prior probability

Approach	Focus	Lang.	Mentions	Method	Context Definition	Features
<b>LCC (Monahan et al., 2011; Monahan &amp; Carpenter, 2012)</b>	Multi-stage approach: (1) disambiguation of mentions, (2) clustering of the mentions using the concept information, (3) additional disambiguation step in which each cluster is linked to a concept if possible	EN, ZH, ES	Proper names	<p><b>Candidate concept identification:</b> lexicon-based</p> <p><b>Disambiguation:</b> ranking-based approach</p> <p><b>Interrelations between mentions:</b> multi-stage approach, clustering information is considered in the final disambiguation step</p> <p><b>NILs:</b> after disambiguation, classification-based approach using logistic regression</p> <p><b>Clustering:</b> agglomerative clustering</p> <p><b>Cross-lingual system:</b> best run disambiguates in the source language and maps the results via cross-language links in Wikipedia to the target language</p>	Surrounding and document-level context	Many features including concept type, mention similarity features, local context features, intra-document coreference information, features that leverage relation extraction techniques
<b>Basistech (Clarke et al., 2012; Merhav et al., 2013)</b>	Concept disambiguation is casted as a cross-document clustering problem; each concept is a cluster	EN, ZH, ES	Proper names	<p><b>Candidate concept identification:</b> lexicon-based with filters, string similarities</p> <p><b>Disambiguation:</b> supervised using a structural SVM to disambiguate the whole coreference chain a mention is part of</p> <p><b>Interrelations between mentions:</b> all mentions within an intra-document coreference chain are disambiguated and clustered together</p> <p><b>Clustering:</b> incremental clustering</p> <p><b>Cross-lingual system:</b> exploiting of the Google Cross-wiki dataset (Spitkovsky &amp; Chang, 2012); translation of mentions</p>	Each mention to be disambiguated and clustered is associated with its intra-document coreference chain; local and document-level context	Context features, prior probability, coreference information, concept type information

Table 9.2: Summary of the approaches we compare to. They have been selected based on the closeness of their method to ours (e.g. modeling of interrelations, context modeling), availability and performance.

## 9.2 Evaluation Metrics

Over the years, different evaluation metrics have been proposed to evaluate concept disambiguation and clustering systems. Depending on the data sets, different evaluation metrics have been established as a standard. In order to be able to compare to other systems, we thus need to report different evaluation metrics. We first describe the metrics for concept disambiguation (Section 9.2.1) and then for concept clustering (Section 9.2.2).

### 9.2.1 Metrics for Concept Disambiguation

In word sense disambiguation, the most common evaluation metrics are precision, recall and F-Measure (Palmer et al., 2006, p. 37). However, the results are usually only reported for the occurrences or mentions with a corresponding concept in the inventory. As we also want to measure the performance of the NILs, we report precision, recall and F-measure separately for mentions with a corresponding concept in the inventory (*inv*) and for mentions with no corresponding concept in the inventory (*NILs*). In the following,  $G_{inv}$  comprises all mentions in the gold standard annotated with a corresponding concept in the inventory.  $T_{inv}$  consists of all mentions in the system output that exactly match a gold mention and that the system assigned a concept in the inventory. Following Hoffart et al. (2014) and our own work (Fahrni et al., 2012), we then calculate precision ( $P_{inv}$ ), recall ( $R_{inv}$ ) and F-measure ( $F_{inv}$ ) for mentions with a corresponding concept in the inventory as follows

$$P_{inv} = \frac{|G_{inv} \cap T_{inv}|}{|T_{inv}|}$$

$$R_{inv} = \frac{|G_{inv} \cap T_{inv}|}{|G_{inv}|}$$

$$F_{inv} = \frac{2 \cdot P_{inv} \cdot R_{inv}}{P_{inv} + R_{inv}}.$$

Analogously, we define the precision ( $P_{NILs}$ ), recall ( $R_{NILs}$ ) and F-measure ( $F_{NILs}$ ) via

$$P_{NILs} = \frac{|G_{NILs} \cap T_{NILs}|}{|T_{NILs}|}$$

$$R_{NILs} = \frac{|G_{NILs} \cap T_{NILs}|}{|G_{NILs}|}$$

$$F_{NILs} = \frac{2 \cdot P_{NILs} \cdot R_{NILs}}{P_{NILs} + R_{NILs}}$$

with  $G_{NILs}$  being the mentions in the gold standard that are annotated as NILs and  $T_{NILs}$

being the mentions in the system output that exactly match a mention in the gold standard and that are annotated as NILs by the system. As indicated at the beginning of this chapter, in all gold standards only some mentions are annotated. We follow the strategy proposed by Resnik & Yarowsky (1999) or Ratinov et al. (2011) and only evaluate on this annotated sample.

In the entity linking task at TAC, the official evaluation metric to measure the disambiguation performance of systems is the accuracy (also known as micro-average score). We thus also report the accuracy calculated via

$$acc = \frac{|G \cap T|}{|G|}$$

where  $G$  comprises all annotated mentions in the gold standard and  $T$  all mentions in the system output that exactly match a gold mention. This measure accounts for both the mentions with a concept in the inventory and the NILs.

For the ACE 2004 data set, the BOT measure has been established as an evaluation metric. BOT stands for bag of titles and is document- instead of mention-based (Milne & Witten, 2008b; Ratinov et al., 2011). For each document, the concepts of all mentions in the gold standard are collected (gold concepts). At the same time, the mentions in the system output that exactly match a mention in the gold standard are identified and their concepts are collected (system concepts). In both the system and the gold concepts duplicates are removed. Then, precision ( $BOTP$ ), recall ( $BOTR$ ) and F-measure ( $BOTF$ ) are calculated based on these two sets as defined above (Ratinov et al., 2011). We use the implementation of Ratinov et al. (2011) to calculate this metric.<sup>2</sup>

For the NTCIR data set, we report the *mean average precision* ( $MAP$ ) and the *precision-at-5* ( $P@5$ ). The *mean average precision* is defined as

$$MAP = \frac{1}{N} \left( \sum_{n=1}^N \frac{\sum_{k=1}^K p_{kn}}{K} \right)$$

with  $N$  being the number of documents and  $K$  being the number of identified concepts in a document.  $p_{kn}$  is the precision for the top  $k$  concepts for a document  $n$ . In order to be able to identify the top  $k$  concepts for a document, the system must rank the concepts for each document (e.g. by their importance for the document).

The *precision-at-5* is the precision for the top 5 concepts for a document and is averaged over all documents. These two evaluation metrics from information retrieval were also used in the cross-lingual link discovery shared task at NTCIR 9 and allow us to compare our results to the shared task results. We use the official implementation from the shared task and restrict ourselves to the file-to-file evaluation based on the Wikipedia ground truth.<sup>3</sup>

<sup>2</sup>[http://cogcomp.cs.illinois.edu/page/software\\_view/Wikifier](http://cogcomp.cs.illinois.edu/page/software_view/Wikifier), 1.8.2014.

<sup>3</sup><http://ntcir.nii.ac.jp/CrossLink/>, 1.8.2014.

## 9.2.2 Metrics for Concept Clustering

For concept clustering, we use the  $B^3$  metric proposed by Bagga & Baldwin (1998a). This metric has also been used in the entity linking task at TAC to evaluate the clustering results and thus allows us to compare our results to others. Given a mention  $m_i$ ,  $C(G_{m_i})$  is its cluster according to the gold standard and  $C(T_{m_i})$  is its cluster according to the system output. For each mention, its precision is given by

$$B^3P(m_i) = \frac{|C(G_{m_i}) \cap C(T_{m_i})|}{|C(T_{m_i})|}$$

and its recall is given by

$$B^3R(m_i) = \frac{|C(G_{m_i}) \cap C(T_{m_i})|}{|C(G_{m_i})|}.$$

The overall precision and recall are given by averaging over all mentions.

In the entity linking task at TAC a further variant is used called  $B^{3+}$  (Ji et al., 2011). In this variant, it is not only checked if the mentions are in the same cluster to calculate  $|C(G_{m_i}) \cap C(T_{m_i})|$ , but also if the mentions with a corresponding concept in the inventory are assigned the correct concept. We report both versions using the the implementation provided by the organizers of the entity linking task at TAC (version 0.7).<sup>4</sup>

## 9.3 Results

In the following, we present the results for our system on the data sets described in Section 8.4. Besides the results for our full joint scope-aware system, we show the results for its different variants (Table 9.1) and compare it to our baselines, upper bounds and the results of related work. To calculate significance between two system outputs, we used two-sided approximate randomization testing (Noreen, 1989) with 10,000 iterations.<sup>5</sup>

### 9.3.1 Monolingual English Results

We evaluate our monolingual English system on the ACE 2005, the ACE 2004 and the English TAC data sets.

**ACE 2005.** The ACE 2005 data set allows us to evaluate the concept disambiguation performance for both common nouns and proper names. As the NILs are not clustered in the gold

<sup>4</sup>[http://www.nist.gov/tac/2013/KBP/EntityLinking/tools/el\\_scorer.py](http://www.nist.gov/tac/2013/KBP/EntityLinking/tools/el_scorer.py), 1.8.2014.

<sup>5</sup>We used the implementation of Sebastian Martschat, which is available from here: <https://github.com/smartschat/art>, 11.10.2014.

	In Inventory			NILs			Acc
	$P_{inv}$	$R_{inv}$	$F_{inv}$	$P_{NILs}$	$R_{NILs}$	$F_{NILs}$	
Upper Bound Mention	90.1	87.8	89.0	84.3	92.5	88.2	88.2
Upper Bound Document	95.8	93.9	94.9	90.7	92.5	91.6	93.8
(1) First Concept	68.6	70.0	69.3	66.0	32.8	43.8	67.3
(2) WM (Milne & Witten, 2008b)	86.4	60.0	70.8 <sup>1</sup>	16.6	78.1	27.4	61.3
(3) ML Dis WM	70.8	72.3	71.6 <sup>1,2</sup>	66.0	32.8	43.8 <sup>2</sup>	69.5 <sup>1,2</sup>
(4) ML Dis+NILs	77.3	76.0	76.6 <sup>1,2,3</sup>	47.6	46.6	47.1 <sup>1,2,3</sup>	73.9 <sup>1,2,3</sup>
(5) ML Dis+NILs+Clust	76.8	76.9	76.9 <sup>1,2,3,4</sup>	55.7	42.5	<b>48.2</b> <sup>1,2,3,4</sup>	74.4 <sup>1,2,3,4</sup>
(6) ML Dis+NILs+Clust Scope	80.3	77.1	<b>78.6</b> <sup>1,2,3,4,5</sup>	42.4	54.6	47.7 <sup>1,2,3</sup>	<b>75.4</b> <sup>1,2,3,4,5</sup>

Table 9.3: ACE 2005: precision ( $P_{inv}$ ,  $P_{NILs}$ ), recall ( $R_{inv}$ ,  $R_{NILs}$ ) and F-measure ( $F_{inv}$ ,  $F_{NILs}$ ) for the non-NILs (*In Inventory*) and the NILs (*NILs*) respectively and the overall accuracy ( $Acc$ ). Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

annotations, it is not suitable to evaluate the clustering performance. However, we can still evaluate the influence of the clustering on the disambiguation.

Table 9.3 shows the results on the ACE 2005 data set. We report precision ( $P_{inv}$ ,  $P_{NILs}$ ), recall ( $R_{inv}$ ,  $R_{NILs}$ ) and F-measure ( $F_{inv}$ ,  $F_{NILs}$ ) for the Non-NILs (*In Inventory*) and the NILs (*NILs*) respectively and the overall accuracy ( $Acc$ ). The highest accuracy and F-measures are bolded, while significance is indicated by superscripts. We applied the randomized significance test to the accuracy and the F-measures. If a system has a significantly higher F-measure or accuracy than another system ( $p < 0.05$ ), the system identifier of this other system is added as a superscript to the respective score. The system identifiers correspond to the numbers in the first column in Table 9.3.

*Upper Bounds and Baseline.* The mention-based upper bound (*Upper Bound Mention*) indicates that a disambiguation system that uses no clustering information can reach at most an overall accuracy of 88.2 given our preprocessing and lexicon. As the document-based upper bound (*Upper Bound Document*) shows this upper bound can be raised to an accuracy of 93.8 if for each mention the candidate concepts of all mentions in the document are considered as its candidate concepts. This document-based upper bound roughly approximates the upper bound for a concept disambiguation system that makes use of clustering information. Besides these two upper bounds, we also report the *first concept baseline* which already achieves an accuracy of 67.3.

*Our Results.* Our basic disambiguation system that reimplements the *WikipediaMiner* features in Markov logic (*ML Dis WM*) significantly outperforms the first concept baseline both with respect to  $F_{inv}$  and  $Acc$  with  $p < 0.01$ . The results for the Non-NILs are the same as for the first concept baseline as NILs are not recognized. Only if a mention has

no candidate concept, it is considered as a NIL. This strategy leads to a recall of 32.8 for NILs ( $R_{NILs}$ ) and a precision of 66.0. Hence, 32.8 of the NILs are trivial to resolve given our candidate concept identification. Our extended version of this system (*ML Dis+NILs*) that jointly disambiguates mentions and recognizes NILs significantly outperforms this basic disambiguation system with respect to all F-measures and the accuracy. While the precision and recall for the Non-NILs improves by more than 6 and 3 points, the recall for NILs increases by almost 14 points, however at the cost of precision.

These results can be further improved by including clustering information. Our joint concept disambiguation and clustering approach (*ML Dis+NILs+Clust*) significantly outperforms the disambiguation-only system (*ML Dis+NILs*) ceteris paribus. In particular, the precision for the NILs increases by more than 8 points at a drop in recall of 4 points. This indicates that mentions that may have no candidate concepts and are thus considered as NILs can be correctly disambiguated by using clustering information.

This joint concept disambiguation and clustering approach uses the same model for all mentions. Our full system, i.e. the joint scope-aware concept disambiguation and clustering approach that distinguishes between three scopes (*ML Dis+NILs+Clust Scope*) significantly outperforms all other system with  $p < 0.01$  with respect to  $F_{inv}$  and  $Acc$ . Hence, by using different models for different mentions, the performance of the Non-NILs and the accuracy significantly improve ceteris paribus. The precision for the non-NILs ( $P_{inv}$ ) is over 80 at a recall of more than 77. The F-measure for the NILs slightly drops compared to the joint concept disambiguation and clustering approach, however this difference is not statistically significant.

*Related Work.* Bryl et al. (2010) report an F-Measure of 71.5 for mentions with a corresponding concept in the inventory on the ACE 2005 data. Their classification-based approach that uses a separate classifier for each lemma thus performs in the range of our basic disambiguation approach. However, these results are not comparable to ours as they use gold mentions and consider a mention as correctly disambiguated only if it links to the first Wikipedia article in the gold standard (Section 8.4.1). In Table 9.3, we report the results of *WikipediaMiner* (Milne & Witten, 2008b) trained with the same training data as our systems and given as input our mentions and lexicon. *WikipediaMiner* (*WM*) is designed as a high precision system and achieves with 86.4 a higher precision than all our systems. However, its recall is low leading to an overall accuracy of 61.3 which is significantly worse than the first concept baseline. Our basic disambiguation system that only uses the features of *WikipediaMiner* has more balanced precision and recall scores and therefore significantly outperforms the *WikipediaMiner* system in terms of accuracy. For a further comparison, we also ran Ratinov et al.’s (2011) and Cheng’s system (2013) on ACE 2005, but it seems that their mention recognition is not designed for ACE 2005. Therefore, we do not report their results here.

*Discussion.* The results on the ACE 2005 data set indicate that concept clustering (Con-

tributions 1 & 3) and scope awareness (Contributions 2 & 3) significantly improve the disambiguation results for common nouns and proper names *ceteris paribus*. While our joint concept disambiguation and clustering system (*ML Dis+NILs+Clust*) outperforms the corresponding disambiguation system with no clustering information (*ML Dis+NILs*), our joint scope-aware system (*ML Dis+NILs+Clust Scope*) obtains higher results than the corresponding scope-ignorant approach (*ML Dis+NILs+Clust*). Compared to the baselines and related work our approach performs favorably. However, the upper bounds have an accuracy of more than 12 points higher than our full system leaving room for further improvements. In particular, the performance for the NILs is rather low and needs to be improved. We provide a more detailed analysis in Section 9.4.4.

**ACE 2004.** Similar to the ACE 2005 data set, the ACE 2004 data set is annotated for both common nouns and proper names. However, it is much smaller. Nevertheless, it allows us to compare the performance of our system to the one of Cheng & Roth (2013). Their system uses a similar problem formulation as our system and also makes use of clustering information.

Table 9.4 shows our results and results of related approaches on ACE 2004. As previous work uses the BOC evaluation metric for this data set, we report these scores using the same implementation (Section 9.2.1).

*Baseline.* Our first concept baseline achieves an *BOCF* score of 78.1. Hence, a simple system that only uses frequency information can already achieve high results on this data set. As we lack mention-based gold annotation information, we cannot calculate our upper bounds.

*Our Results.* Our basic disambiguation system (*ML Dis WM*) does not improve over the first concept baseline. Performing joint concept disambiguation and recognition of NILs slightly improves the results, but the improvements are not statically significant (*ML Dis+NILs*). Including clustering information and performing joint concept disambiguation and clustering leads to a statistically significant gain of 3.7 points in F-measure, while adding scope information leads to an overall document-based *BOCF* score of 86.3. The precision of our approach is higher than 90, while the recall is slightly above 82.

*Related Work.* Our fair comparison with *WikipediaMiner* (Milne & Witten, 2008b) shows similar trends as on the ACE 2005 data set. The precision of *WikipediaMiner* (*WM*) is fairly high (84), while its recall is in the range of the first concept baseline. In contrast to the ACE 2005 data set, our reimplementation in Markov logic leads to lower results. However, our joint scope-aware approach outperforms the *WikipediaMiner* system by 6 points in *BOCF*. Furthermore, we compare our results to the results of Ratinov et al. (2011) and Cheng & Roth (2013). The system of Ratinov et al. (2011) achieves slightly worse results than our basic disambiguation approach (*ML Dis WM*) and our first concept baseline. As their first concept baseline is also lower than ours – the *BOCF* score is 69.52 –, the main difference between their approach and our basic system can be traced back to the concept identification

	<b>BOC P</b>	<b>BOC R</b>	<b>BOC F</b>
(1) First Concept	80.2	76.1	78.1
Ratinov et al. (2011)	n.a.	n.a.	77.3
Cheng & Roth (2013)	n.a.	n.a.	85.3
(2) WM (Milne & Witten, 2008b)	84.7	76.1	80.2
(3) ML Dis WM	80.2	76.1	78.1
(4) ML Dis+NILs	82.7	76.9	79.7
(5) ML Dis+NILs+Clust	86.2	80.8	83.4 <sup>1-4</sup>
(6) ML Dis+NILs+Clust Scope	<b>90.2</b>	<b>82.7</b>	<b>86.3<sup>1-4</sup></b>

Table 9.4: Results on ACE 2004. Significant improvements of a system over another system (*BOCF*) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

component. Their disambiguation approach uses similar features as our extended disambiguation approach (*ML DIS+NILs*), including the same measure to calculate relatedness between concepts. In contrast to us, they use a cascaded approach for disambiguation and recognition of the NILs and do not rely on unambiguous mentions to calculate relatedness. These two differences might explain the difference in their performance. Overall, the results of our disambiguation-only approach (*ML DIS+NILs*) are in the same range as the results of Ratinov et al. (2011). By adding clustering and scope information, we obtain results of almost 10 points higher than the ones of Ratinov et al. (2011). Comparing our results to the results of the state-of-the-art approach of Cheng & Roth (2013) indicates that our joint scope-aware approach leads to state-of-the-art results. In both their and our approach the problem is finally transformed into an ILP and solved using the same solver. Similar as in our approach they also use clustering information. However, while we integrate the clustering information as soft constraints, they use hard constraints. In their ablation study they show that their basic system with clustering information obtains a *BOCF* of 83.4, which corresponds to our results with clustering information (*ML Dis+NILs+Clust*). Their full system uses relation extraction techniques to restrict the context, i.e. they exploit correlations between the entity- and the concept-level cohesive ties. Hence, similar to us they use two techniques to improve the results of their basic system: clustering information and a method to select the disambiguation context. Hence, although they use a different context selection strategy, their results strengthen our claim that clustering information and an improved context selection strategy leads to higher results. Combining their and our context selection strategy may lead to even higher results.

*Discussion.* As the evaluation on the ACE 2004 data set shows, our approach leads to state-of-the-art results. As on the ACE 2005 data set, we observe a similar trend of our different system variants: performing joint concept disambiguation and clustering and using scope information improves the results *ceteris paribus*. Hence, both the results on the ACE 2005 and

ACE 2004 data sets support our claims.

**English TAC Data Sets.** The data sets from the monolingual English entity linking shared task at TAC allow us to further compare our results to the state-of-the art. We restrict ourselves to the TAC data sets 2011, 2012 and 2013, as their gold annotations contain clustering information for the NILs. These annotations allow us to evaluate both the disambiguation and the clustering performance of our system. In contrast to the other data sets, only a few selected challenging mentions are annotated. In the TAC entity linking task, the participants are provided with the mentions to be disambiguated and clustered. To be able to compare our results to the results of the shared task participants, we add these annotated mentions to our mention pool in case our mention recognition did not recognize them anyway.

Tables 9.5, 9.6 and 9.7 show the results on the testing data from the English monolingual entity linking task at TAC 2011, 2012 and 2013 respectively. While significant differences between our system variants are indicated by superscripts, we cannot calculate significance to the results of other systems, as we do not have their system output. The asterisk (\*) behind some of the system names indicate that these are official shared task results. The official TAC evaluation metrics are accuracy ( $Acc$ ),  $B^3P$ ,  $B^3R$  and  $B^3F$  as well as  $B^{3+}P$ ,  $B^{3+}R$  and  $B^{3+}F$ . However, for many systems only the  $B^{3+}F$  score is available. We additionally also report precision, recall and F-measure for Non-NILs and NILs to obtain a more complete picture of our performance.

*Upper Bounds and Baselines.* On all TAC data sets, the upper bounds are determined by the candidate identification component. As we add all mentions to be disambiguated and clustered to our mention pool, the recall for the NILs is 100 for both upper bounds. While the clustering performance for the NILs is thus maximal, the overall clustering performance over Non-NILs and NILs that is reported in the tables is lower as it is constrained by the candidate concept identification step. If for instance two Non-NILs are in the same cluster according to the gold standard, but they are assigned different concepts, they are not clustered together. This can happen if the correct concept is not among the candidate concepts of the mentions. In this case, we randomly assign them one of their candidate concepts, which is not necessarily the same for the two mentions.

On all data sets, the difference between the two upper bounds is substantial, ranging between 9.2 and 16.7 points in accuracy. This indicates that the correct candidate concept is often not among the candidate concepts of the respective mentions, but among the candidates of other mentions in the document, and suggests that clustering is important. While the accuracy of the document-level upper bound is above 90 on all data sets, the results for the first concept baselines strongly vary across the different TAC data sets. On the TAC 2011 and TAC 2013 data sets, the accuracy of the first concept baseline is above 70, whereas it is lower than 40 on the TAC 2012 data set. This indicates that the TAC 2012 data set is more challenging

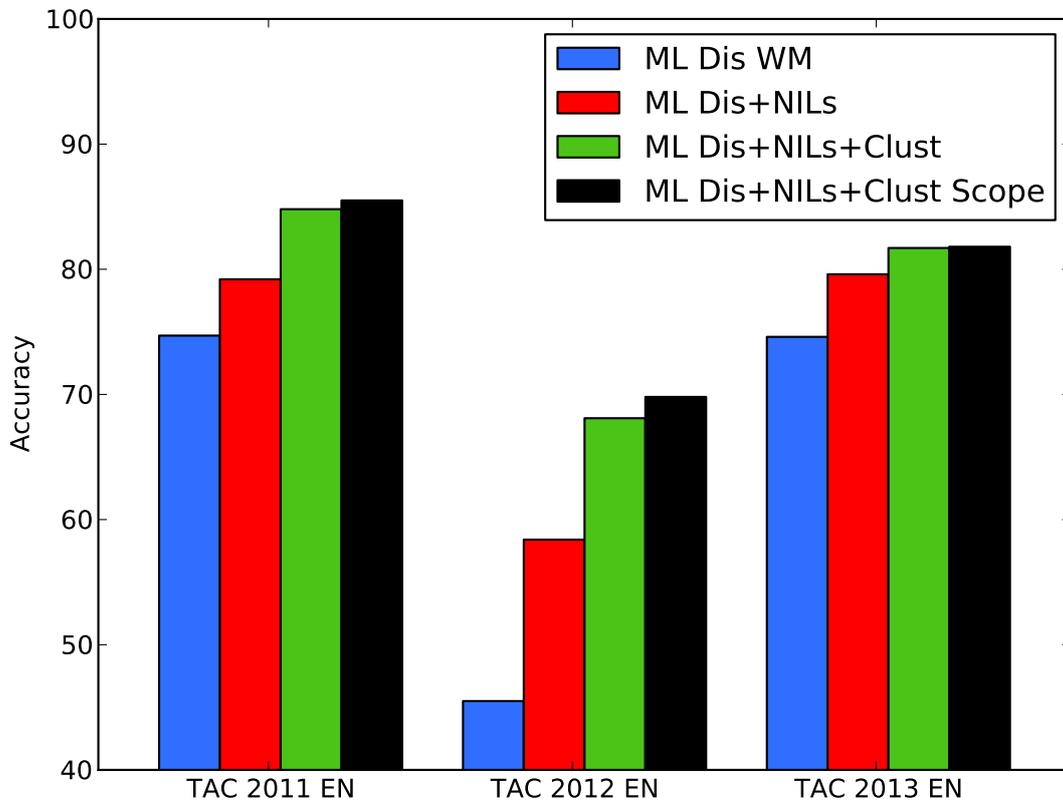


Figure 9.1: Accuracy of our different system variants on the test data sets from the English monolingual entity linking task at TAC 2011-2013. All differences are significant with  $p < 0.05$  except the differences between *ML Dis+NILs+Clust* and *ML Dis+NILs+Clust Scope* on the TAC 2011 and 2013 data sets.

than the other two English TAC data sets considered in this thesis and is in line with the lower results that have been generally achieved by the shared task participants on the 2012 data set. The data sets also vary with respect to the two baselines for the clustering. While on the TAC 2011 data sets string match clustering leads to high results ( $B^3F = 93.8$ ), much lower results can be achieved on the TAC 2012 and 2013 data sets with this heuristic ( $B^3F < 60.0$ ). Mentions with the same string thus tend to denote different concepts in these two data sets and the ambiguity per mention string in the data set is higher. The singleton baseline (*Clustering Singleton*) further shows that on the TAC 2012 data set most mentions are singletons ( $B^3R = 87.2$ ), while this is not the case on the other two data sets (TAC 2011:  $B^3R = 67.3$  and TAC 2013:  $B^3R = 33.1$ ). Hence, while the gold clusters on the TAC 2011 hardly show variability (high scores for string match), the TAC 2012 gold clusters mainly consist of singletons. The TAC 2013 gold clusters however show variability and consist of multiple mentions.

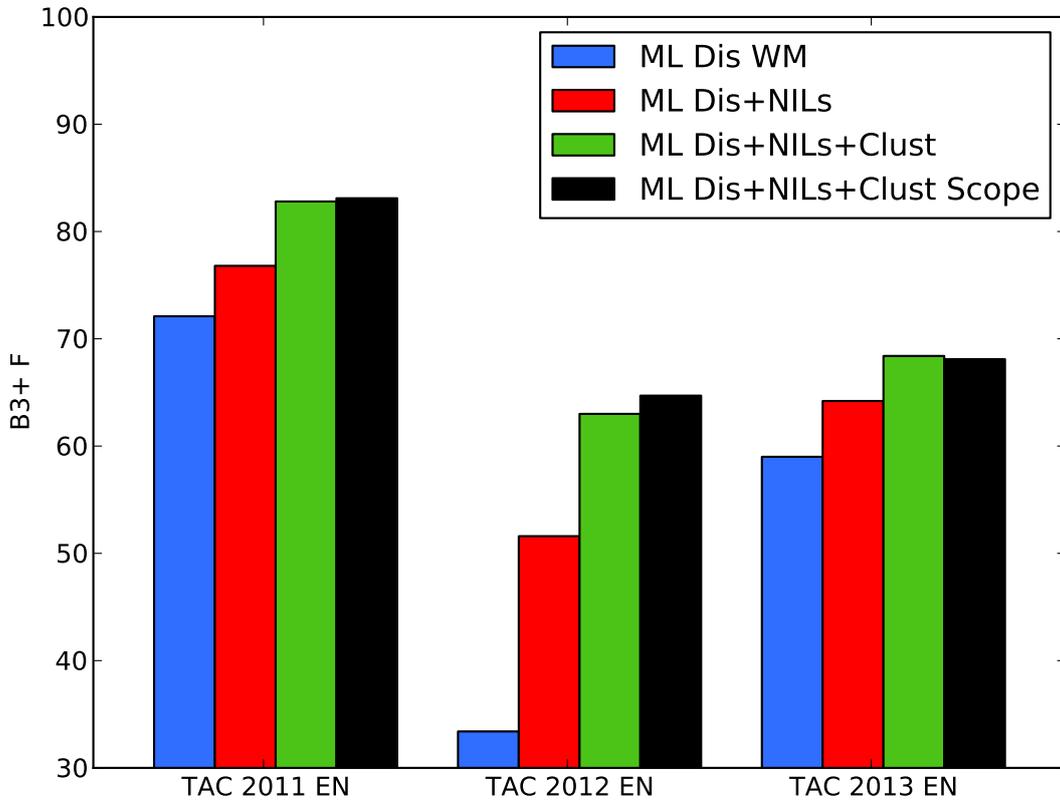


Figure 9.2:  $B^{3+}$  of our different system variants on the test data sets from the English monolingual entity linking task at TAC 2011-2013. All differences are significant with  $p < 0.05$  except the differences between *ML Dis+NILs+Clust* and *ML Dis+NILs+Clust Scope* on the TAC 2011 and 2013 data sets.

*Our Results.* On all three data sets under considerations, the general trends of our systems are the same. While Tables 9.5, 9.6 and 9.7 show all details, Figure 9.1 summarizes the accuracy scores for our different system variants. *ML Dis WM* significantly outperforms the first concept baseline with respect to the accuracy on all data sets. However, its performance is rather low. The extended version (*ML Dis+NILs*) already leads to results of almost 80 on the TAC 2011 and TAC 2013 data sets. These accuracy scores can be significantly improved on all three data set by performing joint concept disambiguation and clustering (*ML Dis+NILs+Clust*). Using scope information only leads to significant improvements on the TAC 2012 data set with respect to the accuracy.

Figure 9.2 shows the performance of the same system variants in terms of  $B^{3+}F$  and thus also accounts for the clustering performance. The joint concept disambiguation and clustering system (*ML Dis+NILs+Clust*) significantly improves the corresponding cascaded system (*ML Dis+NILs*) on all three data sets, while the scope-aware system significantly improves the

results only on the TAC 2012 data sets. As discussed above, the string match and singleton clustering baselines are very strong on the TAC 2011 and TAC 2012 data sets respectively. Our clustering-only system (*ML Clustering*) significantly outperforms both of them only on the TAC 2013 data set. However, our joint concept disambiguation and clustering system obtains significantly higher results than both baselines on all data sets except the TAC 2012 data set. On this data set the  $B^3F$  score is higher for the baseline that considers all mentions as singletons.

Hence, the results on the TAC data sets support our claim that our joint concept disambiguation and clustering approach improves the quality of the disambiguation and the clustering (Contribution 1, 3). In Section 9.4.2, we will discuss why scope information does not consistently improve the results on these data sets.

*Related Work.* The TAC data sets allow us to compare our results to the results of different other approaches. As on all other data sets, we provide a fair comparison with Wikipedia-Miner (Milne & Witten, 2008b) (*WM*). On all three data sets, our full system significantly outperforms WikipediaMiner by more than 7 points in terms of accuracy. As the results of our different system variants suggest, these improvements are to a big portion due to our joint disambiguation and clustering approach. Compared to the results of the systems that participated in the shared task our system performs well. On all data sets, the results of our full system are well above the median ( $>10$  points in terms of  $B^{3+}$ ) and close to the results of the best system in the shared task. In particular, with respect to the accuracy, our system is close to the best system. We participated in the TAC 2012 and TAC 2013 with slight variants of *ML Dis+NILs+Clust* with similar results as reported here. In the TAC 2013 English entity linking task, our system was among the top five performing systems out of 27 systems in terms of  $B^{3+}F$ . It is striking that all top performing system at TAC make use of clustering information during disambiguation (e.g. *LCC* (Monahan et al., 2011; Monahan & Carpenter, 2012), *MS MLI* (Cucerzan, 2012; 2013; Cucerzan & Sil, 2014), *Cogcomp* (Cheng & Roth, 2013; Cheng et al., 2014), *Blender CUNY* (Tamang et al., 2013; Yu et al., 2014)). This strengthens our claim that the two tasks support each other. The best performing system at TAC 2012 and TAC 2013 (*MS MLI*) is based on the approach of Cucerzan (2007) which also uses different context definitions for different mentions. Although their context selection strategy is different from ours, their high performance supports our claim that different mentions require different context definitions. Compared to the best systems, our system uses fewer features. In addition, our parameters are learned on Wikipedia and not on TAC data. As on the ACE 2004 data, our system performs in the same range as the system of Cheng & Roth (2013). As discussed above, this system is closely related to ours and supports our claims. Our full system is superior to the *Basistech* system (Clarke et al., 2012; Merhav et al., 2013) which cast the disambiguation task as a clustering problem. In contrast to our joint system, they first resolve coreference within each document and then cluster the

whole coreference chain. While this approach might be suitable for proper names, it is not applicable to common nouns. The *Blender CUNY* system which obtains better performance than our system on the TAC 2012 data set, but is outperformed by ours on the TAC 2013 data sets, implements a cascaded approach (Chen & Ji, 2011). First mentions are clustered, then all mentions in one cluster are disambiguated to the same concept.

*Discussion.* The evaluation on the English TAC data sets from the entity linking shared task 2011-2013 reveals that our approach is highly competitive. On all data sets, it obtains results close to the results of the top performing systems. We also evaluated our approach on the TAC 2009 and TAC 2010 data sets, which are not annotated for clustering of the NILs. Our full system obtains an accuracy of 83.9 and 87.2 on the TAC 2009 and TAC 2010 data sets respectively. These results, which exceed the results of the best systems at these shared tasks by 1.7 (TAC 2009) and 1.4 points (TAC 2010), further confirm the competitiveness of our approach. The analysis of the contributions of our different components shows that a considerable portion of our performance gains over our basic approach is due to the joint disambiguation and clustering.

	In Inventory			NILs			Acc	Clustering					
	P <sub>inv</sub>	R <sub>inv</sub>	F <sub>inv</sub>	P <sub>NILs</sub>	R <sub>NILs</sub>	F <sub>NILs</sub>		B <sup>3</sup> P	B <sup>3</sup> R	B <sup>3</sup> F	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F
TAC EN 2011													
Upper bound Mention	91.2	75.9	82.9	85.6	100.0	92.3	88.0	98.3	98.3	98.3	87.8	87.8	87.0
Upper bound Document	98.1	94.5	96.2	96.5	100.0	98.2	97.2	99.8	99.6	99.7	97.1	97.1	97.1
(1) First Concept	61.5	55.8	58.5	78.2	85.4	81.7 <sup>2,3,5</sup>	70.6 <sup>2,3,5</sup>	89.2	99.1	93.9 <sup>3,5</sup>	66.0	70.2	68.1 <sup>2,3,5</sup>
(2) String Match Clustering	0.0	0.0	0.0	50.0	100.0	66.7	50.0	89.3	98.8	93.8 <sup>3,5</sup>	45.4	49.6	47.4 <sup>3</sup>
(3) Singleton Clustering	0.0	0.0	0.0	50.0	100.0	66.7	50.0	100.0	67.3	80.4	50.0	35.9	41.8
Median (all participants at TAC) <sup>*</sup>	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	71.6
LCC (Monahan et al., 2011) (Best) <sup>*</sup>	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	86.1	n.a.	n.a.	n.a.	84.4	84.7	84.6
MS MLI (Cucerzan, 2012) <sup>*</sup>	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	86.8	n.a.	n.a.	n.a.	84.8	83.4	84.1
Cogcomp (Ratinov & Roth, 2011) <sup>*</sup>	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	79.4	n.a.	n.a.	n.a.	76.3	77.3	76.8
Cheng & Roth (2013)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	86.1	n.a.	n.a.	n.a.	82.9	84.5	83.7
(4) WM (Milne & Witten, 2008b)	84.1	56.4	67.5 <sup>1,6</sup>	71.2	94.6	81.2 <sup>2,3,5</sup>	75.5 <sup>1-3,5</sup>	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
(5) ML Clustering	0.0	0.0	0.0	50.0	100.0	66.7	50.0	98.4	84.5	91.0 <sup>3</sup>	49.1	45.0	46.9 <sup>3</sup>
(6) ML Dis WM	67.4	62.5	64.8 <sup>1</sup>	81.0	86.9	83.9 <sup>1-5</sup>	74.7 <sup>1-3,5</sup>	91.2	97.3	94.1 <sup>3,5</sup>	70.9	73.3	72.1 <sup>1-3,5</sup>
(7) ML Dis+NILs	85.5	63.7	73.0 <sup>1,4,6</sup>	75.4	94.6	83.9 <sup>1-5</sup>	79.2 <sup>1-6</sup>	95.8	95.3	95.6 <sup>1-3,5,6</sup>	77.0	76.7	76.8 <sup>1-3,5,6</sup>
(8) ML Dis+NILs+Clust	85.9	76.2	80.8 <sup>1,4,6,7</sup>	83.9	93.3	88.4 <sup>1-7</sup>	84.8 <sup>1-7</sup>	95.7	96.8	96.3 <sup>1-3,5-7</sup>	82.5	83.0	82.8 <sup>1-3,5-7</sup>
(9) ML Dis+NILs+Clust Scope	89.3	76.1	82.2 <sup>1,4,6-8</sup>	82.6	94.8	88.3 <sup>1-7</sup>	85.5 <sup>1-7</sup>	96.8	95.3	96.0 <sup>1-3,5-7</sup>	83.6	82.7	83.1 <sup>1-3,5-7</sup>

Table 9.5: Results on the English TAC 2011 data set: besides the results of our approach, we report the results for the best system at TAC 2011 (Best), the median performance of all systems that participated at TAC 2011 (Median) and the results of closely related approaches. An asterisk (\*) besides the system name indicates that the results are official shared task results. Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $B^3F1$ ,  $B^{3+}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

	In Inventory			NILs			Acc	Clustering					
	P <sub>inv</sub>	R <sub>inv</sub>	F <sub>inv</sub>	P <sub>NILs</sub>	R <sub>NILs</sub>	F <sub>NILs</sub>		B <sup>3</sup> P	B <sup>3</sup> R	B <sup>3</sup> F	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F
TAC EN 2012													
Upper bound Mention	70.6	58.9	64.2	84.3	100.0	91.5	78.3	86.8	97.4	91.8	76.7	77.2	76.9
Upper bound Document	93.6	90.5	92.0	96.4	100.0	98.2	95.0	97.0	99.4	98.2	93.8	94.6	94.2
(1) First Concept	21.5	26.5	23.8	69.5	51.6	59.2	38.3	36.4	93.4	52.4	19.1	35.2	24.8
(2) String Match Clustering	0.0	0.0	0.0	47.1	100.0	64.1 <sup>1</sup>	47.1 <sup>1</sup>	36.8	92.4	52.6	17.6	42.5	24.9
(3) Singleton Clustering	0.0	0.0	0.0	47.1	100.0	64.1 <sup>1</sup>	47.1 <sup>1</sup>	100.0	87.2	93.2 <sup>1,2,5-9</sup>	47.1	41.6	44.2 <sup>1,2,5,6</sup>
Median (all participants at TAC)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	60.1	n.a.	n.a.	n.a.	n.a.	n.a.	53.6
MS MLI (Cucerzan, 2013) (Best)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	76.6	n.a.	n.a.	n.a.	n.a.	n.a.	73.0
LCC (Monahan & Carpenter, 2012)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	75.7	n.a.	n.a.	n.a.	n.a.	n.a.	68.9
Blender CUNY (Tamang, 2013)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	68.8
Basistech (Clarke et al., 2012)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	56.9
UMass (Dietz & Dalton, 2013)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	62.6	n.a.	n.a.	n.a.	n.a.	n.a.	56.3
(4) WM (Milne & Witten, 2008b)	55.7	32.9	41.3 <sup>1,6</sup>	59.0	86.1	70.0 <sup>1-3,5,6</sup>	58.0 <sup>1-3,5,6</sup>	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
(5) ML Clustering	0.0	0.0	0.0	47.1	100.0	64.1 <sup>1</sup>	47.1 <sup>1</sup>	89.1	89.6	89.3 <sup>1,2,6,7</sup>	42.3	42.4	42.4 <sup>1,2,6</sup>
(6) ML Dis WM	30.5	37.6	33.7 <sup>1</sup>	73.3	54.3	62.4 <sup>1</sup>	45.5 <sup>1</sup>	46.4	93.2	61.9 <sup>1,2</sup>	27.6	42.3	33.4 <sup>1,2</sup>
(7) ML Dis+NILs	50.8	35.9	42.1 <sup>1,6</sup>	63.0	83.6	71.8 <sup>1-6</sup>	58.4 <sup>1-3,5,6</sup>	76.3	91.9	83.4 <sup>1,2,6</sup>	49.2	54.2	51.6 <sup>1-3,5,6</sup>
(8) ML Dis+NILs+Clust	62.2	57.1	59.5 <sup>1,4,6,7</sup>	73.6	80.5	76.9 <sup>1-7</sup>	68.1 <sup>1-7</sup>	83.3	92.8	87.8 <sup>1,2,6,7</sup>	61.7	64.3	63.0 <sup>1-3,5-7</sup>
(9) ML Dis+NILs+Clust Scope	67.2	58.2	62.4 <sup>1,4,6-8</sup>	72.0	82.8	77.0 <sup>1-7</sup>	69.8 <sup>1-8</sup>	86.0	93.0	89.3 <sup>1,2,6,7</sup>	63.5	65.9	64.7 <sup>1-3,5-8</sup>

Table 9.6: Results on the English TAC 2012 data set: besides the results of our approach, we report the results for the best system at TAC 2012 (Best), the median performance of all systems that participated at TAC 2012 (Median) and the results of closely related approaches. An asterisk (\*) besides the system name indicates that the results are official shared task results. Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $B^3F1$ ,  $B^{3+}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

	In Inventory			NILs			Acc	Clustering					
	$P_{inv}$	$R_{inv}$	$F_{inv}$	$P_{NILs}$	$R_{NILs}$	$F_{NILs}$		$B^3 P$	$B^3 R$	$B^3 F$	$B^{3+} P$	$B^{3+} R$	$B^{3+} F$
TAC EN 2013													
Upper bound Mention	90.3	68.1	77.6	77.5	100.0	87.3	82.7	98.4	86.0	91.8	82.6	77.2	79.8
Upper bound Document	97.6	85.2	91.0	86.9	100.0	93.0	91.9	99.7	95.6	97.6	91.9	90.0	90.9
(1) First Concept	71.1	66.4	68.7	74.5	80.2	77.2 <sup>2,3,5</sup>	72.8 <sup>2,3,5</sup>	86.0	59.3	70.2 <sup>2,3,5</sup>	65.9	48.7	56.0 <sup>2,3,5</sup>
(2) String Match Clustering	0.0	0.0	0.0	46.0	100.0	63.0	46.0	88.7	40.7	55.8 <sup>3</sup>	38.4	19.3	25.7
(3) Singleton Clustering	0.0	0.0	0.0	46.0	100.0	63.0	46.0	100.0	33.1	49.7	46.0	17.2	25.0
Median (all participants at TAC)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	74.6	n.a.	n.a.	n.a.	71.8	49.6	57.4
MS MLI ((Cucerzan, 2014) (Best)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	83.3	n.a.	n.a.	n.a.	82.6	68.9	74.6
UI CCG (Cheng et al., 2014)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	80.5	n.a.	n.a.	n.a.	77.6	62.8	69.4
UMass (Dalton & Dietz, 2014)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	80.6	n.a.	n.a.	n.a.	78.5	58.4	67.0
Blender CUNY (Yu et al., 2014)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	63.7
Basistech (Merhav et al., 2013)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	63.3
(4) WM (Milne & Witten, 2008b)	85.3	59.8	70.3	67.7	91.6	77.9 <sup>2,3,5</sup>	74.4 <sup>1,2,3,5</sup>	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
(5) ML Clustering	0.0	0.0	0.0	46.0	100.0	63.0	46.0	96.2	43.0	59.4 <sup>2,3</sup>	43.4	22.6	29.7 <sup>2,3</sup>
(6) ML Dis WM	73.1	69.4	71.2 <sup>1</sup>	76.1	80.6	78.3 <sup>1,2,3,5</sup>	74.6 <sup>1,2,3,5</sup>	87.6	61.8	72.5 <sup>1-3,5</sup>	68.4	51.9	59.0 <sup>1-3,5</sup>
(7) ML Dis+NILs	86.0	69.6	76.9 <sup>1,4,6</sup>	74.6	91.4	82.1 <sup>1-6</sup>	79.6 <sup>1-6</sup>	94.1	62.6	75.2 <sup>1-3,5,6</sup>	76.1	55.6	64.2 <sup>1-3,5,6</sup>
(8) ML Dis+NILs+Clust	84.1	76.2	80.0 <sup>1,4,6,7</sup>	79.5	88.2	83.6 <sup>1-7</sup>	81.7 <sup>1-7</sup>	93.0	67.7	78.4 <sup>1-3,5-7</sup>	77.5	61.3	68.4 <sup>1-3,5-7</sup>
(9) ML Dis+NILs+Clust Scope	87.3	73.8	80.0 <sup>1,4,6,7</sup>	77.2	91.3	83.7 <sup>1-7</sup>	81.8 <sup>1-7</sup>	94.7	66.2	77.9 <sup>1-3,5-7</sup>	78.6	60.1	68.1 <sup>1-3,5-7</sup>

Table 9.7: Results on the English TAC 2013 data set: besides the results of our approach, we report the results for the best system at TAC 2013 (Best), the median performance of all systems that participated at TAC 2013 (Median) and the results of closely related approaches. An asterisk (\*) besides the system name indicates that the results are official shared task results. Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $B^3 F1$ ,  $B^{3+}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

### 9.3.2 Multi- and Cross-lingual Results

In this section, we report the results of our approach on other languages than English. We focus on Spanish and Chinese, but also report cross-lingual results for English to Japanese and English to Korean. To evaluate our system, we use the data sets from two shared tasks, i.e. from the cross-lingual entity linking task at TAC and the cross-lingual link discovery task at NTCIR.

**Spanish TAC Data Sets.** The Spanish TAC data sets allow us to evaluate our system in a multi- and cross-lingual setting. While most of the texts are in Spanish, a few are in English (Table 8.2, Section 8.4.2). The task is to link mentions to an inventory derived from the English Wikipedia and to cluster mentions across languages. This task has been organized for the first time at TAC 2012. We report our results on all available data sets, i.e. on the Spanish cross-lingual 2012 and 2013 data sets, in Table 9.8.

*Upper Bounds and Baseline.* The mention- and document-based upper bounds for the Spanish cross-lingual data sets are comparable to the English monolingual upper bounds. As in the monolingual setting, the mention-based upper bound is lower than the document-based upper bound. On the Spanish 2012 and 2013 data sets the accuracy for the document-based upper bound is about 90. These results indicate that our candidate identification leads to similar results both for English and Spanish. The first concept baseline on both Spanish data sets is at the same level as in the English monolingual setting. Hence, our prior probabilities are of high quality for Spanish and are comparable with the prior probabilities we extracted for English, although the Spanish Wikipedia is smaller. The results for the clustering baselines are also comparable to the corresponding English monolingual results. While on the Spanish 2012 data set the singleton baseline is strong, the string match baseline leads to good results on the Spanish 2013 data set. Overall, the upper bounds and baselines are portable to the Spanish cross-lingual setting.

*Our Results.* As the aim of the multi- and cross-lingual experiments is to show how our approach scales across languages and less to analyze the contribution of the different components of our system, we only report the results of our joint concept disambiguation and clustering approach (*ML Dis+NILs+Clust*). On both data sets, we trained on our English training data obtained from Wikipedia and only adapted a few features as described in Table 7.4 (Section 7.1). On both Spanish cross-lingual data sets, our results are within the range of our English monolingual results both with respect to the disambiguation and the clustering. The first concept baseline is significantly outperformed by our approach on both data sets. As in the English monolingual setting, the results for the Spanish TAC 2012 data set are lower than for the Spanish TAC 2013 data sets. The lower results for the first concept baseline on the TAC 2012 indicate that this data set is more difficult.

*Related Work.* As in the English monolingual setting, we compare our results to related approaches. An asterisk behind the system name means that the corresponding results are official shared task results. On the Spanish TAC 2012 data sets, our results are close to the one of the best system from LCC (Monahan & Carpenter, 2012), which is an adapted cross-lingual version of their English monolingual system. They also exploit the cross-language links in Wikipedia and perform the disambiguation in the source language. The system from Basistech (Clarke et al., 2012) that models the disambiguation task as a clustering problem performs worse than ours. On the Spanish TAC 2013 data set, a variant of our system with which we participated in this shared task outperformed all other participating systems. The results for our current system are slightly worse, as we use one cross-document clustering feature less for the sake of efficiency. As on the TAC 2012 data set, the results of Basistech are worse than ours.

*Discussions.* Our results for the Spanish cross-lingual entity linking task at TAC indicate that our approach scales well to Spanish. The statistics we obtain from the Spanish Wikipedia (prior probabilities) are of high quality and the results of the upper bounds, baselines and our system are in the range of the English monolingual results. A comparison with the best systems at TAC shows that our results are state-of-the-art, even though we did not retrain on Spanish data. This also indicates that our feature weights scale well from English to Spanish.

	In Inventory			NILs			Acc	Clustering					
	P <sub>inv</sub>	R <sub>inv</sub>	F <sub>inv</sub>	P <sub>NILs</sub>	R <sub>NILs</sub>	F <sub>NILs</sub>		B <sup>3</sup> P	B <sup>3</sup> R	B <sup>3</sup> F	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F
TAC ES 2012													
Upper bound Mention	92.4	76.4	83.6	87.7	100.0	93.5	89.4	94.8	99.1	96.9	89.2	89.2	89.2
Upper bound Document	96.0	90.1	93.0	95.3	100.0	97.6	95.6	98.9	99.5	99.2	95.4	95.4	95.4
(1) First Concept	30.2	32.5	31.3	70.6	66.3	68.4	51.2	38.7	95.9	55.1	21.6	48.6	29.9 <sup>2</sup>
(2) String Match Clustering	0.0	0.0	0.0	55.3	100.0	71.2 <sup>1</sup>	55.3 <sup>1</sup>	40.2	94.8	56.5 <sup>1</sup>	17.7	52.4	26.5
(3) Singleton Clustering	0.0	0.0	0.0	55.3	100.0	71.2 <sup>1</sup>	55.3 <sup>1</sup>	100.0	74.4	85.3 <sup>1,2</sup>	55.3	44.5	49.3 <sup>1,2</sup>
LCC (Monahan & Carpenter, 2012) (Best) <sup>*</sup>	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	64.1
Basistech (Clarke et al., 2012) <sup>*</sup>	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	59.9
(4) ML Dis+NILs+Clust	58.5	58.1	58.3 <sup>1</sup>	79.5	80.0	79.8 <sup>1-3</sup>	70.2 <sup>1-3</sup>	83.3	87.3	85.3 <sup>1,2</sup>	62.5	62.5	62.5 <sup>1-3</sup>
TAC ES 2013													
Upper bound Mention	94.5	79.7	86.5	79.8	100.0	88.8	87.5	98.6	88.5	93.2	87.1	82.8	84.9
upper bound Doc	97.1	91.3	94.1	91.1	100.0	95.4	94.6	99.3	96.5	97.9	94.3	93.1	93.7
(1) First Concept	77.7	70.4	73.9	71.9	82.8	77.0 <sup>2,3</sup>	75.2 <sup>2,3</sup>	89.9	69.4	78.3 <sup>2,3</sup>	71.1	56.1	62.7 <sup>2,3</sup>
(2) String Match Clustering	0.0	0.0	0.0	38.4	100.0	55.4	38.4	93.3	51.4	66.3 <sup>3</sup>	34.1	19.8	25.0
(3) Singleton Clustering	0.0	0.0	0.0	38.4	100.0	55.4	38.4	100.0	46.2	63.2	38.4	18.6	25.1
Ours (Fahrni et al., 2014) (Best) <sup>*</sup>	85.6	77.6	81.4	76.4	87.8	81.7	81.5	93.0	76.0	83.7	78.1	64.9	70.9
Basistech (Merhav et al., 2013) <sup>*</sup>	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	65.1
(4) ML Dis+NILs+Clust	84.6	77.3	80.8 <sup>1</sup>	76.6	87.3	81.6 <sup>1-3</sup>	81.2 <sup>1-3</sup>	92.7	75.6	83.3 <sup>1-3</sup>	77.5	64.4	70.4 <sup>1-3</sup>

Table 9.8: Results on the Spanish cross-lingual TAC 2012 and 2013 data sets: besides the results of our approach, we report the results for the best system at TAC 2012 and 2013 (Best) and the results of closely related approaches. An asterisk (\*) besides the system name indicates that the results are official shared task results. Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $B^3F1$ ,  $B^{3+}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

**Chinese TAC Data Sets.** As the Spanish TAC data sets, the Chinese TAC data sets allow us to evaluate our system in a multi- and cross-lingual setting. Some of the texts are in Chinese, some in English and as in the Spanish cross-lingual task, the aim is to link mentions to an inventory derived from the English Wikipedia and to cluster mentions across languages. This task has been organized for the first time at TAC 2011. We report our results on all available data sets, i.e. on the Chinese cross-lingual 2011, 2012 and 2013 data sets, in Table 9.9 and 9.10.

*Upper Bounds and Baselines.* The mention- and document-based upper bounds are in the same range for Chinese as for English and Spanish. The accuracies for the document-based upper bound are all above 90 and higher than the mention-based upper bound. Hence, although Chinese uses a completely different writing system than English and Spanish, our results are still high. However, as described in Table 7.1 (Section 7.4), we did some customizations for the candidate concept identification step and look up both the simplified and the traditional Chinese version of a string. The accuracies for the first concept baseline are all above 70. This is surprising as the Chinese Wikipedia from where we extracted the prior probabilities is much smaller than for English and Spanish. Nevertheless, the extracted prior probabilities seem to be reliable. However, we cannot derive any conclusions about the number of annotated instances required to obtain reliable statistics for disambiguation, as it could be that the Chinese TAC data is less difficult than the data sets for the other languages. The results of the singleton and string match baselines for the clustering are also comparable to the corresponding results for the other languages. Hence, string match seems to be a strong clustering baseline, even for Chinese.

*Our Results.* On all Chinese data sets, our approach significantly outperforms the first concept baseline in terms of  $F_{inv}$  and  $Acc$ . Except on the Chinese 2011 data set where the first concept baseline in combination with string match clustering obtains a lower, but not statistically significantly different  $B^{3+}F$  score than our joint approach, our joint approach significantly outperforms the first concept baseline also with respect to the clustering. Our Chinese cross-lingual results are in the range of our monolingual English results. Hence, our approach is portable to Chinese with good results. However, some of the clustering features are slightly adapted for Chinese (Table 7.1, Section 7.4).

*Related Work.* Compared to related work, our Chinese cross-lingual approach performs favorably. While on the TAC 2011 data, the approach of LCC (Monahan & Carpenter, 2012) outperforms our approach, our approach shows better results than theirs on the Chinese TAC 2012 data set. As for Spanish and English, our approach also outperforms the clustering-based approach from Basistech (Clarke et al., 2012) on the Chinese TAC 2012 data. On this data set, we obtained the highest results out of all participating systems with a variant of our joint disambiguation and clustering approach. On the Chinese TAC 2013 data set, the results from Basistech are slightly higher than ours. In contrast to us, they use translation information and

the Google Cross-wiki dataset (Spitkovsky & Chang, 2012).

*Discussion.* The Chinese cross-lingual results for the upper bounds, baselines and our system are comparable to the English monolingual results. These results indicate that the statistics we extracted from the Chinese Wikipedia are reliable, even though the Chinese Wikipedia is much smaller than the English and Spanish versions. While the feature weights are also portable from English to Chinese, some of the features and the candidate identification component are slightly adapted for Chinese. Overall, our Chinese cross-lingual system leads to state-of-the-art results.

	In Inventory			NILs			Acc	Clustering					
	P <sub>inv</sub>	R <sub>inv</sub>	F <sub>inv</sub>	P <sub>NILs</sub>	R <sub>NILs</sub>	F <sub>NILs</sub>		B <sup>3</sup> P	B <sup>3</sup> R	B <sup>3</sup> F	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F
TAC ZH 2011													
Upper bound Mention	88.5	77.3	82.5	88.8	100.0	94.1	88.7	98.4	97.4	97.9	88.4	88.1	88.2
Upper bound Document	92.2	86.4	89.2	94.1	100.0	96.9	93.2	99.0	98.0	98.5	93.0	92.5	92.8
(1) First Concept	68.2	66.7	67.5	83.9	85.8	84.9 <sup>2,3</sup>	76.3 <sup>2,3</sup>	86.3	96.2	91.0 <sup>2,4</sup>	68.0	73.9	70.9 <sup>2,3</sup>
(2) String Match Clustering	0.0	0.0	0.0	50.1	100.0	66.8	50.1	87.2	91.3	89.2 <sup>3</sup>	40.6	47.5	43.8 <sup>3</sup>
(3) Singleton Clustering	0.0	0.0	0.0	50.1	100.0	66.8	50.1	100.0	45.0	62.1	50.1	22.5	31.0
LCC (Monahan et al., 2011) (Best)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	78.6	79.0	78.8
(4) ML Dis+NILs+Clust	78.4	69.6	73.7 <sup>1</sup>	80.2	89.2	84.5 <sup>2,3</sup>	79.4 <sup>1-3</sup>	86.7	91.0	88.8 <sup>3</sup>	70.2	75.5	72.7 <sup>2,3</sup>
TAC ZH 2012													
Upper bound Mention	95.2	80.5	87.2	82.1	100.0	90.2	88.6	99.3	86.4	92.4	88.5	83.4	85.9
Upper bound Document	99.2	95.2	97.2	94.6	100.0	97.2	97.2	99.7	97.0	98.3	97.2	96.1	96.6
(1) First Concept	82.4	74.8	78.4	77.1	87.2	81.9 <sup>2,3</sup>	79.9 <sup>2,3</sup>	86.4	77.2	81.6 <sup>2,3</sup>	70.6	67.8	69.2 <sup>2,3</sup>
(2) String Match Clustering	0.0	0.0	0.0	41.6	100.0	58.7	41.6	87.6	47.1	61.3 <sup>3</sup>	31.8	26.9	29.1 <sup>3</sup>
(3) Singleton Clustering	0.0	0.0	0.0	41.6	100.0	58.7	41.6	100.0	23.9	38.6	41.6	14.4	21.4
Ours (Fahrni et al., 2013) (Best) *	88.7	79.4	83.8	79.5	91.2	84.9	84.3	86.3	81.1	83.6	73.8	74.2	74.0
LCC (Monahan & Carpenter, 2012)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	80.2	n.a.	n.a.	n.a.	n.a.	n.a.	66.8
Basistech (Clarke et al., 2012)*	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	56.6
(4) ML Dis+NILs+Clust	90.7	79.0	84.4 <sup>1</sup>	78.7	92.9	85.2 <sup>1-3</sup>	84.8 <sup>1-3</sup>	91.6	78.9	84.8 <sup>1-3</sup>	78.8	72.7	75.6 <sup>1-3</sup>

Table 9.9: Results on the Chinese cross-lingual TAC 2011 and 2012 data sets: besides the results of our approach, we report the results for the best system at TAC 2011 and 2012 (Best) and the results of closely related approaches. An asterisk (\*) besides the system name indicates that the results are official shared task results. Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $B^3F1$ ,  $B^{3+}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

	In Inventory			NILs			Acc	Clustering					
	$P_{inv}$	$R_{inv}$	$F_{inv}$	$P_{NILs}$	$R_{NILs}$	$F_{NILs}$		$B^3 P$	$B^3 R$	$B^3 F$	$B^{3+} P$	$B^{3+} R$	$B^{3+} F$
TAC ZH 2013													
Upper bound Mention	86.3	67.4	75.7	78.2	100.0	87.8	81.8	95.9	80.5	87.5	81.1	75.2	78.0
Upper bound Document	93.6	86.3	89.8	91.0	100.0	95.3	92.3	97.6	92.4	94.9	91.9	89.5	90.7
(1) First Concept	70.9	60.6	65.4	71.7	84.9	77.8 <sup>2,3</sup>	71.3 <sup>2,3</sup>	75.4	66.6	70.7 <sup>2,3</sup>	56.1	53.9	55.0 <sup>2,3</sup>
(2) String Match Clustering	0.0	0.0	0.0	44.0	100.0	61.1	44.0	77.1	41.2	53.7 <sup>3</sup>	28.2	23.9	25.9 <sup>3</sup>
(3) Singleton Clustering	0.0	0.0	0.0	44.0	100.0	61.1	44.0	100.0	15.5	26.8	44.0	8.2	13.8
Basistech (Merhav et al., 2013) (Best) <sup>*</sup>	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	63.1
(4) ML Dis+NILs+Clust	87.3	65.8	75.0 <sup>1</sup>	69.9	91.9	79.4 <sup>1-3</sup>	77.3 <sup>1-3</sup>	83.4	64.8	72.9 <sup>2,3</sup>	64.9	55.8	60.0 <sup>1-3</sup>

Table 9.10: Results on the Chinese cross-lingual TAC 2013 data sets: besides the results of our approach, we report the results for the best system at TAC 2013 (Best). An asterisk (\*) besides the system name indicates that the results are official shared task results. Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $B^3 F1$ ,  $B^{3+}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

**NTCIR 9 Data Sets.** The NTCIR 9 data sets have been used in the cross-lingual link discovery task at NTCIR 9. These data sets allow us to evaluate the capability of our system to link mentions in an English text to concepts in an inventory that is in another language than English. More precisely, we link from English to Japanese, from English to Korean and from English to Chinese. In the cross-lingual link discovery task at NTCIR 9, the systems should only link keywords. Hence, the mentions should be ranked according to their importance. This ranking is considered by the evaluation metrics (MAP and precision-at-5) and highly influences the final performance of a system. We used our joint discourse-aware concept disambiguation and clustering system and ranked the mentions recognized by our mention recognition by their *keyphraseness* (Csomai & Mihalcea, 2008). *Keyphraseness* has been introduced by Csomai & Mihalcea (2008) and is a Wikipedia-based measure to rank keywords. The keyphraseness of a mention  $m$  is given by

$$\text{keyphraseness}(m) = \frac{\text{occ}(m_{\text{linked}})}{\text{occ}(m)}$$

where  $\text{occ}(m_{\text{linked}})$  is the number of times mention  $m$  is linked in Wikipedia and  $\text{occ}(m)$  is the number of times the mention  $m$  occurs in Wikipedia. We use the English Wikipedia to extract these statistics.

The NTCIR 9 data set consists of Wikipedia articles. To allow for a fair comparison, we exclude the information we extracted from the corresponding Wikipedia articles (e.g. hyperlinks) in these experiments. In addition, we made sure that no article in our training data is part of the NTCIR 9 data set. The results are shown in Table 9.11.

*Baseline.* As our upper bounds are not applicable to the NTCIR setting and no clustering information is available, we only report the results of the first concept baselines. The mentions are ranked by keyphraseness. The results of the first concept baseline are high for all languages.

*Our Results.* Our system (*ML Dis+NILs+Clust Scope*) outperforms the first concept baseline for all target languages with respect to *MAP*, although only by a small margin. As our analysis revealed, the ranking of keywords is highly important for this task and affects the results more than the disambiguation method. Hence, the data set is less suitable to evaluate the disambiguation performance, but more to evaluate the performance of the keyword ranking.

*Related Work.* We report the median and performance of the best system from the shared task which is a graph-based approach proposed by us (Fahrni et al., 2011b). The results of this graph-based approach and our current approach are very similar. As stated above, the data set is not suitable to compare different disambiguation algorithms, as these factors are masked by the keyword ranking capabilities. However, our system outperforms all other cross-lingual linking systems at NTCIR in this evaluation setting (Tang et al., 2011) and achieves state-of-the-art results on this task. In contrast to all other systems, we use disambiguation. Hence,

	EN to JA		EN to KO		EN to ZH	
	MAP	P@5	MAP	P@5	MAP	P@5
First Concept	32.7	84.0	46.0	80.8	37.8	<b>83.2</b>
Ours (Fahmi et al., 2011b) (Best)*	31.6	84.0	44.7	<b>84.8</b>	37.3	<b>83.2</b>
Median	23.5	56.8	31.8	68.0	17.9	57.6
ML Dis+NILs+Clust Scope	<b>33.7</b>	<b>84.8</b>	<b>47.4</b>	80.8	<b>39.1</b>	<b>83.2</b>

Table 9.11: Results on the cross-lingual link discovery task at NTCIR 9. We report the results for all three language pairs based on the Wikipedia ground truth. Besides the results of our approach, we report the results for the best system at NTCIR 9 (Best). An asterisk (\*) besides the system name indicates that the results are official shared task results.

disambiguation helps the task, but already a simple method seems to achieve good results.

*Discussion.* Our results on the data from the cross-lingual link discovery task show that our system can easily be adapted for the task of hyperlink insertion. In addition, it shows that our cross-lingual lexicon is of high quality and leads to good results for Japanese, Korean and Chinese.

## 9.4 Analysis

While we showed the performance of our systems on different data sets and the contributions of different components of it to the overall performance in the last section, we analyze some aspects of it in more detail in this section. We focus on the main contributions of this thesis and further investigate the effect of the joint concept disambiguation and clustering (Section 9.4.1), the scope awareness (Section 9.4.2) and further investigate the multi- and cross-lingual strategies (Section 9.4.3). In Section 9.4.4, we discuss some typical errors made by our system. In contrast to the last section, we focus on a few data sets and also show some examples.

### 9.4.1 Joint Concept Disambiguation and Clustering

In the last section, we have shown that our joint approach for concept disambiguation and clustering significantly improves the corresponding cascaded system that first performs disambiguation and then clusters the NILs on different data sets. In this section, we further analyze these results. For brevity, we mainly restrict ourselves to one single data set in this analysis. In contrast to all other data sets, the TAC data sets are annotated for both disambiguation and clustering. We therefore choose one of the TAC data set for our analysis, more precisely the English TAC 2011 data set.

	In Inventory			NILs			Acc
	$P_{inv}$	$R_{inv}$	$F_{inv}$	$P_{NILs}$	$R_{NILs}$	$F_{NILs}$	
TAC EN 2011							
(1) ML Dis+NILs	85.5	63.7	73.0	75.4	94.6	83.9	79.2
(2) ML ClustFirst	82.6	73.3	77.7 <sup>1</sup>	82.6	91.9	87.0 <sup>1</sup>	82.6 <sup>1</sup>
(3) ML Dis+NILs+Clust	85.9	76.2	<b>80.8<sup>1,2</sup></b>	83.9	93.3	<b>88.4<sup>1,2</sup></b>	<b>84.8<sup>1,2</sup></b>

Table 9.12: Clustering first vs. clustering second vs. joint disambiguation and clustering: disambiguation results on the English TAC 2011 data set. Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

**Cascaded vs. Joint Disambiguation and Clustering.** In the last section, we compared our joint concept disambiguation and clustering approach to a cascaded disambiguation and clustering approach in which disambiguation is performed before clustering. However, a cascaded approach could also perform the clustering first and then do the disambiguation. For instance Chen & Ji (2011) proposed to first do the clustering and then the disambiguation. We thus implemented a cascaded approach in Markov logic that uses the same features as our joint (*ML Dis+NILs+Clust*) and the other cascaded approaches (*ML Dis+NILs*) and applied it to our data. We call this system *ML ClustFirst*. It first clusters all the mentions. In the second step, we enforce that all mentions in the same cluster are linked to the same concept or are all NILs.

Table 9.12 shows the disambiguation results of the two cascaded systems (*ML Dis+NILs* and *ML ClustFirst*) and our joint approach (*ML Dis+NILs+Clust*) on the TAC 2011 data set. The cascaded system that first clusters all mentions performs significantly better than the cascaded system that first disambiguates all mentions and then clusters the NILs. Our joint approach significantly outperforms both cascaded systems and leads to the highest disambiguation results. Table 9.13 shows the clustering results for the same systems. While  $B^3F$  scores are not significantly different for the two cascaded approaches, the  $B^{3+}F$  scores are significantly higher for the cascaded approach that first clusters the mentions. The joint approach leads to significantly higher clustering results than both cascaded approaches.

One reason why the clustering first strategy works better than the clustering second strategy is that with the clustering first strategy, candidate concepts are shared between mentions in the same cluster. In the clustering second strategy, each mention is disambiguated separately.

If we only consider the clustering results for the NILs and not all mentions, our joint concept disambiguation approach achieves a  $B^3F$  score of 95.9 on the TAC 2011 test set, while the clustering first strategy leads to an  $B^3F$  score of 94.0.<sup>6</sup> These results show that the overall clustering results are not only better due to a better disambiguation of the joint concept

<sup>6</sup>These results are reported on the true NILs, excluding NIL queries that can be clustered via linking to a newer Wikipedia dump (Cucerzan, 2013).

	Clustering					
	B <sup>3</sup> P	B <sup>3</sup> R	B <sup>3</sup> F	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F
TAC EN 2011						
(1) ML Dis+NILs	95.8	95.3	95.6	77.0	76.7	76.8
(2) ML ClustFirst	93.9	96.5	95.2	78.9	80.8	79.8 <sup>1</sup>
(3) ML Dis+NILs+Clust	95.7	96.8	<b>96.3<sup>1,2</sup></b>	82.5	83.0	<b>82.8<sup>1,2</sup></b>

Table 9.13: Clustering first vs. clustering second vs. joint disambiguation and clustering: clustering results on the English TAC 2011 data set. Significant improvements of a system over another system ( $B^3F$ ,  $B^{3+}F$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

disambiguation and clustering approach, but that also the NIL clustering is better. However, these results should be considered with caution as it is still an open research question how to calculate clustering scores over only some of the mentions.<sup>7</sup>

**Within Document vs. Cross-document Clustering.** With respect to the clustering, we can distinguish between within-document and cross-document clustering. While Chen & Ji (2011) also consider cross-document clustering relations in their clustering first strategy, most other approaches that use clustering information for the disambiguation (e.g. Cheng & Roth (2013)) only apply within document clustering information. Table 9.14 compares the results of our joint concept disambiguation and clustering approach with only within-document clustering to the results with within-document and cross-document clustering on the English TAC 2011 data set. In the within-document clustering approach, the cross-document clustering is performed after disambiguation. Both the  $B^3F$  and  $B^{3+}F$  scores are significantly higher with cross-document information. The accuracy is also significantly higher with cross-document clustering information (+0.9). However, this comparison indicates that the big gains of our joint approach with respect to the accuracy are due to within-document clustering information. These results thus suggest that if time efficiency plays an important role and the main goal is disambiguation, it might be an option to restrict the approach to within-document clustering.

**The Effect of Clustering on Missing Candidate Concepts.** We assume that highly ambiguous mentions such as *Smith* tend to be introduced by a less ambiguous mention in a text. It is thus not necessary that the concept denoted by *Smith* is among the candidate concepts of the mention *Smith*. It is sufficient if it is among the candidate concepts of a less ambiguous

<sup>7</sup>To calculate these results of a subset of mentions, we used the official TAC script, version 0.7. In earlier versions of this script a slightly different method has been used to calculate the clustering score for a subset of mentions. This also reflects that it is still an open research question how to calculate clustering results for a subset of mentions.

	Clustering					
	B <sup>3</sup> P	B <sup>3</sup> R	B <sup>3</sup> F	B <sup>3+</sup> P	B <sup>3+</sup> R	B <sup>3+</sup> F
TAC EN 2011						
(1) ML Dis+NILs+Clust (within-document)	96.1	93.5	94.8	82.5	80.4	81.4
(2) ML Dis+NILs+Clust	95.7	96.8	<b>96.3<sup>1</sup></b>	82.5	83.0	<b>82.8<sup>1</sup></b>

Table 9.14: Within-document vs. cross-document clustering: disambiguation results on the English TAC 2011 data set. Significant improvements of a system over another system ( $B^3 F$ ,  $B^{3+} F$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

mention that is in the same cluster as *Smith*. Via clustering this candidate concept then can be moved to the mention *Smith*.

The substantial differences between the mention- and document-level upper bound on all data sets indicate that the correct concept for a mention is indeed often not part of its candidate concepts, but part of the candidate concepts of other mentions in its document. In the following, we evaluate how often a mention that does not contain the correct concept among its candidates is nevertheless disambiguated to the correct concept. We identified all mentions that denote a concept in the inventory, but do not contain this concept among their candidates in the TAC 2011 data set. In total, this affects 271 mentions out of 2,250. Table 9.15 shows the accuracy for these 271 mentions for our cascaded approach (*ML Dis+NILs*) and our joint approach (*ML Dis+NILs+Clust*).

None of the 271 mentions is correctly disambiguated by *ML Dis+NILs* (Table 9.15). This happens because the correct concept is not among the candidate concepts of these mentions and no clustering information is exploited during disambiguation by *ML Dis+NILs*. However, our joint approach (*ML Dis+NILs+Clust*) disambiguates 46.1% of these mentions (i.e. 125 mentions) correctly. These results are significantly better than the results with no clustering information (*ML Dis+NILs*). These gains are not an artifact of the TAC 2011 data set, but can be observed across data sets. For instance, on the ACE 2005 data set, 7% of the mentions that do not contain their corresponding concepts among their candidates are disambiguated correctly by *ML Dis+NILs+Clust*.

Analogously, we analyzed the gains that can be obtained via joint disambiguation and clustering for mentions that contain the denoted concept among their candidates. We identified 853 Non-NILs in the English TAC 2011 data set for which this is the case. Table 9.16 shows the accuracy scores for these mentions for the two system variants.

Our joint approach significantly improves the accuracy of the system with no clustering information. However, the gains are smaller compared to the ones obtained for mentions that do not contain their corresponding concept among their candidates. Similar trends can also be observed on other data sets including the ACE 2005 data set.

	Acc
TAC 2011 EN	
(1) ML Dis+NILs	0.0
(2) ML Dis+NILs+Clust	<b>46.1<sup>1</sup></b>

Table 9.15: Disambiguation results for mentions that do not contain the denoted concept among their candidate concepts on the English TAC 2011 data set. Significant improvements of a system over another system (*Acc*) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

	Acc
TAC 2011 EN	
(1) ML Dis+NILs	83.9
(2) ML Dis+NILs+Clust	<b>85.3<sup>1</sup></b>

Table 9.16: Disambiguation results for mentions that contain the denoted concept among their candidate concepts on the English TAC 2011 data set. Significant improvements of a system over another system (*Acc*) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

**Improvements.** In this section, we discuss the improvements that are due to joint disambiguation and clustering. To investigate these improvements, we compared the output of our cascaded system (*ML Dis+NILs*) to the output of our joint system (*ML Dis+NILs+Clust*) for the TAC 2011 data set.

Overall, the joint system disambiguated 150 instances (corresponding to more than 6.6% of all instances) correctly that were wrongly disambiguated by the cascaded approach. As discussed above, in many cases (more precisely, in 125), the concept denoted by the mention is not among its candidate concepts. By jointly disambiguating and clustering the mentions, these mentions can be correctly disambiguated, if another mention in the same cluster contains the respective concept among its candidates and shows strong evidence for it. However, clustering can also help if the concept denoted by a mention is among its candidates, as it may provide more evidence from other mentions in the same cluster. To illustrate the gains obtained via joint concept disambiguation and clustering, we show some examples in the following.

One example is the mention *Edwards* in the sentence “Nothing but egomania keeps *Edwards* in the race now.”<sup>8</sup> *Edwards* is highly ambiguous and it is very likely that we miss its corresponding concept during the candidate identification step. However, it is unlikely that *Edwards* is not introduced with a less ambiguous mention unless it has a strong predominant concept (such as e.g. *Merkel*). Indeed, the text from which we extracted this sentence starts

<sup>8</sup>This sentence is from the English testing data set from TAC 2011. The identifier of the article is: eng-NG-31-142692-10073507.

with the sentence “*John Edwards* is a loser.” In addition, *Edwards* is mentioned in many sentences. By jointly disambiguating and clustering concepts, we can exploit all this information. Hence, while *Edwards* is considered as a NIL in the local approach, it is correctly disambiguated in the joint approach.

Another example is the mention *SAFPU* in “As *SAFPU* , we would like to thank ING, academic partners of Johan Cruyff Institute for Sport Studies [...]”<sup>9</sup>. From this sentence, it is difficult to disambiguate *SAFPU* and the cascaded approach wrongly classifies it as a NIL. However, the joint approach clusters it together with the mention *South African Football Players Union* in the sentence “South African Football Players Union in collaboration with Johan Cruyff Academic Institute, University of Johannesburg and Gauteng Department of Sport held a very successful workshop at the University of Johannesburg today.” and correctly disambiguates it as SOUTH AFRICAN FOOTBALL PLAYERS UNION.

While these cases are also correctly resolved by first clustering all mentions and then performing disambiguation, others are more ambiguous and the cascaded approach with the clustering first strategy fails. For instance, the mention *Gnassingbe* in a news text on Togo’s prime minister is much more difficult to disambiguate, as the text mentions both *Faure Gnassingbe* and *Gnassingbe Eyadema*. While with the clustering first strategy all these occurrences are clustered and *Gnassingbe* is wrongly disambiguated as GNASSINGBÉ EYADÉMA, the joint approach manages to disambiguate these mentions correctly, as it considers both disambiguation and clustering information at the same time. In the following example, cross-document evidence leads to the correct disambiguation. In two texts the word *Bata* occurs. In both texts it denotes the concept BATA, EQUATORIAL GUINEA. However, the local approach only disambiguates it correctly in one of the texts which talks about the disappearing of a private plane. In the other text that describes the visit of the Philippine president in Equatorial Guinea, the evidence is too weak and *Bata* is classified as a NIL. However, the joint approach that considers evidence across documents clustered the two mentions and disambiguated both of them correctly. This is also an example of how the clustering is improved by the joint approach.

The joint approach not only improves the disambiguation of proper names as suggested by all previous examples, but also improves the results of common nouns. For instance, one occurrence of *friends* in a text on holiday traditions is wrongly disambiguated as the TV series FRIENDS (SITCOM) by the cascaded approach. However, by also considering the other occurrences of *friends* in the text, the joint approach disambiguated it correctly. Another example is the mention *agency* in the sentence “Asked about the possible war, he said, if tours were curtailed because of a war emergency, the *agency* would refund the portion of the costs covering the remainder of the trip.”<sup>10</sup> The cascaded approach wrongly linked *agency* to the concept

<sup>9</sup>This sentence is from the English testing data set from TAC 2011. The identifier of the article is: eng-NG-31-134077-12158396.

<sup>10</sup>The sentence is from the ACE 2005 data set. The text has the identifier XIN\_ENG\_20030314.0208.

GOVERNMENT AGENCY. However, the joint approach clustered *agency* with *travel agency* in “According to a staff member with one travel agency still doing Middle East business [...]” and correctly linked it to TRAVEL AGENCY.

In a few cases, the joint approach performs worse than the cascaded approach. On the English TAC 2011 test data, the joint approach failed to disambiguate 23 mentions that were correctly disambiguated by the cascaded approach. While the cascaded approach for instance correctly classified *Steward* in a text on Mormons as a NIL, the joint approach clustered *Steward* and *David Stewart* and wrongly linked both of them to the concept DAVID STEWART (SCOTTISH POLITICIAN). Hence, although some errors are introduced by jointly disambiguating and clustering mentions, the overall improvements are substantial.

### 9.4.2 Scope-Aware Approach

Besides a joint concept disambiguation and clustering approach, we also propose a scope-aware approach in this thesis that applies different models to different mentions. We already showed the effectiveness of the scope-aware approach by comparing its results to the results of our scope-unaware approach (Section 9.3). In this section, we further analyze these results and show some examples. We mainly focus on the ACE 2005 data set in this analysis, as in this data set many mentions are annotated for each document and not only a few challenging mentions as in the TAC data sets. This allows us to obtain a better picture how the scope-aware approach performs across mentions in a text.

**Scope-unaware vs. Cascaded Scope-aware vs. Joint Scope-aware Approach.** While we already showed that our scope-aware approach significantly outperforms the corresponding scope-unaware approach on different data sets, we did not discuss how our scope-aware approach performs compared to a cascaded scope-aware approach. In the proposed scope-aware approach, the ambiguity resolution and the scope assignment are performed jointly. We argue that the scope of a mention depends on its concept and vice versa and thus modeled it accordingly. Another way to introduce scope awareness is to first assign each mention a scope and to perform the disambiguation and clustering in a second step by considering the scope assigned to each mention in the first step. In the following, we compare the results of our joint approach to the results of such a cascaded approach.

We implemented a cascaded scope-aware approach in Markov logic. In contrast to the joint approach, the weights for the scope assignment task are learned separately from the weights for the disambiguation and clustering and the predicate *hasScope* is observed (not hidden) during the disambiguation and clustering. All other aspects such as the features are exactly the same in the two approaches.

Table 9.17 shows the results for the scope-unaware approach (*ML Dis+NILs*), the cascaded

	In Inventory			NILs			Acc
	$P_{inv}$	$R_{inv}$	$F_{inv}$	$P_{NILs}$	$R_{NILs}$	$F_{NILs}$	
(1) ML Dis+NILs	77.3	76.0	76.6	47.6	46.6	<b>47.1</b> <sup>2</sup>	73.9
(2) ML Dis+NILs Scope (Casc)	80.1	75.8	77.9 <sup>1</sup>	38.1	55.9	45.3	74.3 <sup>1</sup>
(3) ML Dis+NILs Scope	80.1	76.6	<b>78.3</b> <sup>1,2</sup>	40.4	54.0	46.2	<b>74.9</b> <sup>1,2</sup>

Table 9.17: Results on the ACE 2005 data set: scope-unaware vs. cascaded scope-aware vs. joint scope-aware approach. Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

scope-aware approach (*ML Dis+NILs Scope (Casc)*) and the joint scope-aware approach (*ML Dis+NILs Scope*) on the ACE 2005 data set. As the results indicate, both scope-aware approaches significantly outperform the scope-unaware approach with respect to the  $F_{inv}$  score and the accuracy ( $Acc$ ). Hence, using different models for different mentions thus leads to higher results independently from whether we use a cascaded or a joint approach. However, modeling the scope assignment and the disambiguation jointly leads to higher results than modeling them in a cascaded way. This is in line with our claim that the scope is influenced by the concept denoted by a mention and vice versa.

**Local vs. Intermediate vs. Global Scope.** For the scope assignment task, no gold annotations are available. However, we can investigate the induced scopes and discuss the performance of our disambiguation approach for each scope.

Figure 9.3 shows the distribution of the mentions across induced scopes in the ACE 2005 data set. Mentions with local scope are more frequent than mentions with intermediate scope followed by mentions with global scope. This distribution is in line with previous findings. For instance, in the system of Hoffart et al. (2011b), approximately two thirds of all mentions are disambiguated locally. Although they do not use discourse information to decide which information should be used to disambiguate a mention, their conclusions that most mentions can be disambiguated locally is similar to ours. For each scope, we show the portion of common nouns and proper names. While for the local scope, common nouns are more frequent, for the global scope proper names are more frequent. For the intermediate scope, the portion of common nouns and proper names is more balanced. This is in line with our assumption that common nouns are more likely to be of local scope than proper names. However, as we use the fact whether a mention is a proper name also as a feature for the scope assignment, this assumption is also incorporated into the approach. We thus refrain from drawing conclusions out of the distribution of common nouns and proper names across scopes.

Table 9.18 compares the accuracy of the scope-unaware approach (*ML Dis+NILs+Clust*)

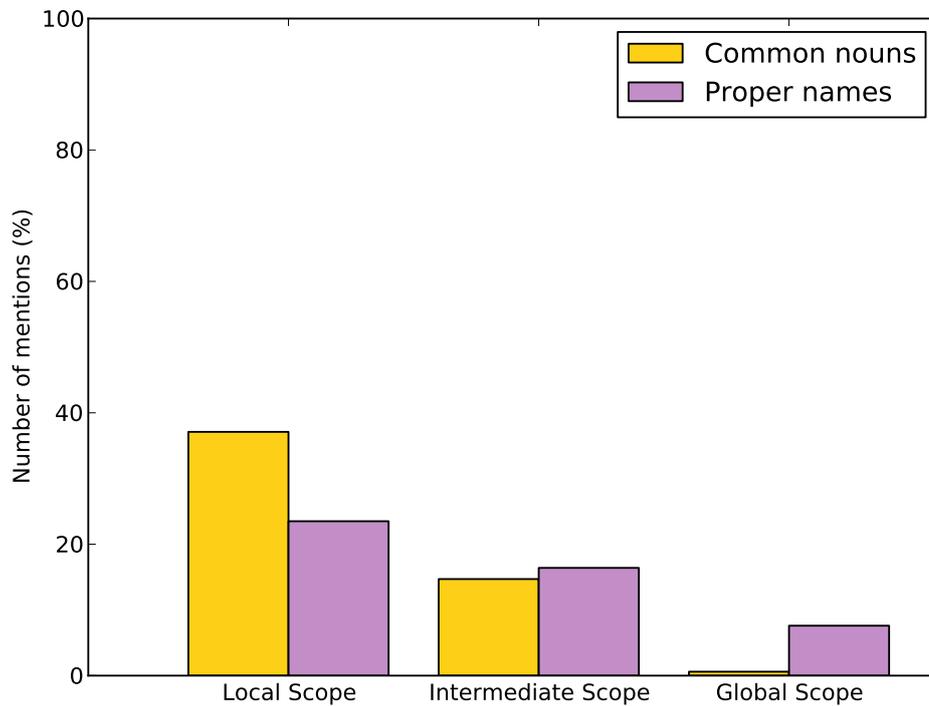


Figure 9.3: Distribution across induced scopes on the ACE 2005 data set.

with the accuracy of the scope-aware approach (*ML Dis+NILs+Clust Scope*). For each scope, the accuracy is reported separately. As mentions that were not recognized by our mention recognition component are not assigned any scope, we only consider mentions that have been actually recognized. The accuracy scores are thus slightly higher than in Table 9.3. The scope-aware approach improves the disambiguation results for mentions of all three scopes. The biggest gain (2.5) is achieved for mentions with induced global scope. The gain for mentions with local and intermediate scope is 1.2 and 0.5 respectively. Except for the intermediate scope, all gains are significant.

While the gains on the ACE 2005 data for the scope-aware approach are substantial, scope awareness does not lead to consistent improvements on the TAC data set. In contrast to the ACE data set, the TAC data sets only contain annotations for a few challenging mentions. Inspecting the scope distribution on the English TAC 2011 data set reveals that most of the mentions (71.1%) were assigned intermediate scope. The gains for mentions with intermediate scope are also not significant on the ACE 2005 data set. Hence, this indicates that the small gains on the TAC data set can be traced back to the selective annotations. However, to make sure that the gains we obtain on the ACE 2005 for scope awareness are not an artifact on this data set, we additionally ran our scope-unaware model and the scope-aware model on the CoNLL 2003 data set. This data set is annotated for proper names. However, in contrast to

	Local Scope Acc	Intermediate Scope Acc	Global Scope Acc
(1) ML Dis+NILs+Clust	75.6	76.3	73.7
(2) ML Dis+NILs+Clust Scope	76.8 <sup>1</sup>	76.6	76.2 <sup>1</sup>

Table 9.18: Evaluation on ACE 2005 data across induced scopes. In contrast to Table 9.3, the accuracies are slightly higher, as we only consider here mentions that have been recognized by our mention identification component. Significant improvements of a system over another system ( $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

the TAC data, all proper names in the text are annotated.

Table 9.19 shows the results of our scope-unaware (*ML Dis+NILs+Clust*) and our scope-aware approach (*ML Dis+NILs+Clust Scope*) on the CoNLL 2003 training, development and testing data. On all three parts the improvements obtained by introducing scopes are significant. Hence, as these results show, the gains on the ACE 2005 data set are not an artifact of this specific data set, but can also be observed for instance on the CoNLL data set. Compared to Hoffart et al. (2011b), who report a precision for the Non-NILs of 81.8 at a recall level of 100.0 or Moro et al. (2014b) who report an accuracy of 82.1 for the Non-NILs on the testing part, our results seem to be low. However, these results are not comparable, as we use system mentions and our task is more difficult, as we also recognize the NILs. The other systems only focus on the mentions with a corresponding concept in the inventory and assign each of them a concept.

**Scope and Prominence.** To further evaluate the quality of the scope assignment, we investigate the scopes in the context of keyword selection. The assumption is that mentions of intermediate and global scope are more likely to be keywords than mentions with local scope, as they show cohesive ties beyond the sentence. As keyword selection is part of the cross-lingual link discovery task at NTCIR-9, we use this data set for the following experiment.

Table 9.20 compares the results of three different systems. The first system selects keywords from all mentions (*All*), the second system selects keywords only from mentions with intermediate or global scope (*Non-Local*) and the third system selects keywords only from mentions with local scope (*Local*). In all system, the mentions within the mention pool to select from are ranked by keyphraseness. As the results indicate, the  $P@5$  scores are much higher if the mention pool is restricted to mentions with global and intermediate scope than if the mention pool is restricted to mentions with local scope.<sup>11</sup> These results indicate that mentions with global or intermediate scope are much more likely to be keywords than mentions with local scope and is in line with our assumption. The differences between *All* and *Non-*

<sup>11</sup>For brevity, we only report here  $P@5$ . The results for  $MAP$  are similar.

	In Inventory			NILs			Acc
	$P_{inv}$	$R_{inv}$	$F_{inv}$	$P_{NILs}$	$R_{NILs}$	$F_{NILs}$	
<b>Test</b>							
(1) ML Dis+NILs+Clust	74.2	74.9	74.5	79.1	61.4	69.1	72.2
(2) ML Dis+NILs+Clust Scope	77.5	76.4	77.0 <sup>1</sup>	74.8	65.2	69.7	74.2 <sup>1</sup>
<b>Dev</b>							
(1) ML Dis+NILs+Clust	75.4	76.0	75.7	77.3	53.3	63.1	71.7
(2) ML Dis+NILs+Clust Scope	77.6	75.8	76.7 <sup>1</sup>	69.8	57.5	63.0	72.3 <sup>1</sup>
<b>Train</b>							
(1) ML Dis+NILs+Clust	76.5	79.2	77.9	83.3	55.2	66.4	74.3
(2) ML Dis+NILs+Clust Scope	79.4	79.1	79.2 <sup>1</sup>	74.7	60.9	67.1 <sup>1</sup>	75.3 <sup>1</sup>

Table 9.19: Results of our scope-unaware (*ML Dis+NILs+Clust*) and our scope-aware (*ML Dis+NILs+Clust Scope*) on the CoNLL 2003 data set. Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

*Local* are rather small or zero for Chinese and Japanese. These results indicate that mentions with local scope tend to have low keyphraseness scores and are therefore not part of the top 5 concepts for a document if all mentions are considered. This can be at least partially traced back to the fact that we also use frequency information as a feature for the scope assignment task (*tf idf* scores from the Gigaword corpus). However, for Korean, an improvement of 4.0 points can be observed by restricting the mention pool to mentions with intermediate and global scope. Hence, although mentions with local scope tend to have lower keyphraseness scores than mentions with intermediate and global scope, some mentions with local scope also have a high keyphraseness score. Removing them from the mention pool thus can improve the results for the keyword selection.

Overall, scope information can improve frequency-based keyword selection. However, the keyphrase baseline is strong and only outperformed for Korean.

**Improvements.** To obtain some insights on the behavior of the scope-aware approach, we investigate some examples. In a text on the 2004 US elections, the mention *Kerry* in “*Kerry* was the clear winner, but victory was snatched from him”<sup>12</sup> is wrongly disambiguated to KERRY GAA, a branch of the Gaelic football association, by the scope-ignorant approach, because the local context strongly prefers an interpretation in the domain of sports. In the scope-aware approach, *Kerry* is assigned global scope, and it is correctly disambiguated to JOHN KERRY, an American politician, as the global relatedness overrules the local context in this model. In another text on U.S. troops in Iraq, the scope-ignorant approach disambiguates

<sup>12</sup>This example is from the ACE 2005 data set. The identifier of the text is: rec.music.makers.guitar.acoustic\_20041228.1628.

	EN to JA	EN to KO	EN to ZH
	P@5	P@5	P@5
All	<b>84.8</b>	80.8	<b>83.2</b>
Non-Local	<b>84.8</b>	<b>84.8</b>	82.4
Local	69.6	67.2	69.6

Table 9.20: Results of scope-informed keyword selection on the NTCIR-9 data set. We report the P@5 for all three languages for the following three systems: *All*: the keywords are selected from all mentions by sorting them by keyphraseness. *Non-Local*: the keywords are selected from mentions that were assigned intermediate or global scope by sorting them by keyphraseness. *Local*: the keywords are selected from mentions that were assigned local scope by sorting them by keyphraseness.

*south* in “Monday’s advances came one day after British forces in the *south* made their deepest push into Iraq’s second largest city”<sup>13</sup> to SOUTHERN UNITED STATES as concepts related to the USA are quite prominent in the text. In the scope-aware approach *south* is considered as being of local scope and is correctly disambiguated as SOUTH. In “we happen to be at a very nice *spot* by the beach where this is a chance for people to get away from cnn coverage”<sup>14</sup> *spot* is disambiguated as SPOT (SATELLITE) in the scope-ignorant approach (misled by CNN), while it has been correctly recognized as NIL by the scope-aware approach in which it is considered as being of intermediate scope.

### 9.4.3 Multi- and Cross-linguality

In Section 9.3.2, we have shown that our system scales well across languages and obtains state-of-the-art results for Spanish and Chinese with minimal adaptations. In this section, we analyze these results in more detail.

**Monolingual vs. Cross-lingual Concept-level Co-occurrence Information.** Concept-level co-occurrence information is crucial for concept disambiguation, as it is a way to approximate concept-level cohesive ties (Section 5.1.2). As the concepts are shared across languages, we can also share concept-level co-occurrence information across languages. For instance, we can extract concept-level co-occurrence information from the English Wikipedia and use it for Spanish. In this section, we compare different strategies.

Table 9.21 compares the use of mono- and cross-lingual concept-level co-occurrence information on the Spanish and Chinese cross-lingual testing data from TAC 2012: the first systems (*ML Dis+NILs+Clust (ES Cooc)* and *ML Dis+NILs+Clust (ZH Cooc)*) use concept-level

<sup>13</sup>This example is from the ACE 2005 data set. The identifier of the text is: APW\_ENG\_20030407.0030.

<sup>14</sup>This example is from the ACE 2005 data set. The identifier of the text is: CNN\_ENG\_20030327\_163556.20. The token *cnn* also written in lowercase in the original text.

co-occurrence information extracted from the Spanish and Chinese Wikipedia dumps respectively; the second systems (*ML Dis+NILs+Clust (EN Cooc)*) use concept-level co-occurrence information extracted from the English Wikipedia dump. In both the Spanish and the Chinese case, the co-occurrences extracted from the English Wikipedia lead to the highest results. Using only the co-occurrences from the respective language leads to a drop in the performance. While for Chinese the difference is small, it is larger for Spanish.

As Table 3.6 in Section 3.3 shows, we can extract much more inlinks – our source for co-occurrence information – from the English Wikipedia than from the other language versions. This at least partially explains why the English co-occurrence information leads to higher results than the Spanish and Chinese co-occurrence information. Additionally, the English co-occurrence information is only available for concepts that have a corresponding concept in the English Wikipedia. Using the English co-occurrence information thus decreases the chances that a mention is linked to a concept which is for instance only available in the Spanish Wikipedia, but increases the chances that it is linked to a concept that is also present in the English Wikipedia. In a cross-lingual setting, this bias towards shared concepts might lead to slightly higher results. That the gains are higher for Spanish than for Chinese by using English co-occurrence information might be explained with the higher coverage regarding the concept mapping for Spanish than for Chinese. The coverage of the cross-lingual concept mapping is 10% higher (absolute) for Spanish than for Chinese (Section 3.3). Only for the concepts that are mapped to an English concept, cross-lingual co-occurrence information can be exploited. Hence, the mapping is crucial for leveraging cross-lingual information. We also experimented with combining the co-occurrence information extracted from different languages. Combining co-occurrence information extracted from English and Spanish decreased the Spanish results substantially, while combining English and Chinese co-occurrence information did not significantly change the Chinese results. However, these results are preliminary and need further inspections including the integration of co-occurrence information from other languages.

Overall, the results in Table 9.21 show that cross-lingual concept-level co-occurrence information leads to good results. Hence, it is not necessary to extract concept-level co-occurrence information from the respective language, but it can be shared across languages. However, accessing concept-level co-occurrence information across languages requires a cross-lingual mapping.

**Language-independent vs. Adaptation-based System.** Besides the concept-level co-occurrence information, our system exploits other information sources. With respect to the disambiguation, our system uses two main other information sources: string-level co-occurrence features (Table 5.1, 5-9) and prior probabilities (Table 5.1, 1). In both cases, language-specific information is necessary, which we obtained from Wikipedia in the respective language version. In the following, we investigate how our system performs without this addi-

	In Inventory			NILs			Acc
	$P_{inv}$	$R_{inv}$	$F_{inv}$	$P_{NILs}$	$R_{NILs}$	$F_{NILs}$	
TAC 2012 ES							
(1) ML Dis+NILs+Clust (ES Cooc)	61.5	48.1	54.0	72.2	85.0	78.1	68.5
(2) ML Dis+NILs+Clust (EN Cooc)	58.5	58.1	<b>58.3<sup>1</sup></b>	79.5	80.0	<b>79.8<sup>1</sup></b>	<b>70.2<sup>1</sup></b>
TAC 2012 ZH							
(1) ML Dis+NILs+Clust (ZH Cooc)	90.1	78.0	83.6	77.9	92.6	84.6	84.1
(2) ML Dis+NILs+Clust (EN Cooc)	90.7	79.0	<b>84.4<sup>1</sup></b>	78.7	92.9	<b>85.2<sup>1</sup></b>	<b>84.8<sup>1</sup></b>

Table 9.21: Monolingual (*ES Cooc*, *ZH Cooc*) vs. cross-lingual (*EN Cooc*) concept-level co-occurrences. Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

tional language-specific knowledge. The clustering features may also require some language-specific adaptations, but more in terms of linguistic hard-coded knowledge and less in terms of resources (e.g. vast amount of annotated data). We thus focus her on the disambiguation.

Table 9.22 shows how our system performs if we remove language-specific resources. The first system (*ML Dis+NILs+Clust*) is our joint concept disambiguation and clustering system that requires both language-specific string-level co-occurrence information and prior probabilities. The second system (*(1)-lexical features*) is the same system, but without string-level co-occurrence information. The third system (*(2)-prior probability*) further drops the prior probabilities and only uses concept-level co-occurrence information extracted from the English Wikipedia. It thus requires no language-specific resources except a language-specific lexicon and a mapping to the concepts extracted from the English Wikipedia. We show the results of these three systems for English (testing data of TAC 2011), Spanish (testing data of TAC 2012) and Chinese (testing data of TAC 2012).<sup>15</sup>

While on the English monolingual data set, the performance without string-level co-occurrence information significantly drops, this is not the case for the Spanish and Chinese cross-lingual data sets. In both cross-lingual cases, the performance without this information is not significantly worse. This does not indicate that this information is not useful in these languages, but rather indicates that the string-level information we extracted for these languages might be insufficient. Discarding also the prior probability information the performance significantly drops in all languages. As already the good results for the first concept baseline indicated, the prior probability statistics are reliable for all languages.

Overall, these results indicate that even with no language-specific information (such as string-level co-occurrence information and prior probabilities) that requires vast amount of annotated data, the accuracy of our system is still above 65 for all languages. Depending on

<sup>15</sup>As we did the whole analysis on the TAC 2011 data for English, we use the same data here as well. For Spanish and Chinese we use the 2012 data sets as in the previous section.

	In Inventory			NILs			Acc
	$P_{inv}$	$R_{inv}$	$F_{inv}$	$P_{NILs}$	$R_{NILs}$	$F_{NILs}$	
TAC 2011 EN							
(1) ML Dis+NILs+Clust	85.9	76.2	<b>80.8</b> <sup>2,3</sup>	83.9	93.3	<b>88.4</b> <sup>2,3</sup>	<b>84.8</b> <sup>2,3</sup>
(2) (1)-lexical features	84.1	72.2	77.7 <sup>3</sup>	81.7	93.2	87.1 <sup>3</sup>	82.7 <sup>3</sup>
(3) (2)-prior probability	89.8	58.2	70.6	71.2	96.3	81.9	77.2
TAC 2012 ES							
(1) ML Dis+NILs+Clust (EN Cooc)	58.5	58.1	<b>58.3</b> <sup>3</sup>	79.5	80.0	<b>79.8</b> <sup>3</sup>	<b>70.2</b> <sup>3</sup>
(2) (1)-lexical features	57.5	56.7	57.1 <sup>3</sup>	78.8	79.7	79.3 <sup>3</sup>	69.4 <sup>3</sup>
(3) (2)-prior probability	68.2	36.1	47.2	66.4	91.7	77.0	66.8
TAC 2012 ZH							
(1) ML Dis+NILs+Clust (EN Cooc)	90.7	79.0	<b>84.4</b> <sup>3</sup>	78.7	92.9	85.2 <sup>3</sup>	<b>84.8</b> <sup>3</sup>
(2) (1)-lexical features	89.8	79.4	84.3 <sup>3</sup>	79.3	92.2	<b>85.3</b> <sup>3</sup>	84.7 <sup>3</sup>
(3) (2)-prior probability	95.0	65.7	77.7	67.4	96.6	79.4	78.6

Table 9.22: Dependencies on language-specific features: joint disambiguation and clustering approach without lexical features and without prior probability. Significant improvements of a system over another system ( $F_{inv}$ ,  $F_{NILs}$ ,  $Acc$ ) with  $p < 0.05$  are indicated by the corresponding system identifiers as superscripts.

the application, such a system might still be useful. However, we focused here on proper names. It requires further investigations whether the behavior is similar for common nouns.

#### 9.4.4 Remaining Errors

In the previous section, we mainly focused on different system variants and compared these results to the results of other systems. In this section, we inspect the difference between the upper bound and the performance of our scope-aware approach (*ML Dis+NILs+Clust Scope*) and discuss some typical errors made by our system. This analysis allows us to identify starting points for future improvements.

Figure 9.4 shows the distribution of different error types for the ACE 2005 and English TAC 2011 data set. On the ACE 2005 data set, the scope-aware approach fails to correctly disambiguate 7,198 mentions out of 29,300. 60.1% of these errors result from an assignment of a wrong concept to a mention (*Non-NIL (wrong concept)*). 21.8% of the errors originate from a misclassification of a mention as a NIL, although it is a Non-NIL (*Non-NIL (wrong NIL)*). 13.5% of the errors lead back to mentions that are wrongly recognized as Non-NILs (*NIL (wrong Non-NIL)*). The remaining errors (< 5%) are due to the mentions recognition (*Mention Recognition*). On the English TAC 2011 data set, 327 out of 2,250 mentions are wrongly disambiguated. 13.5% of these errors result from an assignment of a wrong concept to a mention (*Non-NIL (wrong concept)*). 68.8% of the errors originate from a misclassification of a mention as a NIL, although it is a Non-NIL (*Non-NIL (wrong NIL)*). 17.7% of the errors lead

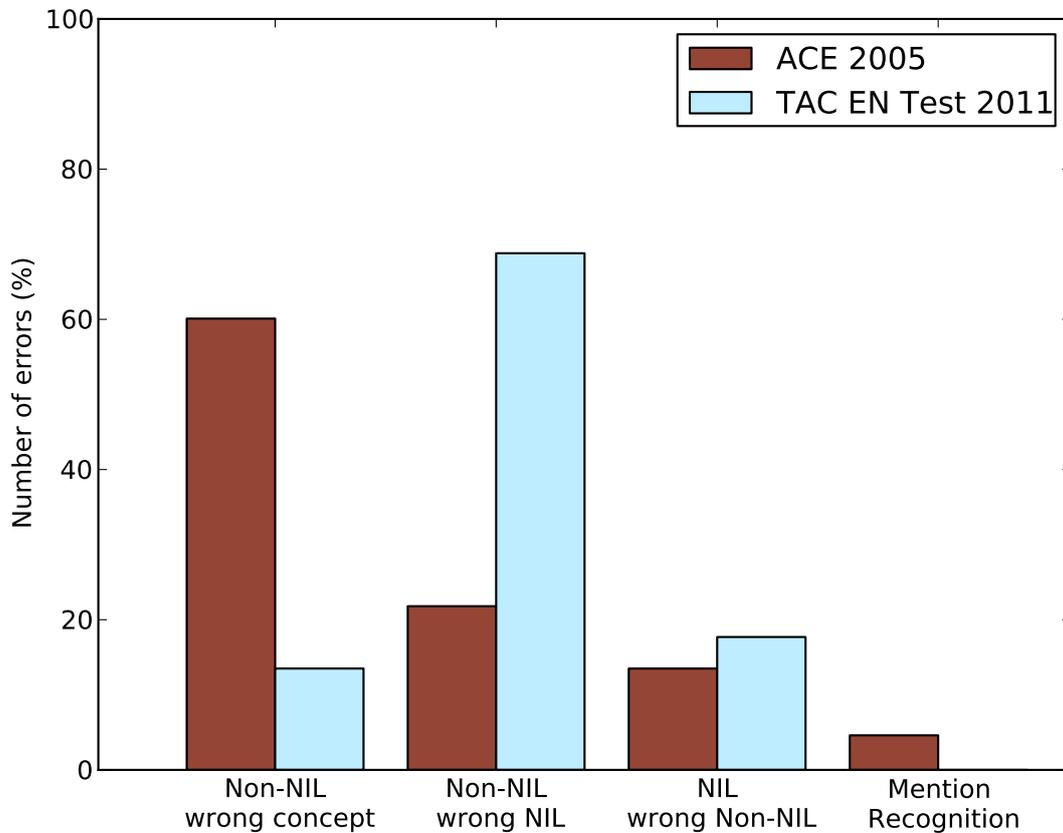


Figure 9.4: Distribution of errors across different types on the ACE 2005 and TAC 2011 data set. On the ACE data set 7,198 out of 29,300 mentions are wrongly disambiguated. On the English TAC 2011 data set 327 out of 2,250 mentions are wrongly disambiguated.

back to mentions that are wrongly recognized as Non-NILs (*NIL (wrong Non-NIL)*). Hence, on both data sets the two main sources of errors are Non-NILs that are assigned a wrong candidate concept or that are wrongly recognized as NILs. We thus focus in the following on these two error categories.

**Disambiguation of Non-NILs.** Screening through the errors for the Non-NILs made by our system, we identified three main categories.

The first category consists of errors where the result of our system is completely wrong. For instance, in a text on SARS, the mention *director* was wrongly disambiguated as FILM DIRECTOR. This is a clear error and might be prohibited with more features.

The second category of errors are errors in the granularity and more difficult to solve. For instance, in a news text on executions in Indonesia the mention *Supreme Court* is wrongly

disambiguated as SUPREME COURT OF THE UNITED STATES, while the concept in the gold standard is SUPREME COURT. Hence, the concept selected by our system is wrong, but not completely wrong. The mention denotes a SUPREME COURT, but not the one in the US, but the one in Indonesia. Such errors are more difficult to address and might require a hierarchical approach where mentions are iteratively assigned to more and more specific concepts.

Finally, the third category of errors occurs between highly related concepts. For instance, in the text on SARS the mention *resident* is disambiguated as PERMANENT RESIDENCY by our system instead of RESIDENCY (DOMICILE) as in the gold standard. Such cases are difficult for our system and are also difficult to be eliminated.

Hence, in the future, we would first address the first category of errors, i.e. errors that are obvious and that can be omitted by adding more features. In a next step, granularity errors might be addressed via a hierarchical approach, while the last category of errors might require more knowledge.

**Recognition of NILs.** Our approach could considerably be improved by investing in the recognition of NILs. As the following example illustrates, it can be a difficult task to identify such NILs: “Some Florida legislators want to give people the right to shoot an attacker in a *public place*. [...] to allow people to shoot to kill in self-defense if they are attacked `in any other *place* where he or she has a right to be.’”<sup>16</sup> We are interested in the two (highlighted) mentions *public place* and *place*. Our approach correctly disambiguates *public place* as PUBLIC PLACE. The mention *place* is also linked to the concept PUBLIC PLACE, although it is a NIL according to the gold standard, as it denotes a more general concept that is not part of Wikipedia. Recognizing such subtle differences is already difficult if this more general concept were part of our inventory. It is even more difficult if it is not part of our inventory, as our approach lacks any positive evidence for NILs. More precisely, our approach recognizes NILs by deciding that none of the candidate concepts is a valid candidate. It does not use positive evidence that a mention should be a NIL. Hence, in this and many more examples, the evidence against one of the candidate concepts (here PUBLIC PLACE) is too weak to prevent linking it to this concept.

To address such cases, it would be interesting to identify sources of positive evidence for the recognition of NILs. Some first work in this direction has been done by Han et al. (2011) and Hoffart et al. (2014). More research is required to provide a robust solution for NILs.

Although our approach is robust enough to disambiguate and cluster mentions in noisy texts, its performance is still considerably better on news texts. Table 9.23 show the accuracy of our scope-aware approach across different text sources on the English testing data of TAC

<sup>16</sup>From the ACE 2005 data set. The identifier of the text is: AGGRESSIVEVOICEDAILY\_20050224.1207.

	<b>News Acc</b>	<b>Web Texts Acc</b>	<b>Discussion Forums Acc</b>
ML Dis+NILs+Clust Scope	88.3	77.0	73.9

Table 9.23: Accuracy of our approach across different text sources on the TAC 2013 testing data.

2013. We chose this data set, as it contains not only news and web texts, but also texts from discussion forums. As the results illustrate, our approach performs more than 10 points better on news texts than on the web texts. The performance further drops if discussion forums are considered which are even noisier than other web texts. As these results suggest, it might be promising in the future to take into account the source of the text and use different models for different sources.

## 9.5 Summary

In this chapter, we evaluated our approach on different data sets and compared the results to the results of related approaches. We showed that our approach leads to state-of-the-art results. In addition, we evaluated the impact of jointly disambiguating and clustering concepts (Contribution 1, 3) and the impact of the scope awareness (Contribution 2, 3). In both cases, we could observe a significant improvement of the results. Furthermore, we evaluated our approach on other languages than English and analyzed the impact of language-specific adaptations (Contribution 4). We showed that our approach scales well across languages and also achieves state-of-the-art results on languages other than English.

# Chapter 10

## Related Work

In 2013 and 2014, the most important natural language processing and information retrieval conferences – *SIGIR 2013*, *WWW 2013*, *WSDM 2014* and *ACL 2014* – hosted tutorials on concept disambiguation under the name *Entity Linking and Retrieval* (Meij et al., 2013a; 2013b; 2014) and *Wikification and Beyond: The Challenges of Entity and Concept Grounding* (Roth et al., 2014). These tutorials show that concept disambiguation and clustering is considered as relevant in both the natural language processing and the information retrieval communities. While resolving lexical ambiguity has been a topic in natural language processing since its beginning, the use of Wikipedia and related resources as an inventory has given the field a new spin. Since 2007, shared tasks with different foci have been organized. The most prominent one is the *Entity Linking Task* at *TAC* that takes place every year since 2009 (McNamee & Dang, 2009; Ji et al., 2010; 2011). But also the *Cross-lingual Link Discovery Task* at *NTCIR 9* (Tang et al., 2011) and *NTCIR 10* (Tang et al., 2013), the *INEX Link-the-Wiki track* from 2007 to 2009 (Huang et al., 2008; 2010), the *Multilingual Word Sense Disambiguation Task* at *SemEval 2013* (Navigli et al., 2013), and more recently, the *Entity Recognition and Disambiguation Challenge* at *SIGIR 2014*<sup>1</sup> and the *Making Sense of Microposts* at *WWW 2014* (Cano et al., 2014) have attracted many participants. Considering the funding agencies behind these shared tasks reveals that both the governmental (e.g. *DARPA*, the *US Defense Advanced Research Projects Agency*) and the industrial sector (e.g. *Google* and *Microsoft*) are interested in these topics.

In this chapter, we embed our proposed approach in the broader research context. We mainly focus on approaches that use Wikipedia as an inventory, but also sketch the relation to previous work that uses other resources such as WordNet. By surveying recent work, we also close a gap: while in-depth research surveys exist from the perspective of the late nineties (Ide & Véronis, 1998) and the late noughties (Agirre & Edmonds, 2006; Navigli, 2009), no extensive survey reviews the more recent work that uses Wikipedia as an inventory. Ji &

---

<sup>1</sup><http://web-ngram.research.microsoft.com/erd2014/>, 14.7.2014.

Grishman (2011) and partially Hachey et al. (2013) summarize such approaches, but they both focus on work in the context of the TAC shared tasks. Some other studies (Rizzo et al., 2012; Cornolti et al., 2013) compare the performance of different systems that use Wikipedia as an inventory, while Moro et al. (2014b) draw some connections between previous work with WordNet and more recent work with Wikipedia. However, none of these studies is extensive. Cornolti et al. (2013) discuss how different task definitions entail each other, but they do not reveal the underlying assumptions. They also propose a framework to improve system comparison. In the same vein, some other publicly available frameworks have been proposed (Miller et al., 2013; Ceccarelli et al., 2013) and a repository for system outputs has been built (Hachey et al., 2014).

In the following, we analyze related approaches by comparing them based on selected aspects that are relevant to this thesis. We first present prevalent task definitions (Section 10.1) that not only illustrate different facets of the problem, but also show the common and diverse interest of different sub-communities currently working on concept disambiguation and clustering. We then contrast different methods (Section 10.2). As one of the main contributions of this thesis is the proposed context model, we show how it is related to other context definitions in the literature (Section 10.3). While these aspects address how the problem is formalized, commonly used features are discussed in Section 10.4. Multi- and cross-lingual approaches are then picked up in Section 10.5. The closest related approaches are summarized in Table 9.2 in Section 9.1.4.

## 10.1 Task Definitions

Ambiguities of nouns have been studied by different communities with different applications in mind, consequently leading to different task definitions. Although these task definitions vary in certain aspects and assumptions, they are borne by the shared idea of reducing ambiguities.

In this section, we compare different task definitions to embed our proposed approach in the broader research context. Shared tasks play an important role in this context, as they have heavily shaped the common task definitions.

### 10.1.1 Lexical Semantics

Resolving lexical ambiguities has been one of the core tasks in natural language processing since its beginning. Weaver (1955) reasons how lexical ambiguity could be solved in the context of machine translation and optimistically concludes that by considering a sufficiently large context, a machine would be able to resolve it (Section 1). This optimism is not shared

anymore by Bar-Hillel (1960) who judges it as a difficult problem involving a lot of knowledge. While in these first considerations in the context of machine translation the problem and its determining factors are identified (e.g. Kaplan (1955)), no clear task is yet defined. This changes in the eighties and nineties with the availability of machine readable dictionaries and the statistical paradigm shift in natural language processing. During these years two main research lines emerge: *word sense disambiguation* and *word sense discrimination*. In more recent years, an additional task known as *unknown sense detection* has evolved. We briefly introduce these tasks and finally discuss their relation to our approach.

**Word Sense Disambiguation.** In word sense disambiguation, resolving lexical ambiguities is formulated as a labeling task. Given a predefined inventory containing senses, the task is to link each occurrence of an open-class word – i.e. nouns, verbs, adjectives and adverbs – to the sense in the inventory that best fits its context (Ide & Véronis, 1998; Navigli, 2009). In contrast to other labeling tasks such as part of speech tagging where the set of labels is fixed, the set of labels depends on the token to disambiguate (Navigli, 2009). Hence, Ide & Véronis (1998) define the task as a two step process: (1) determination of relevant candidate senses for each word; (2) selection of the appropriate candidate sense for each token. By assuming a complete predefined inventory that contains a mapping from senses or labels to lexical realizations, the first step is usually reduced to a lexicon look-up step and not further considered. The focus lies on the second step, the effective disambiguation. Lesk (1986) is one of the first that defines the task in this way. For each word to disambiguate the sense from an inventory – e.g. from the *Oxford Advanced Learners Dictionary of Current English* – is selected that best fits the context of the word. Meanwhile, several shared tasks at SenseEval and its successor SemEval (Kilgarriff & Rosenzweig, 2000; Edmonds & Cotton, 2001; Mihalcea & Edmonds, 2004) have shaped the definition of word sense disambiguation.

Essentially, two variations of the word sense disambiguation task are distinguished (Palmer et al., 2006). In the *lexical sample task*, the aim is to disambiguate occurrences of a predefined set of open-class words in different texts (Kilgarriff & Rosenzweig, 2000; Kilgarriff, 2001; Palmer et al., 2001; Mihalcea et al., 2004a; Pradhan et al., 2007, inter alia). For instance, in the English lexical sample task at Senseval-2, between 75 and 300 occurrences of 73 noun, adjective and verb types are tagged (Kilgarriff, 2001; Palmer et al., 2001). These occurrences are split into a training and a test set. In the *all words task*, the aim is to disambiguate all open-class words in a text (e.g. Palmer et al. (2001), Snyder & Palmer (2004), Pradhan et al. (2007)). For instance, in the English all words task at Senseval-3, more than 2,000 open-class words have been annotated in three texts to evaluate the participating systems (Snyder & Palmer, 2004). Hence, in the all words setting a meaning representation is added to as many open-class words as possible in a text, which might be useful for downstream tasks such as summarization or machine translation. Evaluating a word sense disambiguation system in

this setting allows to estimate the average performance of the system on many different types ranging from highly to hardly ambiguous ones. In contrast, the lexical sample task enables to evaluate a word sense disambiguation system on a few selected types that can be highly ambiguous and estimate its performance on these challenging cases.

Most work in English word sense disambiguation uses WordNet as an inventory, hence word sense disambiguation is often equated with disambiguation with WordNet. With respect to noun phrases, the focus lies on common nouns. Proper nouns are usually not part of the selected types in the lexical sample task and not annotated with a sense in the all-word setting (e.g. Snyder & Palmer (2004), Miller et al. (1993)). This can be traced back to the widespread assumption that proper names lack a denotation and only refer (Section 2.1.2), but also to the commonly used resource WordNet in English, which only contains a few proper names (Section 3.1.2). However, current trends point into a new direction. For instance, the latest edition of the multilingual all-word disambiguation task at SemEval 2013 uses *BabelNet* (Navigli & Ponzetto, 2012a) – a multilingual resource consolidating Wikipedia and WordNet – as inventory (Navigli et al., 2013) and includes the disambiguation of both common nouns and proper names.

**Unknown Sense Detection.** Already early work in the field of automatic text understanding broaches the issue that an inventory might be incomplete (e.g. Granger (1982)). Despite this awareness, the underlying assumption in word sense disambiguation is generally that the inventory is complete. Although NILs are marked accordingly in data sets – e.g. in SemCor 3.5% open-class words not including proper names are marked as NILs (Miller et al., 1993) –, they are usually ignored in word sense disambiguation. For instance, the research survey by Agirre & Edmonds (2006) only contains a chapter on word sense discrimination (Pedersen, 2006), but lacks a chapter on unknown sense detection in the framework of disambiguation with respect to an inventory. In 2014, Lau et al. (2014) calls the detection of senses that are not described in the inventory a novel task. Indeed, only few work addresses this question (Erk, 2006; Cook et al., 2013; Lau et al., 2014) and it is not yet embedded in the overall word sense disambiguation task.

**Word Sense Discrimination.** A concurrent task definition has been developed based on the distributional hypothesis stating that words occurring in similar contexts have similar meanings (Harris, 1968). Instead of using a predefined inventory, word tokens are clustered so that all occurrences that have the same meaning are in the same cluster (e.g. Schütze (1998) and Pedersen (2006)). This alternative task definition is highly influenced by the vector space model from information retrieval (Salton & McGill, 1983) and often called *sense discrimination* or *sense induction*. In this thesis, we are mainly interested in *token-based approaches*, i.e. the clustering of tokens. Token-based approaches are an alternative to disambiguation in case

a downstream application only requires to distinguish if some occurrences have the same or a different meaning. One such example is the clustering of web search query results – an evaluation setting that has been used in a recent shared task at SemEval 2013 (Navigli & Vannella, 2013). Instead of seeing ambiguity resolution as a labeling task, it is addressed as a clustering task. Token-based approaches have so far mainly been evaluated based on occurrences of a few types, but not in an all-words setting. As in word sense disambiguation, proper nouns are not considered.

Another research direction are *type-based approaches* where word types with similar meanings are clustered (Landauer & Dumais, 1997; Lin, 1998; Agirre et al., 2006, inter alia). However, such type-based approaches are not an alternative to word sense disambiguation approaches, but rather suitable to construct an inventory in a corpus-driven manner that is then used by a disambiguation approach (Schütze, 1998; Agirre et al., 2006). Two shared task at SemEval have been designed along this research line: the systems automatically induce sense clusters that are then used as an inventory to label unseen occurrences (Agirre & Soroa, 2007; Manandhar et al., 2010). While this cascaded scenario has been introduced early (e.g. Schütze (1998)), it has – to our knowledge – never been proposed to boost sense disambiguation by combining it with sense discrimination or to use sense discrimination after disambiguation to cluster the NILs.

**Discussion.** Work in the area of lexical semantics focuses on common nouns, verbs, adjectives and adverbs. As already indicated, proper names are excluded due to the wide-spread assumption that they lack a denotation. In addition, WordNet, the most commonly used inventory, contains only a few proper names. In contrast, we focus on all nouns including both common nouns and proper names. This is in line with current trends in e.g. Moro et al. (2014b) where both common nouns and proper names are jointly disambiguated using the same representation for both noun types. The inclusion of proper names and with it the choice of a different inventory is the key distinction between concept disambiguation and clustering, as we propose it in this thesis, and work in lexical semantics. However, as we assume that proper names behave similarly as common nouns, the methods and features that have been proposed in the area of lexical semantics are highly relevant to concept disambiguation and clustering.

On the other hand, work in lexical semantics can also benefit from the proposed approach, but also from advances in information extraction (Section 10.1.2). For instance, the task addressed in this thesis – inspired by the entity linking task at TAC (Ji et al., 2011) – goes beyond the setting in lexical semantics. Disambiguation and clustering are not seen as two rivaling research paradigms as in lexical semantics (Agirre & Edmonds, 2006), but as complementary tasks that are both necessary to provide an integrated solution for non-NILs and NILs. In this thesis, we even go one step further, by modeling disambiguation and clustering jointly. Although word sense discrimination approaches are often evaluated against sense

annotated data (Navigli & Vannella, 2013; Agirre & Soroa, 2007), one might object – from the perspective of lexical semantics – that such an integrated view suffers from potential mismatches between induced clusters and predefined senses in an inventory (Kilgarriff, 1997; McCarthy, 2006a). However, we argue that by using a more coarse-grained inventory than WordNet (e.g. Wikipedia) such mismatches are negligible.

### 10.1.2 Information Extraction

Currently, ambiguity resolution is a hot topic in the field of information extraction. In contrast to work in lexical semantics, the focus lies on proper names (Ji & Grishman, 2011; Ji et al., 2011). The assumption is that proper names refer to real world entities. Consequently, disambiguating proper names is understood as identifying the real world entities referred to (e.g. Dredze et al. (2010)) represented by the entries in an inventory. Analogously, the clustering of proper names is conceived as grouping them according to the real world entities referred to. Hence, as in lexical semantics, two main task definitions have emerged, although on entity and not on concept (or sense) level.

**Cross-document Coreference Resolution.** The entity clustering task has been introduced by Bagga & Baldwin (1998b) under the name of cross-document coreference resolution. As the name already suggest, the task is defined at the reference level and considered as an extension of the within-document coreference resolution task across documents. While the task definition of Bagga & Baldwin (1998b) includes persons, places, events and concepts, they only evaluate their approach for person names, as most of the later work (Mann & Yarowsky, 2003; Gooi & Allen, 2004, *inter alia*). One of the motivations to focus on persons is that person name queries are frequent in web queries. The Web People Search task, which has been organized three times since 2007, focuses exactly on this scenario and asks the participating systems to cluster web pages for person name queries into different entities (Artiles et al., 2007; 2008; 2009; 2010). Hence, while in cross-document coreference resolution all mentions are clustered, only selected ones are considered in the Web People Search task. Rao et al. (2010) and Singh et al. (2011) extend the cross-document coreference task by including organizations and locations in their evaluation, while Lee et al. (2012) propose an approach to cluster entities and events. In most work (e.g. Bagga & Baldwin (1998b), Mann & Yarowsky (2003)), including publications related to the Web People Search task (Artiles et al., 2010), the focus lies on ambiguity resolution. Only later work also accounts for variability of mentions, but only to a certain extent (e.g. Baron & Freedman (2008), Rao et al. (2010), Singh et al. (2011)).

In contrast to work in word sense discrimination, scalability has been considered early. Gooi & Allen (2004), Rao et al. (2010) and Singh et al. (2011) aim to create large-scale

evaluation data sets to test the scalability of their proposed approaches.

**Entity Disambiguation and Detection of Unknown Entities.** Entity disambiguation is the task of linking proper names to their corresponding entries in an inventory (Ji & Grishman, 2011). Bunescu & Paşca (2006) are the first to propose to use Wikipedia as an inventory for person name disambiguation. One of the seminal approaches in the area of entity disambiguation is described in Cucerzan (2007). In this work, the task is extended to all proper names. Almost all work in entity disambiguation uses Wikipedia or a derived resource as inventory with a few exceptions (Stoyanov et al., 2012; Sil et al., 2012). In contrast to word sense disambiguation, the detection of unknown entities is from the beginning incorporated in the task definition and addressed by most work (Bunescu & Paşca, 2006; Dredze et al., 2010; Zhang et al., 2010, *inter alia*). This strong awareness for unknown entities might be caused by the higher productiveness of proper names in contrast to e.g. common nouns. One act of an *initial baptism* (Kripke, 1980, p. 95) is usually sufficient to introduce a new usage pattern for a proper name, while the meaning of common nouns only changes slowly. Recently, research interest has grown even stronger in this direction (Li et al., 2013; Hoffart et al., 2014).

Since 2009, research in entity disambiguation has been heavily boosted and shaped by the entity linking shared task at TAC (McNamee & Dang, 2009; McNamee, 2010; Ji & Grishman, 2011; Ji et al., 2011). The entity linking task is defined in the style of a “lexical” sample task, i.e. the systems are required to disambiguate a few selected challenging proper names of type person, organization and geo-political entity per text. The task comprises three subtasks (Ji & Grishman, 2011; Hachey et al., 2013): candidate entity identification, entity disambiguation and recognition of NILs. In contrast to word sense disambiguation, the candidate identification step has received more considerations – probably because the ambiguity tends to be higher for proper names than for common nouns and variations in writing, e.g. abbreviations or short forms, are more common for proper names than common nouns (Hachey et al., 2013). Being part of the knowledge base population track at TAC, the extension of the inventory with unknown entities has become an essential part of the task: since 2011, the entity linking task incorporates entity clustering as a fourth additional subtask (Ji et al., 2011). By not only detecting, but also clustering unknown entities, more information can be collected about them which can then serve as a basis for a new entry in the inventory. To our knowledge, the entity linking task at TAC is the first attempt to use clustering complementary to disambiguation to address proper names that refer to unknown entities.

2014 is the first time that the entity linking task at TAC requires to recognize and disambiguate all proper names in a text. Similar, the *Entity Recognition and Disambiguation Challenge* at SIGIR 2014 also demands to recognize the proper names and to disambiguate them. However, unknown entities are not part of the latter evaluation campaign.

**Discussion.** In entity disambiguation, the close relationship to word sense disambiguation methods has been recognized since the beginning (e.g. Cucerzan (2007), Ratinov et al. (2011), Hachey et al. (2013), Charton et al. (2014), Moro et al. (2014b)). The two tasks mainly vary in their underlying linguistic assumptions: word sense disambiguation focuses on the concept level while work in proper name disambiguation aims to resolve references to (real-world) entities. Whereas in proper name disambiguation it is often assumed that textual references directly point to real-world entities, we argue that proper names behave similar to common nouns.

As also indicated by Moro et al. (2014b), work tends to be duplicated in the word sense and entity disambiguation areas. We argue that not only work in the field of information extraction could benefit from work in word sense disambiguation, e.g. in terms of disambiguation algorithms and features, but also vice versa. The higher ambiguity and variability of proper names (Hachey et al., 2014; Moro et al., 2014b) lead researchers in this area to focus on the incompleteness of the inventory and on the candidate identification – both aspects that have been neglected in word sense disambiguation. This thesis is a step towards exploiting insights from both areas.

### 10.1.3 Information Retrieval

With the availability of large-scale inventories such as Wikipedia, DBPedia and FreeBase in the last years, work on ambiguity resolution has boomed in the field of information retrieval. Some of the most seminal papers in concept disambiguation with Wikipedia such as Milne & Witten (2008b) and Kulkarni et al. (2009) have been published in the field of information retrieval.

Disambiguation in this area often aims to obtain a concept-based text representation that overcomes shortcomings of the bag of words approach such as ambiguity and variability (Milne & Witten, 2008b; Kulkarni et al., 2009). While Kulkarni et al. (2009) aim to link as much as possible in a text, Milne & Witten (2008b) focus on the disambiguation of keywords that represent the main topics of a text. The linking of keywords is an application on its own, known as automatic hyperlink insertion, link discovery, smart tagging. Examples of such systems are Microsoft Smart Tags, Google's Autolink or the movie linking system of Drenner et al. (2006). If Wikipedia is used as a resource, the task is called wikification (Csomai & Mihalcea, 2008; Milne & Witten, 2008b). Both in the INEX link-the-Wiki track (Huang et al., 2008; 2010) and the NTCIR cross-lingual link discovery task (Tang et al., 2011; 2013), participating systems are required to insert hyperlinks to Wikipedia articles in the same or a different language.

In contrast to work in lexical semantics or information extraction, work in the area of information retrieval tends to address both common and proper names. Disambiguation is

primarily seen as a linking task allowing to interrelate resources. This view affects the evaluation setup in at least three ways. (1) Not only the denotation or reference of a word is of interest, but also entries that are related, relevant or useful given a keyword or a text. For instance, in the link discovery task, participating system can return up to five links per keyword (Tang et al., 2011; 2013). (2) Systems often rank the provided annotations by a confidence score. This allows to balance between recall and precision (Tang et al., 2011; 2013). (3) The aim is often to provide an annotation or representation of the whole text and not of tokens as in information extraction or word sense disambiguation. Accordingly, it is common to evaluate systems on document level instead of on token level (Tang et al., 2011; 2013; Milne & Witten, 2008b). NILs are usually not part of the evaluation.

Although the task definition is different in information retrieval, the methods and features are similar to the ones in lexical semantics and information extraction. The approach proposed in this thesis is inspired by e.g. Milne & Witten (2008b) and Kulkarni et al. (2009).

#### 10.1.4 Discussion

As this overview indicates, all the above mentioned perspectives shade a different light on ambiguity resolution of open-class words by stressing different aspects. Considering all these setups helps to better understand the problem. Currently, the exchange between these communities as reflected in citations is rather limited. With this overview, we hope to help improving this exchange.

Of course, this overview is not exhaustive. For instance, the database community works on ambiguity resolution in the field of record linkage (Fellegi & Sunter, 1969; Winkler, 2006) or author name disambiguation (Ferreira et al., 2012). Other work specializes on disambiguation in a specific domain such as in the gene normalization task at BioCreative (Lu et al., 2011).

## 10.2 Methods

In the following we compare approaches based on their respective problem formulation. In word sense disambiguation, it is common to distinguish between *supervised methods* that need annotated data (Màrquez et al., 2006), *knowledge-based methods* that only require an inventory, but no annotated data (Mihalcea, 2006) and *unsupervised or knowledge-lean methods* (Pedersen, 2006) that neither make use of an inventory nor of annotated data. The amount of supervision an approach requires is highly relevant and discussed in Section 10.4. However, we argue that a comparison based on different task formalizations sheds more light on the underlying linguistic assumptions and thus focus on it.

We first discuss disambiguation approaches (Section 10.2.1) and approaches to recognize NILs (Section 10.2.2), before we review clustering approaches (Section 10.2.2). To better

compare the methods, we use the same terminology and notation throughout the whole section independent of the terminology of the respective paper.

### 10.2.1 Methods for Concept Disambiguation

Concept disambiguation can be described as a labeling problem. The mentions are the objects to label, while the concepts form the labels. What distinguishes disambiguation from many other labeling tasks in natural language processing such as named entity recognition or part of speech tagging is its enormous size of the label space. In case of Wikipedia, the label space is in the millions. If all concepts in the inventory were considered as candidate labels for each mention, it would be extremely inefficient and almost impossible to solve the task. Therefore, all disambiguation approaches first prune the label space for each mention to a few candidate concepts leading to mention-specific label spaces (Navigli, 2009). In the following we first discuss different formulations of the pruning problem, i.e. the candidate concept identification step, before we compare different formalizations of the actual disambiguation step.

**Candidate Concept Identification.** Comparative studies, e.g. Hachey et al. (2013), have shown that the identification of candidate concepts is crucial for the final performance of a system. It influences the upper bound of the final performance, which should be as high as possible, and the average ambiguity, which is ideally only as high as necessary. While the candidate concept identification is considered as a simple lexicon look-up step in word sense disambiguation with WordNet, it has obtained more attention in disambiguation with Wikipedia. The candidate concept identification step comprises three decisions, i.e. the *selection of sources*, the *generation of string variations* and the *look up strategy*.

*Selection of Sources.* In contrast to WordNet, Wikipedia does not contain a predefined lexicon that maps linguistic realizations to concepts. However, such a mapping can be extracted from different sources. A commonly used source are *anchors* (Cucerzan, 2007; Csomai & Mihalcea, 2008; Milne & Witten, 2008b; Zhang et al., 2010; Zhou et al., 2010a; Ferragina & Scaiella, 2012, inter alia). As they can be noisy, some systems – including ours – clean them and only consider them if they serve more than  $n$  times as an anchor in Wikipedia (Cucerzan, 2007; Milne & Witten, 2008b; Csomai & Mihalcea, 2008; Ferragina & Scaiella, 2012). Wikipedia article titles are also useful lexicalizations. Hachey et al. (2013) showed that a system that only looks up Wikipedia article titles and returns NIL if none can be found, achieves 71% accuracy in the TAC 2009 test set. Redirects that are considered by most systems (e.g. Bunescu & Paşca (2006), Zhang et al. (2010), Han & Sun (2011)), are of similar high quality. Other sources that are noisier are for instance disambiguation pages (Bunescu & Paşca, 2006; Zhang et al., 2010) or bold terms extracted from the first paragraph in Wikipedia (Varma et al., 2009). In addition, external sources have

been used. For instance Zhou et al. (2010a) extract lexicalizations from query logs by identifying search terms that end in a click on a Wikipedia article. Sometimes the Wikipedia search engine (Zhang et al., 2010) or another search engine (Fader et al., 2009; Dredze et al., 2010; Guo et al., 2011) is used directly to obtain the top  $n$  pages for a mention. However, as the use of commercial search engines is problematic in a scientific context – reproducibility is not guaranteed and the ranking algorithm is not public –, they are forbidden in the TAC shared tasks.

*Look up Strategy.* Besides the sources for lexicalization, the look-up strategy is relevant. The simplest strategy is to require an exact match between a look-up string and a lexicon entry (Milne & Witten, 2008b; Csomai & Mihalcea, 2008; Ferragina & Scaiella, 2012). Other approaches also allow fuzzy matches (e.g. Dredze et al. (2010)) leading to 78 candidates per mention on the TAC 2009 data set. Hachey et al. (2013) compare different candidate concept identification strategies for the TAC data and conclude that a cascaded look up (backoff strategy), where first safe sources (e.g. articles, redirects) followed by less safe sources are queried, leads to good results. Recently, Dalton & Dietz (2013) propose a joint approach for the candidate look up and the disambiguation (see below). They define the whole concept disambiguation problem as an information retrieval problem and consider the context of a mention while looking up and ranking candidates. Cheng & Roth (2013) follow a similar strategy and use relational queries to extend the candidate concepts of mentions. Given two mentions that are in a textual relation, they fix the concept of one of them and search for matching relation triples obtained from DBpedia and from the internal link structure from Wikipedia.

*Generation of String Variations.* Finally, the effective query string is relevant for the results. While we query both the token string of a mention and its lemma, other approaches first correct misspellings (e.g. Zhang et al. (2010)). Hachey et al. (2013) show that including string variations obtained from coreferent mentions improve the final results for proper name resolution. In contrast to all other approaches (Cucerzan, 2007; Guo et al., 2011; Hachey et al., 2013; Cheng & Roth, 2013; Dalton & Dietz, 2013, inter alia) that first resolve coreference to e.g. obtain the full version of a person name or an acronym or to get additional string variations, we look up each mention separately, but allow to move candidates from one mention to another one during disambiguation.

The candidate concept identification is crucial and can affect the overall performance more than the disambiguation step. Our approach is rather conservative. However, we do not require that the correct candidate must be among the candidates of each mention, but allow the candidates to move from one mention to another via clustering during the disambiguation. The assumption is that a highly ambiguous mention is usually introduced in a text by a less ambiguous mention. This strategy may fail on short text snippets as e.g. on Twitter data.

**Concept Disambiguation.** Most of the work in concept disambiguation focuses on the effective disambiguation step leading to a wide range of proposed approaches. One of the main research contributions of Cucerzan (2007) and later Ratinov et al. (2011) is a problem formalization that is fairly general and captures most approaches. Given two  $n$ -tuples  $\mathbf{m} = (m_1, m_2, \dots, m_n)$  and  $\mathbf{c} = (c_1, c_2, \dots, c_n)$  with  $c_i$  being the disambiguation result for mention  $m_i$ , disambiguation can be described as a maximization problem with two main components: (1) a *local function*  $\phi(c_i, m_i, d)$  that measures the compatibility of a candidate concept of a mentions in dependence of some observed features and (2) a *global function*  $\psi(\mathbf{c}, d)$  that accounts for the interrelations between the concepts.  $d$  stands for all resources that are considered to disambiguate a mention, including the inventory and the current document collection. It does not include the candidate concepts of other mentions in the document collection. Assuming a linear combination between the two components (Cucerzan, 2007; Ratinov et al., 2011), disambiguation can be formalized via

$$\mathbf{c}^* = \arg \max_{\mathbf{c}} \lambda \cdot \left( \sum_{i=1}^n \underbrace{\phi(c_i, m_i, d)}_{\text{local component}} \right) + \mu \cdot \underbrace{\psi(\mathbf{c}, d)}_{\text{global component}}. \quad (10.1)$$

The two parameters  $\lambda$  and  $\mu$  represent weights. Starting from this problem formulation, we classify different methods based on how they realize these two components. This allows us to compare them based on how they model interrelations between concepts.

**Local Approaches for Concept Disambiguation.** Local approaches completely ignore interrelations between concepts and disambiguate each mention separately via first computing

$$c_i^* = \arg \max_{c_i} \phi(c_i, m_i, d). \quad (10.2)$$

for every  $i \in \{1, \dots, n\}$  and then setting  $\mathbf{c}^* = (c_1^*, \dots, c_n^*)$ . Local approaches are the most common approaches in concept disambiguation and already lead to good results (Ratinov & Roth, 2011). Most of the leading systems at SensEval and SemEval are local (Yarowsky, 2000; Veenstra et al., 2000; Mihalcea & Moldovan, 2001b; Hoste et al., 2001; Yarowsky et al., 2001; Strapparava et al., 2004; Decadt et al., 2004; Cai et al., 2007; Chan et al., 2007). But also in disambiguation with Wikipedia, local approaches have been successful (Bunescu & Paşca, 2006; Csomai & Mihalcea, 2008; Fader et al., 2009; Han & Sun, 2011; Gottipati & Jiang, 2011; Li et al., 2013; He et al., 2013). The chosen function for  $\phi(c_i, m_i, d)$  ranges, among others, from a single similarity measure (Lesk, 1986; Csomai & Mihalcea, 2008; He et al., 2013), decision lists (Yarowsky, 2000), memory-based approaches (Veenstra et al., 2000; Hoste et al., 2001), SVMs (Chan et al., 2007; Zhang et al., 2010), Bayesian approaches (Csomai & Mihalcea, 2008; Fader et al., 2009; Han & Sun, 2011; Li et al., 2013; Jin et al., 2014)

to the combination of different classifier results (Csomai & Mihalcea, 2008; Yarowsky et al., 2001). For a discussion of different local approaches in the field of word sense disambiguation the reader is referred to Mårquez et al. (2006) and Navigli (2009).

Investigating local approaches that use Wikipedia as inventory, we identify two main developments.

First, the problem formalization has changed. Local supervised approaches in lexical semantics consider the disambiguation of each word type as a separate classification task involving a separate classifier (Navigli, 2009). This leads to an enormous amount of classifiers requiring even more training instances. Occurrences of types for which no or only a few annotated training data are available cannot be disambiguated with such an approach. Using Wikipedia as an inventory changes the problem formulation, as an underlying assumption made by WordNet based approaches is suddenly not taken for granted anymore. More precisely, the stable relation between candidate senses and lexicalizations that WordNet suggests is unmasked as shaky considering the dynamics of Wikipedia, particularly with respect to proper names. Thus, an approach that is based on a stable relation between candidate concepts and lexicalization such as the separate classifier approach is not suitable anymore. In addition, scalability becomes a more and more important aspect when it comes to web-based applications. Hence, only early work in disambiguation with Wikipedia trains a separate classifier for each type (Csomai & Mihalcea, 2008; Bryl et al., 2010). Most classification-based approaches that use Wikipedia as an inventory apply one single binary classifier to all mentions. For each mention to disambiguate, all candidate concepts are classified as valid or invalid. If more than one candidate concept of a mention is classified as valid, the one with the highest confidence value (Milne & Witten, 2008b)<sup>2</sup> or some highest feature value (Zhang et al., 2010) is selected as the correct disambiguation. Such a binary classifier requires less training instances to learn the model parameters and can also be applied to mentions for which no training data is available. In addition, it does not assume a stable relation between lexicalization and candidate concepts. Framing disambiguation as a binary classification task has also changed the features. Lexical features that are heavily used in work with separate classifiers (Csomai & Mihalcea, 2008; Bryl et al., 2010), need to be transformed into numerical scores if a binary classifier is used (Zhang et al., 2010; Fader et al., 2009).

Second, in work with Wikipedia supervised ranking algorithms become more prominent, which may be partially due to the influence of information retrieval. Bunescu & Paşca (2006) use a kernel-based ranking approach. Dredze et al. (2010) and Zheng et al. (2010) both use  $SVM_{Rank}$  (Joachims, 2002), while Chen & Ji (2011) compare different learning-to-rank algorithms. ListNet (Cao et al., 2007), a listwise ranking approach, gives the best results that

<sup>2</sup>Milne & Witten (2008b) include global information, hence we discuss this work in the next subsection.

can only be beaten if the output of several rankers are combined using voting (Chen & Ji, 2011).

In summary, local approaches assume that observed features such as lexical co-occurrences and derived similarity scores, part-of-speech tags and prior probabilities are sufficient to disambiguate a mention. By not modeling any interrelations between concepts, they are more time efficient than global approaches at the cost of lost information.

**Global Approaches for Concept Disambiguation.** In recent years, global approaches have been booming. We consider all disambiguation approaches that model interrelations between concepts from different mentions as global – independently from whether they are aggregative, iterative or joint (Section 4.1). Hence, global approaches consider at least the second component in Equation 10.2.1, although they usually take into account both. From a purely methodological perspective, different global approaches vary in how they realize the global component  $\psi(\mathbf{c}, d)$  and combine it with the local one. In most cases, the local and global components are combined via addition. Given  $\psi(\mathbf{c}, d)$  stands for the interrelations between the concepts of all mentions, solving Equation 10.2.1 is a NP-hard problem (Cucerzan, 2007) and requires in most cases an approximation.

We classify global approaches based on how the global component is realized, approximated and combined with a local component. Before we discuss some approaches in more detail, we give a broad overview over global approaches by discussing the classification shown in Table 10.1.

*Pairwise vs. Non-pairwise Approaches.* Most global approaches that have been proposed for concept disambiguation are pairwise, i.e. the interdependencies between concepts is modeled pairwise via

$$\psi(\mathbf{c}, d) = \sum_{\langle c_i, c_j \rangle \in \text{pair}(\mathbf{c})} \psi_{pw}(c_i, c_j, d) \quad (10.3)$$

with  $\text{pair}(\mathbf{c})$  consisting of all unique concept pairs. The function  $\psi_{pw}(c_i, c_j, d)$  takes two concepts  $c_i$  and  $c_j$  as input and returns a measure that indicates the relatedness between the two (e.g. Milne & Witten (2008a)). All pairwise approaches can be represented by a graph where the candidate concepts form the vertices and the interrelations between them are indicated by weighted or unweighted edges. How this graph exactly looks like depends on the respective approach. The local score of a candidate concept  $\phi(c_i, m_i, d)$  can for instance be represented by vertex weights (Equation 10.2.1). Alternatively, the mentions can be included as additional vertices into the graph connected to their respective candidate concepts by edges that are weighted by the local score. The only non-pairwise approaches are Bayesian, e.g. Han & Sun (2012) and Sen (2012). They model interrelations via topics. Each mention is associated with a topic that determines the distribution over its candidate concepts. The more often a topic is

picked up, the more important it becomes for a document and the more likely it is that it is assigned to a mention. Our proposed approach is largely pairwise. Only the cluster transitivity constraint renders it in a non-pairwise approach setting it apart from most other approaches. However, as it is most similar to pairwise approaches, we ignore the transitivity constraint in the following discussion and thus consider our approach as pairwise.

*Iterative vs. Non-iterative Approaches.* Non-iterative approaches disambiguate all mentions in one pass and consider the interrelations between all candidate concepts. While *joint approaches* use exact or approximated inference techniques to obtain the optimal combination of concepts and only take into account the interrelations between selected concepts, *aggregative approaches* account for the relations between all connected candidate concepts, independently from whether they are part of the final solution. The latter method can misguide the disambiguation process, as a candidate concept can be highly related to non-selected candidates (Section 2.2). Iterative approaches first obtain a (partial) disambiguation result based on a local approach that then builds the disambiguation context for the actual global disambiguation. In this way, less noise is introduced than in the non-iterative aggregative approach. However, it requires that the first disambiguation step is of sufficient quality. All currently proposed iterative approaches are aggregative. The approach proposed in this thesis uses a hybrid strategy and models safe cases non-iteratively jointly, while unsafe interrelations are modeled similar to Milne & Witten (2008b) using an iterative aggregative approach (Section 4.1).

*Weighted vs. Unweighted Combination of the Local and Global Component.* Pairwise approaches can either consider the local and global factors as equally important (unweighted combination) or assign them different weights for the final score ( $\lambda$  and  $\mu$  in Equation 10.2.1). In the latter case, some annotated data is necessary to tune or learn the weights, as in our proposed approach. However, if an approach is unweighted, this does not imply that the approach is unsupervised. For instance, Kulkarni et al. (2009) train the local component independent from the global one on annotated data.

After this broad overview, we discuss some representative and prominent approaches from each category in Table 10.1 in more details. Approaches that we do not explain in the following are methodically similar to the other ones in their respective category.

*Iterative Pairwise Approaches.* An example for this category is the seminal work of Milne & Witten (2008b) which focuses on the identification and disambiguation of keywords in a text. As all other approaches in this category, Milne & Witten (2008b) approximate the global component of the NP-hard problem (Equation 10.2.1) via

$$\psi(\mathbf{c}, d) = \sum_{i=1}^n \sum_{c_j \in \mathbf{c}'} \psi_{pw}(c_i, c_j, d) \quad (10.4)$$

Non-Iterative				Iterative Aggregation	
Aggregation		Joint Inference		Unweighted Combination	Weighted Combination
Unweighted Combination	Weighted Combination	Unweighted Combination	Weighted Combination	Unweighted Combination	Weighted Combination
<b>Pairwise Models</b>					
Han et al. (2011) Cucerzan (2007) Guo et al. (2011) Moro et al. (2014b)	Ferragina & Scaiella (2012) Chen & Ji (2011) Zhou et al. (2010a)	Turdakov & Lizorkin (2009) Kulkarni et al. (2009)	<b>ML Dis+NILs+Clust Scope</b> <b>ML Dis+NILs+Clust</b> Cheng & Roth (2013) Dai et al. (2011) Hoffart et al. (2011b)	Charton et al. (2014)	<b>ML Dis+NILs</b> Milne & Witten (2008b) Dalton & Dietz (2013) Ratinov et al. (2011)
<b>Non-Pairwise Models</b>					
		Han & Sun (2012) Sen (2012)			

Table 10.1: Classification of global methods according to how they model interrelations between concepts.

with  $c'$  being the disambiguation context. The disambiguation context consists of concepts that are obtained during a preceding disambiguation step. In the corresponding graph in Table 10.1, the concepts that are selected as disambiguation context are marked by stripes. Milne & Witten (2008b) assume that a text contains a sufficient number of unambiguous mentions and use their denoted concepts as disambiguation context. For each candidate concept of the ambiguous mentions, relatedness is then calculated to each concept in the disambiguation context and the scores are aggregated. The local and global features are combined using machine learning, more precisely with a Bagged C4.5 algorithm (Witten et al., 2011). Our approach, in particular our relatedness features, is based on Milne & Witten (2008b). The approach of Milne & Witten (2008b) has been highly influential. For instance, both Ratinov et al. (2011) and Ferragina & Scaiella (2012) build upon the ideas in Milne & Witten (2008b) and redefine the disambiguation context to avoid relying on unambiguous mentions which might be risky in shorter texts.

To obtain a disambiguation context, Ratinov et al. (2011) first apply a local disambiguation model to all mentions and use the disambiguation results as disambiguation context. All mentions are then disambiguated again, this time by also considering the global component. For each candidate concept of each mention, relatedness is calculated to all concepts in the disambiguation context and aggregated by choosing the average and the maximum score. Instead of only using one relatedness measure such as Milne & Witten (2008b), several measures are

used.

All these aggregated relatedness scores are then linearly combined with the features from the local model. The weights for the different features are learned with a linear ranking support vector machine. The results of this global model are mixed. The final model that accounts for interrelations between mentions fails to outperform the local one on some data sets. The authors conclude that sometimes the local, sometimes the global model leads to the correct results. This is a hint that not all mentions are equally important for their mutual disambiguation and that it is not promising to use the same context definition for all mentions (Section 10.3).

Dalton & Dietz (2013) propose a method rooted in information retrieval that jointly identifies and ranks candidate concepts of a mention. They assume that one single mention in a document needs to be disambiguated, but that context mentions can help to identify and rank candidate concepts. By exploiting unambiguous mentions – their relatedness function returns 1 if a context mention unambiguously denotes a concept which is an outgoing or incoming link –, they manage to integrate the global component into their retrieval engine framework and thus to perform joint candidate identification and disambiguation with a global model. The top  $n$  candidate concepts returned by this linear model are then reranked using an additional supervised ranking model with more features.

*Non-iterative Pairwise Approaches based on Aggregation.* Cucerzan (2007) is an example of an aggregative approach with an unweighted combination of the local and global component. It approximates the NP-hard problem via

$$\mathbf{c}^* = \arg \max_{\mathbf{c}} \sum_{i=1}^n \left( \underbrace{\phi(c_i, m_i, d)}_{\text{local component}} + \underbrace{\sum_{j=1 \wedge j \neq i}^n \sum_{c_{j,k} \in \text{cand}(m_j)} \psi_{pw}(c_i, c_{j,k}, d)}_{\text{global component: } \psi(\mathbf{c}, d)} \right) \quad (10.5)$$

with  $\text{cand}(m_j)$  denoting all candidate concepts of the mention  $m_j$ . The function  $\psi_{pw}(c_i, c_{i,k}, d)$  measures the pairwise relatedness between two concepts. Instead of summing over the pairwise relatedness scores between the selected concepts, the scores between all candidate concepts of different mentions are taken into account. With this aggregative approach, each mention can be solved independently. Cucerzan (2007) proposes an unsupervised vector space model for proper name disambiguation using Wikipedia as an inventory. Each concept  $c$  in Wikipedia is represented with two binary vectors with contextual surface clues  $\text{clues}_c$  and category tags  $\text{cat}_c$  as dimensions respectively. Given a document to disambiguate, the frequency counts of the contextual surface clues are stored in a document vector  $\text{clues}_{\text{doc}}$ . Each mention is then disambiguated separately taking the relations to all candidate concepts of the other mentions in the document into account with the local component defined as

$$\phi(c_i, m_i, d) = \langle \text{clues}_{c_i}, \text{clues}_{\text{doc}} \rangle \quad (10.6)$$

and the global component given by

$$\sum_{j=1 \wedge j \neq i}^n \sum_{c_{j,k} \in \text{cand}(m_j)} \psi_{pw}(c_i, c_{j,k}, d) = \sum_{j=1 \wedge j \neq i}^n \sum_{c_{j,k} \in \text{cand}(m_j)} \langle \text{cat}_{c_i}, \text{cat}_{c_{j,k}} \rangle. \quad (10.7)$$

In contrast to other approaches in the same vein that have been proposed for word sense disambiguation with WordNet and are explicitly called graph-based (Navigli & Lapata, 2010; Ponzetto & Navigli, 2010), the approach in Cucerzan (2007) is not introduced as such, although it could be represented as an  $n$ -partite graph.

Another prominent approach in this category is the voting-based method of Ferragina & Scaiella (2012) focusing on short texts and efficiency. Given a mention  $m_i$  to disambiguate and a candidate concept  $c_i$ , each other mention has one vote that is obtained by aggregating relatedness scores between all corresponding candidate concepts. The voting-based relatedness score is then combined with the local model, i.e. the prior probability of a concept, via multiplication, i.e.

$$\arg \max_{c_i} \phi(c_i, m_i, d) \cdot I \left( \sum_{j=1 \wedge j \neq i}^n \sum_{c_{j,k} \in \text{cand}(m_j)} \psi_{pw}(c_i, c_{j,k}, d) \right). \quad (10.8)$$

$I \left( \sum_{j=1 \wedge j \neq i}^n \sum_{c_{j,k} \in \text{cand}(m_j)} \psi_{pw}(c_i, c_{j,k}, d) \right)$  is an indicator function that returns 1 if the obtained score is close to the score of the most related candidate concept of a mention and 0 otherwise. They also experiment with machine learning methods, but this more simple approach that only requires for a threshold leads to similar results.

In contrast to all other approaches discussed so far, Han et al. (2011) propose to aggregate relatedness scores via a propagation algorithm. They build a directed graph including mentions and candidate concepts as vertices. The edges between the candidate concepts are weighted by a relatedness score normalized by the out degree, while the edges between the mentions and candidates are weighted by the score of the local model again normalized by the respective out degree. Each vertex is then associated with a score which indicates its importance given the graph structure calculated via a personalized PageRank algorithm (Agirre & Soroa, 2009). For initialization, each mention vertex is assigned its tf-idf score reflecting its importance. Via the personalized PageRank algorithm this information is propagated through the graph (Section 10.3). To obtain the final ranking of the candidate concepts the corresponding vertex score is multiplied with the score of the local model.

To our knowledge, the collaborative ranking approach of Chen & Ji (2011) is the first

approach that leverages cross-document clustering relations for disambiguation. They use a pipeline-based approach. Given a mention to disambiguate, first at most 300 additional documents that contain the same mention string are identified and clustered using a bag-of-words document representation. The assumption is that the target strings in all documents that are in the same cluster as the target document denote the same concept and are thus so called collaborators. Instead of only disambiguating the target mention in the target document, also its collaborators are disambiguated using a ranking-based approach. Depending on the ranking algorithm, all mentions in the cluster are disambiguated separately and their disambiguation results are aggregated (unweighted combination) or features from the collaborators are aggregated and combined in a supervised way (weighted combination). In their final system, they combine the output of different rankers and apply both strategies. Considering cross-document information can improve the results by 1.4% in case of the best ranker, i.e. ListNet. These results are in line with ours.

*Non-iterative Pairwise Approaches based on Joint Inference.* In the field of disambiguation with Wikipedia, Kulkarni et al. (2009) is the first approach that does not simplify the NP-hard problem by aggregation, but perform joint inference using hill-climbing and rounding ILPs. Their objective (without NILs) is given by

$$\mathbf{c}^* = \arg \max_{\mathbf{c}} \frac{1}{n} \left( \sum_{i=1}^n \phi(c_i, m_i, d) \right) + \frac{1}{2} \sum_{\langle c_i, c_j \rangle \in \text{pairs}(\mathbf{c})} \psi_{pw}(c_i, c_j, d) \quad (10.9)$$

where  $\text{pairs}(\mathbf{c})$  consists of all unique pairs in  $\mathbf{c}$ . The local weight vector is learned separately using a max-margin technique. Similar to Ratinov et al. (2011), their experimental results are mixed. On the IITB data set, the local approach already achieves high results. Depending on the inference algorithm, the global approach slightly lowers the results or improves the results by almost one percent. One of the reasons for these rather mixed results could be the assumption that all concepts are equally relevant for their mutual disambiguation, as also indicated by the authors (Section 10.3).

Hoffart et al. (2011b) approximate the NP-hard problem with a graph-based approach. Given a document, an undirected graph is built with candidate concepts and mentions as vertices. The mentions are connected to their respective candidates via edges weighted by the local component score. The edges between the candidate concepts are weighted by a relatedness score. Similar to Moro et al. (2014b), Hoffart et al. (2011b) combine aggregative and approximate inference techniques. First, the graph is pruned by removing weakly connected candidate concepts until the number of candidate vertices is at most five times bigger than the number of mentions. Second, an approximate algorithm to identify a subgraph with maximum density is applied with the restriction that for each mention one candidate concepts remains. The density of a subgraph is not measured by the summed edge weights, but by its minimum

weighted degree, as this better allows to account for weaker connected candidate concepts. As Kulkarni et al. (2009) and Ratinov & Roth (2011), they already achieve high results with their local model for proper name disambiguation on the CoNLL data. Including the global component further improves these results. Their strategy to activate the global component only for some mentions plays a significant role in these results (Section 10.3).

Another joint approach, which is close to ours, is the one of Cheng & Roth (2013). They formulate the problem as an ILP. In contrast to Kulkarni et al. (2009) and Hoffart et al. (2011b), Cheng & Roth (2013) only model a few selected interrelations between concepts (Section 10.3). They use relation extraction techniques and coreference resolution to decide if the interrelations between two mentions should be activated (Section 10.3). The sparsity of their interrelations and the improved performance of state-of-the-art ILP solvers – like us they use *Gurobi* (Gurobi Optimization, 2014) –, allow them to perform exact inference. Their ILP formulation combines the global component linearly with the system of Ratinov & Roth (2011) – which consists by itself of a local and a global aggregative component – using ILP. The intuition behind the approach of Cheng & Roth (2013) is similar to ours, i.e. to focus on a few high precision interrelations between concepts in the joint modeling part and to model the other interrelations aggregatively. In contrast to us, they do not learn weights for the global relational features, but set them manually. Their results on various data sets show that the coreference relations lead to improvements on the MSNBC data sets, but not on ACE 2004 or the AQUAINT data sets. Their relational component improves the results on ACE 2004 and MSNBC, but not on the AQUAINT data sets. It is unclear if these improvements are significant.

Similar to Cheng & Roth (2013) and us – and in contrast to most other approaches –, Dai et al. (2011) distinguish between different types of interrelations between concepts. They aim to link gene mentions jointly to the EntrezGene database using Markov Logic. For each interrelation type a separate global first-order logic formula is designed. In total, Dai et al. (2011) use three different such global formulas. Two of them address cohesive ties of type identity, whereas one models cohesive ties of type relatedness. For each of these global formulas a separate weight is learned. The weighted global scores are then linearly combined with the scores of the weighted local formulas. As Cheng & Roth (2013) and we, they focus on a few high precision relations and only disambiguate these mentions jointly.

Turdakov & Lizorkin (2009) propose an HMM-based model with multiple lexical chains for disambiguation with Wikipedia. Each mention can either join some previous lexical chains or start a new chain. To decide if a mention joins some previous chains, pairwise relatedness between the concepts of a fixed number of mentions (so called active mentions) in each chain and the candidates of the current mention is calculated.

*Discussion.* Most recently proposed approaches model a local and a global component. In approaches that use Wikipedia as an inventory, we can observe three main trends that

are connected with each other. First, the widely spread “assumption of maximal cohesiveness” (Section 2.2.2) tends to be given up. It says that the more and stronger cohesive ties can be established between selected concepts in a text or text part the better. While earlier work that uses Wikipedia as an inventory such as Cucerzan (2007), Kulkarni et al. (2009) or Ratinov & Roth (2011) still models this intuition, Kulkarni et al. (2009) discuss that mentions in a document may belong to different topical clusters or even singletons that share no cohesive ties with other mentions. Already Milne & Witten (2008b) assume that not all mentions are equally relevant to disambiguate each other and Hoffart et al. (2011b) only activate the global component if the features values of the local component contradict each other. Also recent generative approaches follow this trend by modeling a document as a mixture of topics (e.g. Han & Sun (2012), Sen (2012)). We will discuss this development in more detail in Section 10.3. Second, instead of trying to capture all cohesive relations between mentions, more recent approaches restrict themselves to a few high precision relations such as for instance between coreferent mentions (e.g. Chen & Ji (2011), Dai et al. (2011), Cheng & Roth (2013)). The underlying assumption is that a few good relations are sufficient for disambiguation. Third, given the advances in joint inference and given the tendency to only model a few high precision cohesive relations which leads to sparseness, it becomes feasible to resolve the NP-hard inference problem (Cheng & Roth, 2013). In Table 10.1, the different versions of our approach are added. While our aggregative approach (*ML Dis+NILs*) is most similar to Milne & Witten (2008b), our joint concept disambiguation and clustering approach (*ML Dis+NILs+Clust*) shares most commonalities with Cheng & Roth (2013) and Chen & Ji (2011) respectively. As we combine aggregative and joint inference techniques similar to Cheng & Roth (2013), our approach is at the intersection of iterative aggregative and non-iterative joint approaches. To our knowledge, our joint approach (*ML Dis+NILs+Clust*) is the first approach that performs disambiguation and clustering jointly.

### 10.2.2 Methods for NIL Recognition

Recently, the NIL recognition task – i.e. detecting mentions that denote concepts that are not part of the inventory (*NILs*) – has become more and more popular. Hoffart et al. (2014) distinguish between two types of NILs: (1) NILs for which no candidate concepts can be retrieved from the inventory (e.g. *tiger selfie*), and (2) NILs for which candidate concepts can be retrieved from the inventory (e.g. *Tinder*). As this distinction indicates, the capability of a system to recognize NILs not only depends on its NIL detection algorithm, but also on its candidate identification component (Hachey et al., 2013; Gottipati & Jiang, 2011).

In the following we focus on the recognition of NILs for which candidate concepts can be retrieved from the inventory. We classify previous approaches for these tasks into three main groups and discuss each of them.

**Threshold-based Approaches.** The most common way to recognize NILs is via a tuned threshold  $\tau_{NIL}$ . The intuition is that if the evidence for the candidate concept selected by the disambiguation component is only weak, it is most likely a NIL. Hence, the score returned by the disambiguation component for this candidate concept is compared to a threshold and if it is lower, the mention is considered as a NIL. While for instance Zhou et al. (2010a), Guo et al. (2011) and Han & Sun (2012) tune this threshold separately, Bunescu & Paşca (2006) and Kulkarni et al. (2009) integrate it into their disambiguation function. Both introduce a NIL concept  $c_{NIL}$  which is per default part of the candidate concepts for a mention. Bunescu & Paşca (2006), who use a linear ranking function for disambiguation, add an additional NIL feature to their model which is 1 for the candidate  $c_{NIL}$  and 0 otherwise. The weight for this feature is learned together with all other feature weights. While threshold-based approaches are straightforward to implement, it is not guaranteed that the threshold generalizes across mentions and data sets (Hoffart et al., 2014).

**Classification-based Approaches.** Instead of only relying on a threshold, classification-based approaches use a binary classifier to decide whether a candidate concept is a valid concept for a mention. If no candidate concept is valid, the mention is a NIL. While in classification-based approaches such as Zhang et al. (2010) this classifier is also used for disambiguation, ranking-based approaches require a separate validation classifier. For instance, Zheng et al. (2010), Dredze et al. (2010) and Ratinov et al. (2011) use a supervised ranker for disambiguation. After the disambiguation step, a separate binary SVM classifier decides if the highest ranked candidate concept for a mention is a valid target. This validation classifier uses similar features as the ranking module including the score returned by the ranker (Zheng et al., 2010). As this two step process can lead to error propagation, Dai et al. (2011) propose a joint approach for disambiguation and recognition of NILs using Markov Logic similar to ours. They introduce an additional hidden predicate *isSuitableforLinking* and use domain-specific features – they only focus on the disambiguation of gene mentions – to recognize NILs. As in our case, the joint approach outperforms the two-stage approach *ceteris paribus*. While their precision for the joint approach decreases, recall increases. In contrast to them, our implementation only requires one hidden predicate for the disambiguation and recognition of NILs without any loss in expressive power. Via a hard cardinality constraint we define that for each mention zero or one candidate concepts are selected and the features for recognizing NILs are defined with respect to the single hidden predicate.

**Disambiguation-based Approaches.** All approaches discussed so far – including ours – use negative evidence to recognize NILs. Han et al. (2011) are the first to detect NILs via positive evidence. Similar to Bunescu & Paşca (2006) and Kulkarni et al. (2009), they introduce a single NIL concept, but – in contrast to them – associate this NIL concept with the same

information that is available for all other concepts in the inventory, i.e. a NIL popularity score, a NIL mention string distribution and a NIL context model. For instance, the popularity of a regular concept is estimated by the relative frequency of it in an annotated corpus. By using add-one smoothing, they assign unseen regular concepts and the NIL concept a non-zero popularity score. The NIL mention distribution and NIL context models are estimated in a similar spirit via a general language model. This information is then used in their generative disambiguation approach that disambiguates each mention separately. For each candidate concept, including the NIL concept, the probability is calculated that it is denoted by the mention, combining the three information sources mentioned above. The candidate concept with the highest probability – this can also be the NIL concept – is then selected.

While the approach of Jin et al. (2014) is similar to the one of Han et al. (2011), Hoffart et al. (2014) introduce a separate NIL concept for each mention string. Each of these NIL concepts is associated with keyphrase information, which is the main information source in their disambiguation approach (Hoffart et al., 2012). Given this keyphrase information they can treat the NIL concepts in the same way as regular concepts during disambiguation. Their disambiguation scenario is for instance news streams. To extract the keyphrase information for the NIL concepts, they leverage the redundancy in texts from the same time period as the text to disambiguate. For each mention string, they harvest keyphrases from these additional texts and calculate the set difference between them and the keyphrases of the corresponding candidate concepts in the inventory. The set difference is then associated with the corresponding NIL concept. Their results are mixed. Whereas the performance for the NILs increases, the overall disambiguation performance including NILs and non-NILs decreases, in particular if global information is used for disambiguation.

**Discussion.** Recognizing NILs is a challenging task that can still be improved. It would be interesting to enrich our joint approach with some positive evidence similar to for instance Han et al. (2011). Especially for global approaches the recognition of NILs is important, as they also affect the performance of the Non-NILs (Hoffart et al., 2014).

Recognizing NILs is one necessary step in supporting concepts that are not in the inventory. To create new concepts and add them into the inventory as in ontology learning or knowledge base population, the ambiguity and variability of NILs have to be resolved and information about them needs to be extracted. While ambiguity and variability can be addressed via clustering and is part of the thesis, the extraction of additional information is not part of it. Work in this direction is for instance Lin et al. (2012) or Nakashole et al. (2013) that assign fine-grained semantic types such as e.g. GUITAR PLAYER or CONCERTS to proper names with no corresponding entry in the inventory.

### 10.2.3 Methods for Concept Clustering

In the following, we focus on token-based clustering approaches in which occurrences – and not types – are clustered (Pedersen, 2006). Since Schütze (1998), a common practice is to first induce concepts or senses from a corpus – either type- or token-based – and then label new instances with respect to these induced clusters (Véronis, 2004; Agirre et al., 2006; Agirre & Soroa, 2007; Manandhar et al., 2010). As in this thesis the aim is to cluster on the fly, we only discuss the first part of such approaches and only if they are token-based.

The main focus of token-based clustering approaches lies on ambiguity. Only a few approaches address variability, although only variability that can be captured by string edit distance and substring matching (e.g. Baron & Freedman (2008) and Rao et al. (2011)). This is also the case for our proposed approach.

Since Bagga & Baldwin (1998b) and Schütze (1998), concept clustering approaches consists of three integral parts. First, each mention is associated with a *context representation*. Then all mentions that share some string similarity are *compared in a pairwise mode*. Finally, the mentions are *clustered* based on the pairwise scores. While the context representation is discussed in the next section, we focus here on pairwise comparison functions and clustering techniques.

**Pairwise Comparison.** The most common approaches are vector space models (see Turney & Pantel (2010) for an overview). Given two vectors, each of them representing the context of a mention, the similarity between them is calculated via a similarity measure such as the cosine similarity (Schütze, 1998; Bagga & Baldwin, 1998b; Mann & Yarowsky, 2003; Purandare & Pedersen, 2004; Pedersen et al., 2005). While Rao et al. (2011) linearly combine different similarity measures into one score, Han & Zhao (2009) first align each dimension in a vector to the most related dimension in the other vector. Then the similarities are calculated based on these alignments respectively and averaged. In contrast to the standard cosine similarity, this approach does not require that the dimensions of the vectors are identical, but only related and is supposed to overcome sparsity. A graph-based extension of this approach has been proposed by Han & Zhao (2010).

A few approaches consider the pairwise comparison as a binary classification problem. For instance, Fleischman & Hovy (2004), Niu et al. (2004) and Niu et al. (2005) use a maximum entropy model and Finin et al. (2009) a SVM. In contrast to vector-based approaches, classification-based methods require training data, but allow to combine features based on different context representations. Our proposed approach is also supervised. However, in contrast to most other approaches – one exception is the generative approach of Andrews et al. (2014) that jointly learns similarities between mentions and performs clustering –, we consider transitive relations during weight learning and testing.

**Clustering.** The pairwise scores build the input for the clustering. Agglomerative bottom-up clustering has led to good results and is the most often used technique (Bagga & Baldwin, 1998b; Gooi & Allen, 2004, *inter alia*). Compared to partitive clustering approaches – for instance Pedersen (2006) uses k-means clustering –, the number of clusters, i.e. concepts, is not required to be known a priori. However, although agglomerative clustering leads to good results, it is inefficient. Thus, recently more scalable approaches have been proposed, such as e.g. streaming clustering where mentions are clustered online (Rao et al., 2011; Clarke et al., 2012; Merhav et al., 2013). In the same vein, Singh et al. (2011) exploits distributed inference across different nodes. A similar approach could be used to scale up the approach proposed in this thesis.

**Interrelations between Disambiguation and Clustering.** Only few approaches consider relations between disambiguation and clustering. Chen & Ji (2011) first cluster mentions across documents and jointly disambiguate them enforcing a common disambiguation result. A similar pipeline-based approach is used by Cheng & Roth (2013), but by exploiting within-document clustering (Section 10.2.1). The closest approach to ours with respect to interrelations between disambiguation and clustering is the one of Monahan et al. (2011). They first disambiguate the mentions, then cluster them and use the cluster information to reconsider the disambiguation decision by voting for a concept within each cluster. Another related approach that addresses both disambiguation and clustering is described in Clarke et al. (2012) and Merhav et al. (2013). In contrast to us, they formulate disambiguation as a clustering problem. For each concept in the inventory representative mentions are obtained from the concept description. These mentions are then considered as clusters and treated in the same way as the clusters that have been obtained from texts. Then a streaming clustering algorithm is applied that links a mention either to one of the clusters from the inventory or to new clusters obtained from the text.

**Discussion.** Schütze (1998) suggests that clustering is an easier problem than disambiguation as it only requires to decide if two occurrences denote the same concept, but not which one. We argue that this is only partially true. In contrast to clustering, disambiguation is much more informed, as it can exploit the information associated with each concept in an inventory. By modeling them jointly, we can benefit from both worlds.

## 10.3 Context Definitions

The context is *the* determining factor for concept disambiguation and clustering. Already Weaver (1955) asked how much context is needed to disambiguate a token (Chapter 1). In this

thesis, we claim that the context relevant to disambiguate a mention depends on its embedding into discourse and consists of all text parts (mentions, other lexical units) with which it shares concept-level cohesive ties (Definition 2.2.2). This comprises the local embedding into the sentence – for instance a verb or prepositions can constrain the semantic type of a mention –, but also cohesive relations across the whole text. However, due to the reciprocal relationship between cohesive ties and denotation – the cohesive ties depend on the respective concept and vice versa –, it is extremely difficult to identify the relevant context.

In this section, we discuss how different approaches approximate the relevant context and compare their respective assumptions. Figure 10.1 shows aspects of the context that have been studied:

**Linguistic Level.** The context can be directly modeled at the concept level. We classify all approaches that exploit such concept-level cohesive ties directly as global, as they model interdependencies between the labels (Section 10.2.1). As the concept level is challenging to access, many approaches try to exploit correlations to cohesive ties on other linguistic levels that might be easier accessible. For instance, the one sense per collocation assumption (Yarowsky, 1993) suggests that frequently co-occurring lexical strings tend to show a parallel co-occurrence relation on the concept level. Hence, given large concept annotated corpora such as Wikipedia, lexical units that co-occur with a concept can be identified and used for disambiguation (Kulkarni et al., 2009; Ratnov & Roth, 2011; Hoffart et al., 2012, inter alia). Recent work (e.g. (Cheng & Roth, 2013; Dalton & Dietz, 2013)) combines different linguistic levels including the discourse entity level (Section 10.3.1).

**Context Selection Function.** Given a linguistic level, a context selection function needs to be defined. In a pairwise model, we can define this function in the case of concept-level cohesive ties as

$$\psi_{pw}(c_i, c_j, d) = \tau_\psi(i, j, d) \cdot crel(c_i, c_j, d)$$

where  $crel(c_i, c_j, d)$  is a context-independent function that measures the relation between two concepts  $c_i$  and  $c_j$ .  $\tau_\psi(i, j, d)$  is a context-specific weight. In the discussion of different methods (Section 10.2), we have considered  $\psi_{pw}(c_i, c_j, d)$  as a black box. However, its inner structure is the heart of context modeling: it defines which context is selected and how it is weighted. This can affect the method. For instance, depending on how the context is selected, exact inference suddenly becomes possible (Cheng & Roth, 2013).

Analogously, we can define a pairwise function for the string-level cohesive ties as

$$\phi_{pw}(c_i, l_j, d) = \tau_\phi(i, j, d) \cdot lrel(c_i, l_j, d)$$

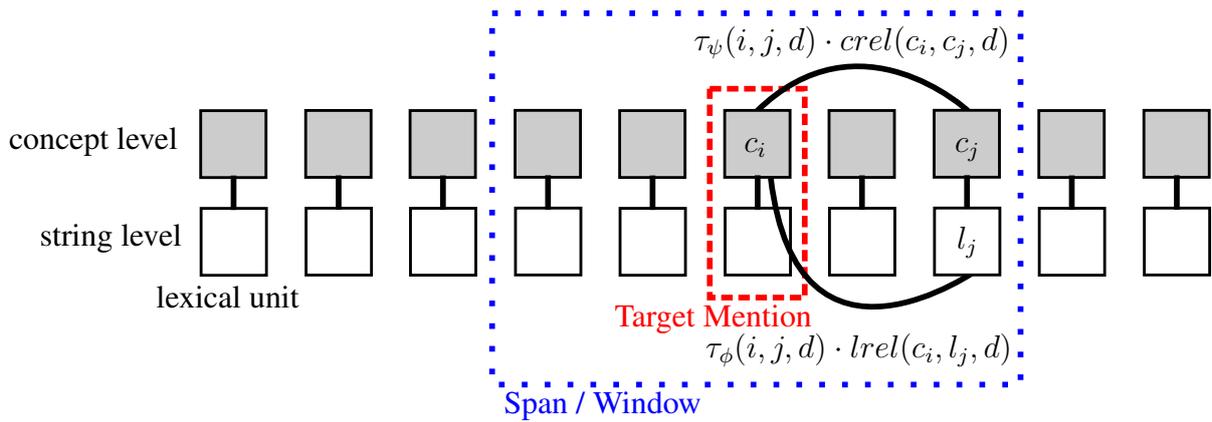


Figure 10.1: Factors that affect the modeling of the context.

where  $lrel(c_i, l_j, d)$  is a context-independent function that measures the relation between a concept  $c_i$  and a lexical unit  $l_j$  that occurs in its context.  $\tau_\phi(i, j, d)$  is a context-specific weight. Similar as for the global component, the local component  $\phi(c_i, m_i, d)$  can include different features that are formulated in a pairwise way via  $\phi_{pw}(c_i, l_j, d)$ .

**Span/Window.** The context span or window is the part of the text which is considered relevant to disambiguate or cluster a mention. It can comprise the whole text or only parts of it. The context window is determined by the context selection functions  $\tau_\psi(i, j, d)$  and  $\tau_\phi(i, j, d)$  which return 0 if a context concept or lexical unit lies outside this window.

Based on these context aspects we can refine our classification of approaches (Table 10.1). We distinguish between the following categories:

**Weighted vs. Unweighted Context Definition.** The context selection function can either contain a context-dependent weight (weighted context definition) or use a context-independent weight (unweighted context definition). Given the formulas  $\psi(c_i, c_j, d)$  and  $\phi(c_i, m_i, d)$  for the pairwise case, we can say that a pairwise approach has a weighted context definition if  $\tau_\psi(i, j, d)$  and  $\tau_\phi(i, j, d)$  are not always set to 1 and an unweighted context definition if  $\tau_\psi(i, j, d)$  and  $\tau_\phi(i, j, d)$  are always set to 1. In the latter case, the weight returned by  $\psi(c_i, c_j, d)$  and  $\phi(c_i, l_j, d)$  respectively only depends on the context-independent weights  $crel(c_i, c_j, d)$  and  $lrel(c_i, l_j, d)$ . The selection of  $\tau_\psi(i, j, d)$  and  $\tau_\phi(i, j, d)$  determines the span and reveals which context an approach considers as relevant. Approaches with a weighted context definition such as ours can be distinguished by the information they consider to determine  $\tau_\psi(i, j, d)$  and  $\tau_\phi(i, j, d)$ .

**Uniform vs. Binwise vs. Individual Context Weighting.** Approaches with a weighted context definition can further be classified by how many different weighting functions  $\tau_\psi(i, j, d)$

and  $\tau_\phi(i, j, d)$  they use. In Section 2.3, we identified three different strategies that we can also use to classify related work: One option is to use one single function  $\tau_\psi(i, j, d)$  and  $\tau_\phi(i, j, d)$  for each  $\psi$  and  $\phi$  respectively (*uniform context weighting*). For instance the context-independent score can be weighted by the importance of the context mention (Han et al., 2011). The second option is to use a different weighting function  $\tau_\psi(i, j, d)$  or  $\tau_\phi(i, j, d)$  for each single token to disambiguate (*individual context weighting*). However, such a strategy is not practical as it does not generalize. We therefore disregard this option. The third option is to group the tokens to disambiguate into different bins and use for each group of mention a different weighting function (*binwise context weighting*). Our approach is an example for such a binwise weighting.

**Predefined vs. Iterative vs. Joint Context Selection.** In the case a binwise context weighting strategy is chosen, we can distinguish between different strategies to select a context weighting function for a token to be disambiguated. The bin for a token to disambiguate can be determined before disambiguation (*predefined context selection*). An example of this strategy is our cascaded scope-aware approach (*ML Dis+NILs Scope (Casc)*, Table 9.17, Section 9.4.2). Another way is to iteratively select a context weighting function and disambiguate until a certain stopping criterion is met. The third option is to perform the identification of the context selection function for a token and its concept jointly. Our joint scope-aware approach is an example of a joint approach.

As our approach is global, we mainly focus on other global approaches that model concept-level cohesive ties (Section 10.3.1). However, the same classification schema also applies to approaches that model string level cohesive ties briefly discussed in Section 10.3.2. While we only consider the text-level context, Stoyanov et al. (2012) propose to explicitly model the whole communication context including e.g. the time or the target reader and to exploit it for the disambiguation. Their proposed model is preliminary and has not been implemented yet.

### 10.3.1 Concept-level Context Modeling

Table 10.2 extends Table 10.1 by the previously defined aspects related to the context definition (vertical dimensions). In the following, we discuss the context modeling of different approaches in more detail. Approaches we do not discuss here are similar to the other approaches in the same category.

**Approaches with an Unweighted Context Definition.** The assumption behind all approaches in this category (e.g. Kulkarni et al. (2009) or Ratinov & Roth (2011)) is that all mentions are equally important for their mutual disambiguation. The concepts of all mentions in the whole document are considered and the weight only depends on the context-independent

weight obtained for instance from the link structure in Wikipedia. Hence, these approaches assume that the concepts should be selected so that the cohesion in a document is maximal. None of these approaches achieve consistently higher performance if the concept-level cohesive ties are considered compared to a local approach.

**Approaches with a Uniformly Weighted Context Definition.** Most approaches use a uniformly defined weighting function  $\tau_\psi(i, j, d)$ .

Milne & Witten (2008b) is a prominent example in this category. They assume that the more a concept of a mention contributes to the main topics in a text, the more important it is to disambiguate other mentions. To model the prominence of a context concept  $c_j$ , i.e.  $\tau_\psi(i, j, d) = \tau_\psi(c_j)$  two factors are taken into account: (1) its average relatedness to all other candidate concepts and (2) the keyphraseness of its mention string measured by the probability that the mention is linked in Wikipedia. Hence, the final weight of  $\psi_{pw}(c_i, c_j, d)$  with  $c_i$  being a candidate concept of a mention to disambiguate and  $c_j$  being a context concept depends on the context independent relatedness score  $crel(c_i, c_j, d)$  and the prominence of  $c_j$  in the document.

Han et al. (2011) add three types of information into their graph: (1) the uncertainty of candidate concepts – measured by their respective probability given the local model –, (2) the tf-idf scores of the mentions reflecting their prominence, (3) the context-independent pairwise relatedness scores. All this information is propagated through the graph via a personalized PageRank algorithm. The final aggregated relatedness scores reflect all these three factors.

Ferragina & Scaiella (2012) define a sliding window consisting of 10 mentions. Given a target mention to disambiguate, only its concept relations within this window are considered, while all others are set to 0. This is mainly done to speed up the disambiguation process. However, it also implements the assumption that concept-level cohesive ties to closer mentions are more important than to distant mentions for disambiguation. Besides the distance, the  $\tau_\psi(i, j, d)$  also accounts for the uncertainty of a context candidate concept by taking the prior probability of it and the ambiguity of the respective mention into account.

			Non-Iterative				Iterative	
			Aggregation		Joint Inference		Aggregation	
			Unweighted Combination	Weighted Combination	Unweighted Combination	Weighted Combination	Unweighted Combination	Weighted Combination
			Pairwise Models					
Weighed	Binwise	Joint			Turdakov & Lizorkin (2009) Kulkarni et al. (2009)*	ML Dis+NILs+Clust Scope		
		Iterative			Cucerzan (2007)			
	Predefined					ML Dis+NILs+Clust Scope (Case) Hoffart et al. (2011b)		
	Uniform		Han et al. (2011) Guo et al. (2011) Moro et al. (2014b)	Ferragina & Scaiella (2012) Chen & Ji (2011)	ML Dis+NILs+Clust Cheng & Roth (2013) Dai et al. (2011)		Charton et al. (2014)	ML Dis+NILs Milne & Witten (2008b) Dalton & Dietz (2013)
Unweighted			Zhou et al. (2010a)		Kulkarni et al. (2009)		Ratinov et al. (2011)	

Table 10.2: Global approaches for disambiguation: they are classified based on their way to model interrelations between concepts and based on their context modeling strategy.

This information does not model any textual properties, but rather accommodates for the noise introduced by their non-iterative aggregative strategy.

In contrast to all approaches discussed so far, the following approaches use a binary function  $\tau_\psi(i, j, d)$ . Given a mention pair and a concept relation type, certain context-dependent conditions need to be met. If they are met the weight for this concept relation type is set to 1 and all corresponding concept relations between the two mentions are activated. Otherwise the weight is set to 0 and they are deactivated.

An example for such an approach is Cheng & Roth (2013). They use relation extraction techniques and coreference resolution to decide if the interrelations between two mentions should be activated. Only if two mentions match some relation extraction pattern, relations between their corresponding candidate concepts are considered. We can write this weighting function as

$$\tau_\psi(i, j, d) = I_\psi(m_i, m_j) \quad (10.10)$$

leading to

$$\psi_{pw}(c_i, c_j, d) = I_\psi(m_i, m_j) \cdot crel(c_i, c_j, d) \quad (10.11)$$

with  $I_\psi(m_i, m_j) = 1$  if the two mentions match an extraction pattern and 0 otherwise. The score  $crel(c_i, c_j, d)$  is the retrieval score – they index DBpedia triplets and links from Wikipedia with Lucene. Coreference clusters are identified in advance using string match heuristics. Hence, Cheng & Roth (2013) leverage correlations to cohesive ties on the discourse entity level and use them to weight the concept-level cohesive ties.

A similar approach is chosen by Dalton & Dietz (2013). They exploit correlations to cohesive ties on the lexical level to weight concept-level cohesive ties. In particular, they estimate the co-occurrence probability of the target and the context mention string in topically similar documents and use this value as  $\tau_\psi(i, j, d)$ . The intuition is that the more the target and a context mention co-occur in similar documents the more important they are for their mutual disambiguation. To calculate these co-occurrence weights, they first identify topically related documents that contain the target mention (or a variation of it). These similar documents are obtained by applying the same IR-based model that is used for disambiguation, but to another document collection than Wikipedia and by weighting the context mentions with their relative frequency in the target document. The co-occurrence score between a target mention  $m_i$  and a context mention  $m_j$  is then calculated by summing over all relative frequencies of the context mention  $m_j$  in this document collection. These relative frequencies are weighted by the relative retrieval score of the respective document to give higher weight to frequency scores from more similar documents, as in these documents it is more likely that the target and context mentions denote the same concepts as in the target document. The evaluation results on the

TAC testing data (2009-2012) show that weighting context mentions by co-occurrence scores across documents leads to significantly higher results than weighting them by their relative frequency in the target document in most data sets. The co-occurrence weighting based approach of Dalton & Dietz (2013) is orthogonal to our scope-aware approach. While the idea of using co-occurrence information is appealing, it is also expensive to identify additional documents that are similar for each mention to disambiguate. In contrast, we only exploit cross-document information within the document collection to disambiguate.

Han & Zhao (2009) and Han & Zhao (2010) are two clustering approaches that exploit a weighted concept-level context representation. They first disambiguate the words in the context of a mention to cluster – they do not disambiguate the target mention – and then derive a concept-based vector representation similar to the one we use for cross-document clustering features. The concepts are weighted by considering their relatedness to other concepts in the context (Han & Zhao, 2009) and – in Han & Zhao (2010) – in the context of the mention to compare to.

**Approaches with a Binwise Context Definition.** Our scope-aware approach is a binwise approach. Other binwise approaches are thus the closest related approaches to our work with respect to context modeling (Table 10.2). In word sense disambiguation, local binwise approaches are the rule. Each word type forms a bin and is disambiguated with a separate classifier (Navigli, 2009). The selection of the bin thus does not depend on the discourse structure as in our case and can be selected before disambiguation. Dhillon & Ungar (2009) also train a separate classifier for each word type, but exploit information from other words with similar senses for feature selection. Only a few global approaches that use Wikipedia as an inventory are binwise.

One example is Hoffart et al. (2011b). They distinguish between two main bins: local and global. Mentions that are part of the local bin are disambiguated locally and only the resulting concept is added to the graph. The mentions of the global bin are disambiguated by considering the global component. For each mention, it is decided to which bin it belongs before the graph-based disambiguation. This decision depends on the local feature values for all candidate concepts of a mention. They calculate the difference between the two used local feature scores for all candidate concepts of a mention. If this difference is higher than a certain threshold, i.e. if they contradict each other, the global component is used to disambiguate it. Otherwise it is disambiguated locally. This strategy leads to a significant improvement over the method in which the global component is always activated. In their test data, two thirds of the instances are disambiguated locally. Hoffart et al. (2011b) explain that a permanent activation of the global component for all mentions can be problematic if not all mentions belong to the same coherence cluster. However, in contrast to us their model does not include discourse information to decide which features to activate to disambiguate a mention.

Cucerzan (2007) proposes an iterative binwise approach. He uses the following three different binary weighting functions that all use distance information and iteratively apply them:

$$\tau_{\psi}(i, j, d) = \tau_{\psi\_doc}(m_i, m_j) = \begin{cases} 1 & \text{if } m_i \text{ and } m_j \text{ are in the same document} \\ 0 & \text{otherwise} \end{cases}$$

$$\tau_{\psi}(i, j, d) = \tau_{\psi\_paragraph}(m_i, m_j) = \begin{cases} 1 & \text{if } m_i \text{ and } m_j \text{ are in the same paragraph} \\ 0 & \text{otherwise} \end{cases}$$

$$\tau_{\psi}(i, j, d) = \tau_{\psi\_sentence}(m_i, m_j) = \begin{cases} 1 & \text{if } m_i \text{ and } m_j \text{ are in the same sentence} \\ 0 & \text{otherwise.} \end{cases}$$

In particular, first  $\tau_{\psi\_doc}(m_i, m_j)$  is applied to disambiguate a mention. If more than one candidate concept of this mention has a score higher than a certain threshold,  $\tau_{\psi\_paragraph}(m_i, m_j)$  is used for its disambiguation. If still more than one candidate has a higher score than this threshold, the  $\tau_{\psi\_sentence}(m_i, m_j)$  is applied and the disambiguation is repeated. Cucerzan (2007) calls this strategy *context shrinking*.

The HMM-based model with multiple lexical chains of Turdakov & Lizorkin (2009) is a joint binwise approach. Each lexical chain is a bin. To select the relevant lexical chains, the mention needs to be disambiguated and vice versa. However, as most approaches in lexical chaining, they solely rely on relatedness measures to decide which chain a mention should join. However, as also wrong candidate concepts between mentions can be related (Section 2.2), this might not be sufficient. Kulkarni et al. (2009) briefly discuss an approach similar to the one of Turdakov & Lizorkin (2009) where mentions are clustered according to their concept-level cohesive ties and disambiguated at the same time. They report that this did not significantly improve the results. In contrast to our approach, they only rely on concept-level cohesive ties, but do not take into account other features to model the embedding of mentions into the discourse structure.

**Discussion.** Comparing different global approaches based on their weighting function for concept-level cohesive ties helps to analyze their implemented context models. We identify two current trends: while earlier work such as Kulkarni et al. (2009) or Ratinov & Roth (2011) are unweighted approaches and consider all mentions as equally important for their mutual disambiguation, more recent approaches are weighted (Table 10.2). The trend goes towards a weighting of concept-level cohesive ties by exploiting correlations in cohesive ties

on other linguistic levels (Cheng & Roth, 2013; Dalton & Dietz, 2013). Our approach is in line with this development and also exploits such correlations.

What sets our approach apart from others is that we use different weighting functions. The selection of the weighting function and the disambiguation is performed jointly. In contrast to e.g. Turdakov & Lizorkin (2009), we do not rely on concept-level cohesive ties to decide on the weighting function, but select it based on the embedding of the mention into discourse. This embedding is modeled as a separate classification task combining different features.

### 10.3.2 String-level Context Modeling

To classify approaches with respect to their context definitions on the string level, we can use the same scheme as for the concept level. However, as the weighting strategies are comparable and usually do not go beyond a distance- (Csomai & Mihalcea, 2008; Dalton & Dietz, 2013; Sen, 2012), frequency- (Kulkarni et al., 2009; Ratinov & Roth, 2011) or co-occurrence-based weighting (Hoffart et al., 2012), we rather focus on three outstanding approaches.

Jin et al. (2014) propose an iterative context selection approach with string-level features. Given a mention to disambiguate, they iteratively select the most discriminative feature. As soon as the probability of a candidate concept given the selected features is higher than a certain threshold, it is returned as the disambiguation result. This iterative selection of lexical features improves the result of an approach where all lexical features are taken into account and is similar to the approach of Yarowsky (1995) based on decision lists.

The generative approach for concept disambiguation of Han & Sun (2012) models a mutual reinforcement between the concepts and the string-level context. Each mention is assigned a topic. Which topic is selected depends on the probability that this topic appears in the document – and thus on the assignments of the topic to the other mentions (concept-level context) – and on the probability of the denoted concept given this topic. The assignment of a concept to a mention in turn depends on the probability of the concept given the selected topic, the probability of the mention string given the concept and on the context tokens that are assigned to this concept (string-level context). More precisely, for each token in the document a target concept is selected. Which target concept is selected depends on how often it has been assigned to a mention in the document and on the probability of the token string being in the context of it. Hence, the concept-level context influences the string-level context and vice versa: the more often a concept is selected for a mention, the more likely it is that a token is assigned this concept. The more often a concept is assigned to a context token, the more likely it is that it is selected for a mention. This proposed approach improves a window-based approach, which does not model this reinforcement, by 2% F1 on the IITB data set. This approach is in so far similar to ours as it also assumes that the selection of the context and the disambiguation influence each other. However, in contrast to us, they do not model any

information related to the text structure and would obtain the same results, if the mentions were scrambled.

In contrast to Han & Sun (2012), Sen (2012) exploits distance information in its generative approach for person name disambiguation. Similar to Han & Sun (2012), they estimate a distribution for each token that measures how likely it is that the token is associated with a concept. This distribution takes into account distance information. To estimate these distributions, they associate a token with 50% probability to the same concept as its previous token and with 50% probability to a concept according to its own distribution. In another experiment, they assign a token with 33.3% probability each to a concept that appears in the same sentence, in the same paragraph or in the same document respectively. Hence, the closer a concept, the more likely a token is associated with it. Their results are rather low. The second distribution estimation strategy improves the first one by more than 10% on Wikipedia data.

In concept clustering, the lexical units in the context are often not directly used. Instead, their respective co-occurrences – smoothed using singular value decomposition – are employed (Schütze, 1998). While the former are called first-order co-occurrences, the latter are second-order co-occurrences. Purandare & Pedersen (2004) compare such a second-order context representation to a first-order representation and conclude that the former performs better in most cases. In contrast to first order co-occurrences, second order co-occurrences are less sparse and also account for similar or related words. Véronis (2004) goes further and also exploits co-occurrences of co-occurrences, i.e.  $n$ -order co-occurrences. Most clustering approaches use a unified context weighting strategy. An exception is the binwise approach of Popescu (2010) and similar Chen et al. (2012). Popescu (2010) extracts all occurrences to cluster – they all have the same string – and reweights the context, so that discriminative lexical units obtain more weights. In Popescu (2009), the context definition depends on the ambiguity of a proper name. The idea is that it needs more context to cluster ambiguous proper names than unambiguous ones.

**Discussion.** In general, the tendency is to use the whole document as context for concept-level features and a smaller window size for string-level features. Most approaches use a static window size, although of different size. For instance, Csomai & Mihalcea (2008) use a three-tokens, Bunescu & Paşca (2006) a 55-tokens window size and Han & Sun (2011) a 50-tokens window size for concept disambiguation with Wikipedia. Gale et al. (1992d) analyze different window-sizes and show that the broader context is informative for the disambiguation of nouns. The approaches discussed in more details define the string-level context in a more dynamic way. Our approach defines the context still window-based, but weight different window sizes differently dependent on the embedding of the mention into the discourse.

## 10.4 Modeled Information

Many features have been proposed in the field of disambiguation. Instead of describing these features directly, we discuss in the following the different underlying linguistic phenomena that they describe. As these phenomena are often unobserved, they need to be approximated. Hence, for each modeled linguistic phenomenon, we show different approaches to implement it. Different approximations require different resources. This section thus also sheds light on the resources used by different approaches.

We first discuss common linguistic information modeled in concept disambiguation and NIL detection (Section 10.4.1), before we review the modeled information in concept clustering (Section 10.4.2).

### 10.4.1 Modeled Information for Concept Disambiguation and Recognition of NILs

In the following, we focus on information modeled by disambiguation approaches that use Wikipedia as an inventory. This overview is not comprehensive, but highlights the information used by most approaches. Agirre & Stevenson (2006) discuss the modeled information – they call it knowledge sources – that has been useful in disambiguation with for instance WordNet.

By reviewing previous work for concept disambiguation and recognition of NILs, we identify three main linguistic phenomena that have been considered as useful and thus have been modeled: the prominence of a concept, the type of a concept and co-occurrence and domain information. Table 10.3 summarizes different approximations of these three phenomena, lists the respective required resources and mentions some work that uses these approximations.

**Prominence of a Concept.** Some concepts occur more frequently than others. This information can be exploited for disambiguation by biasing the disambiguation results more towards frequent concepts. For instance, Cucerzan (2007) implicitly models the prominence of a concept by its description length assuming that more prominent concepts are more extensively described. While this method only requires concept descriptions, an annotated corpus is needed to estimate the relative frequency of concepts (Dredze et al., 2010; Ratinov et al., 2011). However, Hoffart et al. (2011b) conclude that the prior probability, i.e. the prominence of a concept given a mention, leads to better results. The prior probability is a strong feature and has been used since Gale et al. (1992a) as a hard-to-beat baseline. To overcome the need for an annotated corpus to estimate the prior probability, McCarthy et al. (2007) have proposed an unsupervised method to estimate the predominant sense of a word. Measuring the string distance between the canonical concept realization and the mention string is another way to model the mention string dependent prominence of a concept (Dredze et al., 2010;

Name	Description	Required Resources	Used in
<b>Prominence of a Concept</b>			
<b>Prior probability</b>	Measures the probability of a concept $c$ given a mention estimated via $p(c m) = \frac{\text{count}(c,m)}{\sum_{c'} \text{count}(c',m)}$ with $\text{count}(c, m)$ being the number of times mention $m$ denotes concept $c$ in an annotated corpus	Large annotated corpus	Kulkarni et al. (2009), Milne & Witten (2008b), Turdakov & Lizorkin (2009), Ratinov et al. (2011), Hoffart et al. (2011b), Ferragina & Scaiella (2012)
<b>Topic-specific prior</b>	A topic- or group-specific prior of a concept in generative approaches	Large (annotated) corpus	Han & Sun (2012), Sen (2012)
<b>Description length</b>	The length of the description is taken as an indicator for its prominence; this measure is independent of a mention string	Inventory with concept descriptions	Cucerzan (2007) (implicitly), Dredze et al. (2010)
<b>Relative occurrence</b>	Concepts that appear more often in more texts in a corpus, are more probable; this measure is independent of a mention string	Large annotated corpus	Dredze et al. (2010), Ratinov et al. (2011)
<b>Search engine</b>	Rank of concept (i.e. Wikipedia page) in a search engine query with the mention string	Inventory indexed by a search engine; search engine	Dredze et al. (2010), Fader et al. (2009)
<b>Type of Concept</b>			
<b>Named entity type</b>	Named entity type obtained from a named entity recognizer	Named entity recognizer; inventory with type information	Dredze et al. (2010), Zhang et al. (2010)
<b>Appositions</b>	Cosine similarity between apposition and description of concept	Inventory with concept descriptions	Cheng & Roth (2013)
<b>Syntax-based features</b>	Approximation of selectional restrictions with local context words, surrounding verbs or statistics which concepts are for instance typical subjects of a verb	Large (annotated) corpus	Hoffart et al. (2011b), Csomai & Mihalcea (2008)
<b>Co-occurrence and Domain Information</b>			
<b>String-level similarity</b>	Similarity between a string-level concept representation and a string-level context representation: concept representation is based on e.g. keyphrases (Hoffart et al., 2011b; 2012), tokens in first paragraph in Wikipedia (Kulkarni et al., 2009) or whole page (Kulkarni et al., 2009), values in infoboxes (Dredze et al., 2010), frequently co-occurring tokens (Ratinov & Roth, 2011; Kulkarni et al., 2009); text representation is for instance based on the surrounding words or keyphrases in the whole text; common similarity measures are the cosine similarity or the dot product	Large annotated corpus or inventory with extensive concept descriptions	Kulkarni et al. (2009), Hoffart et al. (2011b), Bunescu & Paşca (2006), Zhou et al. (2010a), Ratinov et al. (2011), Zhang et al. (2010), Dredze et al. (2010), Dai et al. (2011), Fader et al. (2009), Zheng et al. (2010), Han et al. (2011)
<b>Concept co-occurrence score</b>	Relatedness between two concepts based on shared incoming links in Wikipedia; as two concepts share an inlink if they co-occur in the same document, this is a cooccurrence score; the most popular measure is the one from Milne & Witten (2008a): $\text{crel}(c_i, c_j) = \frac{\log(\max( \text{in}(c_i) ,  \text{in}(c_j) )) - \log( \text{in}(c_i) \cap \text{in}(c_j) ))}{\log( C ) - \log(\min( \text{in}(c_i) ,  \text{in}(c_j) ))}$	Large annotated corpus	Milne & Witten (2008b), Kulkarni et al. (2009), Turdakov & Lizorkin (2009), Hoffart et al. (2011b), Ratinov et al. (2011), Ferragina & Scaiella (2012), Han et al. (2011)
<b>Learned score</b>	Learned combination of different measures	Large annotated corpus	Ceccarelli et al. (2013)
<b>Kore</b>	Relatedness between two concepts based on a keyphrase representation of the concepts	Inventory with concept description	Hoffart et al. (2012), Hoffart et al. (2014)
<b>Concept signature</b>	Relatedness based on random walk through concept network	Highly connected concept network	Moro et al. (2014b)
<b>Category-based relatedness</b>	Relatedness based on category or type information	Inventory with type or category information	Cucerzan (2007), Zhou et al. (2010a)
<b>Reader-based relatedness</b>	Concepts (i.e. Wikipedia articles) that have been opened in the same browsing session	User data	Zhou et al. (2010a)

Table 10.3: Commonly modeled information in concept disambiguation and NIL recognition.

Zhou et al., 2010a). The assumption is that the more distant the string is, the less likely it is that it denotes the concept. In contrast to the prior probability this is rather a negative indicator to exclude candidate concepts and to detect NILs. Another common assumption is that the domain constrains concepts (Madhu & Lytel, 1965; Agirre & Stevenson, 2006), which for instance Han & Sun (2012) approximate with topic-dependent priors of concepts.

**Concept Type.** The type of a concept such as e.g. PERSON or MACHINE mainly helps to reduce the probability of candidate concepts that fail to match the selectional restrictions of the governing verb, noun or adjective of a mention. In word sense disambiguation with WordNet, this information is helpful (Agirre & Stevenson, 2006), while in disambiguation with Wikipedia it has been less helpful. For instance Hoffart et al. (2011b) show that despite they generalized proper name concepts to overcome data sparsity, this information is not helpful in proper name disambiguation. One reason could be that this information is indiscriminative if many candidates are of the same type which is often the case in Wikipedia.

**Co-occurrence and Domain Information.** Co-occurrence and domain information can be difficult to separate from each other and thus are often modeled together. Two concepts that show high co-occurrence scores tend to belong to the same domain and vice versa. Both kinds of information can be approximated on different linguistic levels (e.g. string-level co-occurrences vs. concept-level co-occurrences) using different context selection strategies (e.g. window-based weighting strategy, exploitation of correlated co-occurrences on other linguistic levels). As we already discussed these aspects (Section 10.3), we only describe here different functions for  $crel(c_i, c_j)$  and  $lrel(m_i, m_j)$  (see also Fernández García et al. (2014)). These functions mainly define if rather co-occurrence or domain information is modeled. For instance, the most common chosen function for  $crel(c_i, c_j)$  is the relatedness measure of Milne & Witten (2008a). It is based on the inlinks in Wikipedia. As two concepts share an inlink in Wikipedia if they occur together in an article, this measure accounts mainly for concept co-occurrences, but at least partially also models domain information. This measure has shown good performance (e.g. Milne & Witten (2008b) and Hoffart et al. (2011b)) and is also used in our approach. It leads to better results than for instance an inlink-based cosine similarity (Ratinov et al., 2011). However, it requires an annotated corpus and is less accurate for rare concepts as only a few inlinks are available for them. Hoffart et al. (2012) propose another relatedness measure *kore* that shifts the modeling from the concept to the string level and compares two concepts based on their shared keyphrases instead. In contrast to concept-level co-occurrences, co-occurring keyphrases are easier to identify, but also noisier (Hoffart et al., 2014). Ceccarelli et al. (2013) combine different measures into one score in a supervised way – similar to Fahrni et al. (2011b) – with promising results. By replacing the inlink-based relatedness measure of Milne & Witten (2008a) they obtain 5%

improvement in precision at a recall of 10% *ceteris paribus*. In contrast to inlink-based measures, measures that build upon the category structure of Wikipedia and thus rather model domain than co-occurrence information have shown to be less effective (Kulkarni et al., 2009; Ratinov et al., 2011). Cucerzan (2007) nevertheless obtains high results with such a domain-based relatedness measure. He also uses information extracted from list pages and enumerations. On the string-level, the most common measures are the cosine similarity between a text representation vector and a concept representation vector and the dot product (e.g. Bunescu & Paşca (2006), Kulkarni et al. (2009) and Ratinov & Roth (2011)).

**Discussion.** Almost all state-of-the-art systems model the prominence of a concept and co-occurrence information. As both aspects are hidden, often many features are used to approximate them. As most other state-of-the-art approaches, we combine information from all three categories. In addition, we also exploit clustering information as for instance Cheng & Roth (2013) or Charton et al. (2014). These aspects are discussed in the next section on concept clustering.

Our discussion only gives a broad overview, but shows that almost all approximations including the approximations in this thesis require a large annotated corpus. Hence, it is often more important to consider the features to decide how much supervision a system requires than to screen how many annotated data has been used to tune the parameters. For instance, many knowledge-based approaches that rely on the internal hyperlink structure in the English Wikipedia, need as much annotated data – at least for English, it might be different in a multilingual setting –, as a system that uses the whole Wikipedia to tune its parameters. As it is expensive to manually annotate a corpus, several approaches have been proposed to (semi-)automatically annotate corpora with concepts or senses. For instance, one technique is to identify text snippets with unambiguous words and use them as examples for ambiguous words that denote the same concepts (Leacock et al., 1998; Mihalcea & Moldovan, 1999; Agirre & Martínez, 2004). This method is only applicable if the assumption that the inventory is complete holds, which is already critical for common nouns and even more for proper names. Other common approaches to automatically annotate data are bootstrapping (Yarowsky, 1995; Mihalcea, 2002) or the use of translation equivalents in parallel texts (e.g. Gale et al. (1992b), Diab & Resnik (2002), Ng et al. (2003), Diab (2004), Section 10.5). As we use Wikipedia as an inventory, the data bottleneck is less critical than in disambiguation with e.g. WordNet due to the huge amount of internal hyperlinks (Mihalcea, 2007) and is not addressed in this thesis. However, as for instance Hoffart et al. (2014) discuss, not for all Wikipedia concepts sufficient annotated data is available. To obtain high performance for such concepts at the long tail, it is thus indispensable to overcome this bottleneck using unsupervised approaches.

### 10.4.2 Modeled Information for Concept Clustering

In concept clustering, roughly the same information is modeled as in concept disambiguation, although with less additional resources.

**Mention String Compatibility.** String similarities between mentions are extensively used in concept clustering. Similar to our approach, two mentions are only clustered if they show sufficient similarity with respect to their strings. Cross-document clustering approaches often rely on the one-sense-per-discourse hypothesis (Gale et al., 1992c) stating that within the documents all mentions with the same or similar string denote the same concept (e.g. Gooi & Allen (2004)).

**Prominence.** In concept clustering, the prominence is hardly ever modeled. One way to consider prominence is described in Baron & Freedman (2008). They check if a mention string is unambiguous in Wikipedia. The assumption is that in this case, the respective occurrences must have a predominant concept and therefore cluster together. In our case, prominence is modeled via disambiguation features.

**Concept Type.** Fleischman & Hovy (2002) assume that each mention is assigned with a fine-grained semantic type such as e.g. *LAWYER*. Given two mentions, they measure how compatible their respective types are. For instance a *POLITICAN* is often a *LAWYER*. They also leverage modifier compatibility.

**Co-occurrence and Domain Information.** Context-based features are the most important features for concept clustering. Compared to concept disambiguation, factual information such as e.g. birth date or occupations identified via information extraction techniques are by far more common (Mann & Yarowsky, 2003; Lefever et al., 2007, inter alia). However, Chen & Martin (2007) indicate that such information is rather rare.

## 10.5 Multi- and Cross-linguality

Since Wikipedia has been used as an inventory, its potential for multilingual disambiguation has been highlighted (e.g. Cucerzan (2007)). While most work in disambiguation and clustering still focuses on English, multi- and cross-lingual tasks have gained popularity over the years. Since the nineties, i.e. long before multi- and cross-lingual aspects have been addressed in the context of Wikipedia, they had been a topic in lexical semantics (Brown et al., 1991; Gale et al., 1992b; Diab & Resnik, 2002, inter alia).

As our approach indicates, multi- and especially cross-linguality comes in different flavors. Not only the task itself can be multi- or cross-lingual, but also a monolingual approach can for instance exploit multi- and cross-lingual features (Chapter 7). In this section, we discuss the relation of this thesis to other work that addresses multi- and cross-linguality in the context of disambiguation and clustering. In particular, we focus on three aspects, i.e. multi- and cross-lingual task definitions (Section 10.5.1), multi- and cross-lingual strategies (Section 10.5.2) and multi- and cross-lingual features (Section 10.5.3).

### 10.5.1 Multi- and Cross-lingual Tasks

In disambiguation and clustering, multiple multi- and cross-lingual tasks exist. In fact, the motivation to address a multi- and cross-lingual task varies between different works leading to different task definitions. In the following, we show how the multi- and cross-lingual task definitions used in this thesis relate to other variants. The relations are summarized in Table 10.4. As we focus in this section on multi- and cross-lingual aspects, no monolingual tasks are listed in Table 10.4 (left upper cell **(1)**). The tasks that are highlighted in bold are at least partially based on Wikipedia.

**Multilingual Disambiguation and Clustering.** Multilingual disambiguation and clustering tasks do not show much variance. Commonly, a system that has been developed for one language is ported to at least one other language (e.g. Khapra et al. (2009)). The ported system then performs disambiguation and clustering within this language. The aim of multilingual approaches is to balance between costs (e.g. in terms of time and annotations) and quality, which is especially challenging for languages with only a few resources. Khapra et al. (2010) propose an approach to reach an optimal quality cost ratio. In this thesis, we proposed different variants of our systems that require different degrees of adaptations.

**Cross-lingual Disambiguation and Clustering.** In contrast to purely multilingual tasks, cross-lingual tasks have been proposed in many more variants. We distinguish between cross-lingual tasks that are not multilingual – i.e. that consider one single source language – and cross-lingual tasks that are multilingual and address multiple source languages (Section 1.1.3). Moreover, cross-lingual disambiguation approaches have been studied for different purposes, as discussed in the following.

*Cross-lingual representation.* In this thesis, we perform cross-lingual concept disambiguation and clustering to obtain a common representation of texts in different languages. Such a representation allows for instance to extract information from different languages and is also the main goal of the entity linking task at TAC (Ji & Grishman, 2011; Ji et al., 2011). The assumption is that concepts are at least to a certain degree shared across languages. The cross-

lingual link discovery task at NTCIR slightly deviates from this purpose, but still preserves the idea that one can point from a mention in one language to a predefined concept in another language.

*Translation-based inventory.* In lexical semantics, cross-lingual disambiguation has been studied from completely different perspectives since the early nineties. With the paper of Brown et al. (1991) that proposes to distinguish between different concepts or senses of a word based on translational equivalences, the idea of using a cross-linguistically motivated inventory has been born and given rise to cross-lingual studies (Resnik & Yarowsky, 1999; Ide, 2000; Chugur et al., 2002; Dyvik, 2004). Assuming that different meanings are realized differently across languages, the hope is to determine the granularity of an inventory in a data-driven way. In the meantime, several cross-lingual shared tasks have been organized that define the different concepts or senses of words via translational equivalences (Ide et al., 2002; Chklovski et al., 2004; Jin et al., 2007; Lefever & Hoste, 2010; 2013). However, although translational equivalences are useful to calculate relatedness between the denoted senses or concepts of a word (Resnik & Yarowsky, 1999; Chugur et al., 2002), many questions are still unclear. For instance, it is still open how many language pairs should be considered, which language families should be combined (Resnik & Yarowsky, 1999; Ide, 2000) and if for instance also metaphoric meaning differences can be captured (Chugur et al., 2002). Compared to disambiguation with WordNet or Wikipedia, approaches based on translation equivalencies do not capture variability or only to a certain extent. Such an approach is also not applicable for proper names, as in most cases the ambiguity is preserved across languages.

*Translation-based acquisition of annotated data.* Starting from the same assumption as the translation-based approaches, i.e. that different meanings of an ambiguous word are lexicalized differently across languages, it has been proposed to automatically acquire training data from parallel data for a given inventory such as WordNet (e.g. Diab & Resnik (2002), Zhong & Ng (2009)). In contrast to translation-based approaches, such approaches map translational equivalences either automatically (Zhong & Ng, 2009) or semi-automatically (Ng et al., 2003) to a predefined inventory. Hence, the disambiguation itself is not cross-lingual, but only the acquisition of training data is cross-lingual.

### 10.5.2 Multi- and Cross-lingual Strategies

To adapt a monolingual system for multi- and cross-lingual concept disambiguation and clustering, different strategies are possible (Chapter 7). In this section, we classify different related approaches based on their strategy and show their relation to our approach. We focus here only on the disambiguation and clustering and not on the preprocessing, which may require additional adaptations. Table 10.5 summarizes disambiguation approaches.

	Not Cross-lingual	Cross-lingual Representation	Cross-lingual Translation-based Inventory	Translation-based Data Acquisition
<b>Not Multilingual</b>	(1):	(2): <b>Cross-lingual Link Discovery Task at NTCIR-9 (Tang et al., 2011)</b>	Gale et al. (1992d)  Lexical Sample Tasks at Senseval and SemEval with Translational Equivalences (e.g. Chklovski et al. (2004), Jin et al. (2007), Lefever & Hoste (2010), Lefever & Hoste (2013))	Diab & Resnik (2002) Ng et al. (2003) Zhong & Ng (2009)
<b>Multilingual</b>	(3): Khapra et al. (2009) Khapra et al. (2010) Khapra et al. (2011)  <b>Multilingual Word Sense Disambiguation at SemEval-2013 (Navigli et al., 2013)</b>	(4): <b>Cross-lingual Linking of Person Names (Mayfield et al., 2011)</b> <b>Cross-lingual Entity Linking Task at TAC 2011, 2012, 2013 (Ji &amp; Grishman, 2011; Ji et al., 2011)</b> <b>Cross-lingual Link Discovery Task at NTCIR-10 (Tang et al., 2013; 2014)</b>	Brown et al. (1991) Li & Li (2002) Dagan & Itai (1994)	Tufis et al. (2004)

Table 10.4: Different multi- and cross-lingual tasks. The tasks highlighted in bold use at least partially Wikipedia as an inventory.

**Multilingual Approaches.** The most common multilingual approaches are adaptive and language-independent. In concept disambiguation, graph-based approaches that exploit the structure of the inventory and do not require any training data to tune the parameters are popular (Moro et al., 2014b; Gutiérrez et al., 2013). Assuming that the structure in the inventory is similar (Khapra et al., 2009) or even the same across languages as in BabelNet (e.g. Moro et al. (2014b)), such approaches can be considered as language-independent, in the sense that they do not require any adaptations. More critical are features that are derived from corpora that involve linguistic realizations. For instance, prior probabilities of a concept given a linguistic realization or lexical similarity features require language-specific annotated data. Khapra et al. (2009) propose an approach to project such information from a resource-rich to other resource-poor languages. Assuming not only that the corresponding concepts in the two languages are interlinked, but also the realizations within the synsets, they project the prior probabilities across languages. While the performance is in most cases lower than if an annotated corpus in the respective language is used, the costs are much lower. In later work, they propose an approach to balance between quality and costs (Khapra et al., 2010). Another approach they

		Not Cross-lingual	Cross-lingual		
		(1):	Target-sided	Source-sided	Cross-lingual
Not Multilingual			(2): <i>Kim &amp; Gurevych (2011)</i> <i>Kim &amp; Gurevych (2013)</i>	Guo & Diab (2010) <i>Tang et al. (2012b)</i> <i>Navigli &amp; Ponzetto (2012a)</i>	Silberer & Ponzetto (2010) van Gompel (2010) Lefever et al. (2011) Carpuat (2013) Gale et al. (1992d) <i>Knoth et al. (2011b)</i> , <i>Knoth et al. (2011a)</i>
Multilingual	<b>Translation-based</b>	(3):	(4): Guo et al. (2012) <i>Tang et al. (2012a)</i>		
	<b>Adaptation-based</b>	Khapra et al. (2010) Khapra et al. (2011)	<i>McNamee et al. (2011)</i> <i>Mayfield et al. (2011)</i>	<i>Monahan et al. (2011)</i> <i>Monahan &amp; Carpenter (2012)</i> <i>Knoth &amp; Herrmannova (2013)</i>	Brown et al. (1991) Li & Li (2002) <i>Chang et al. (2011)</i> <i>Spitkovsky &amp; Chang (2011)</i> Dagan & Itai (1994) <b>ML Dis+NILs+Clust</b>
	<b>Language-independent</b>	Khapra et al. (2009) <i>Gutiérrez et al. (2013)</i> <i>Moro et al. (2014b)</i>	Tufis et al. (2004) Zhang et al. (2013) <i>Cassidy et al. (2011)</i>		

Table 10.5: Comparison of different multi- and cross-lingual disambiguation approaches based on their strategies. The variants of our proposed approaches are bolded. Approaches that also use Wikipedia or a derived resource as an inventory are in italics.

propose assumes that some annotated data is available in both languages. By performing bilingual bootstrapping and projecting the parameters across languages, they achieve better results than with monolingual bootstrapping (Khapra et al., 2011). The results of Khapra et al. (2009), Khapra et al. (2010), Khapra et al. (2011) are in line with others. Already good results can be achieved without any adaptations that can be improved by injecting language-specific information. In concept clustering, Pedersen et al. (2006) show that their approach based on second-order co-occurrences scales well across different languages without major adaptations. While the topic-model based approach of Kozareva & Ravi (2011) also scales across languages, Rozovskaya & Sproat (2007) obtain lower results for Russian, a morphologically-rich language, than for English. They conclude that for such languages, lemmatization is essential.

**Cross-lingual Approaches.** Target-based cross-lingual approaches first translate or transliterate text parts into the target language and then select a concept or cluster for the mentions in the target language. Such approaches are common in cross-lingual link discovery (Kim & Gurevych, 2011; 2013, inter alia), but also at TAC (McNamee et al., 2011). Although McNamee et al. (2011) disambiguate in the target language, they train a separate model for each language pair and thus require adaptations. However, they also show results in which the same model is used for languages with the same script without a big drop in performance. Source-sided cross-lingual approaches disambiguate on the source side and then map the results to the concept or word in the target language. Such approaches have lead to top performing

results in different shared tasks (Guo et al., 2013b; Knoth & Herrmannova, 2013). Closely related to a source-sided strategy are cross-lingual strategies. Usually, the only difference is that the translation step is not an additional step, but integrated in the disambiguation. A common cross-lingual strategy that is also used in this thesis is to use a multilingual resource that makes the translation step obsolete (Spitkovsky & Chang, 2011).

In cross-lingual concept clustering, a shared representation between the languages becomes even more important, as only in this way context-based features can be used across languages. The most common approach is to translate the context (Clarke et al., 2012; Merhav et al., 2013; Monahan & Carpenter, 2012). Green et al. (2012) compare a machine translation based approach to polylingual topic models obtained from Wikipedia, i.e. a comparable corpus. Their evaluation for cross-lingual English and Arabic cross-document coreference resolution indicates that a context representation based on polylingual topics lead to lower results than using machine translation. However, while the polylingual topic models can be obtained from comparable corpora, a statistical machine translation system requires parallel corpora which are rarer.

### 10.5.3 Multi- and Cross-lingual Features

Independent from whether a system addresses a mono-, multi- or cross-lingual task, its features can be monolingual, i.e. derived from one language, or cross-lingual, i.e. derive from one or multiple other languages (Chapter 7). In the following, we discuss different cross-lingual features that have been exploited by related work.

**Concept-level Co-occurrence Features.** Concept-level co-occurrences are strong features (Section 10.4). In order to extract them, concept-annotated data is necessary. However, if the concepts are mapped across languages as in our case, we can extract co-occurrences across languages.

**Translational Equivalences of Mentions.** Starting from the common assumption that ambiguities of words are lexicalized differently across languages, Navigli & Ponzetto (2012b) propose to consider the translations of words to disambiguate in different languages. Using the multilingual resource BabelNet, they first obtain all candidate concepts for a word to disambiguate and then all translations are obtained from BabelNet. For each translation and the original word string a separate ranking of the candidate concepts is obtained via a graph-based approach. The final disambiguation decision aggregates all these rankings. Both in a monolingual and a cross-lingual task, the cross-lingual features improve the results *ceteris paribus*.

**Lexical Co-occurrences in Different Languages.** Banea & Mihalcea (2011) and Dandala et al. (2013) extend the idea of using translational equivalences to the whole context and translate it using a machine translation system to several languages. Lexical features are then extracted from the original and all translations. Compared to a system that only uses monolingual features, the results are higher. While the idea of exploiting multilingual contexts is appealing, it is disputable if several machine translation systems should be applied to perform disambiguation given the assumption that machine translation might be one of the applications of disambiguation. We therefore do not exploit cross-lingual lexical features, but stick to concept-level features extracted from different languages.

#### 10.5.4 Discussion

Concept disambiguation and clustering is an unsolved problem and one could argue that one should first solve it for one language, e.g. English, before moving on to other languages or perform it across languages. However, as discussed in this section, looking at the problem from a multi- and cross-lingual perspective reveals new opportunities and helps to reach a better understanding of the problem. Work in lexical semantics mainly focuses on translational equivalences to obtain a better understanding of concepts and their boundaries (e.g. Resnik & Yarowsky (1999), Ide (2000)) and as a source to cheaply acquire training data (Diab & Resnik, 2002; Ng et al., 2003; Zhong & Ng, 2009). Recently, large-scale multilingual inventories, e.g. derived from Wikipedia, open a new additional perspective. Assuming that concepts are shared across languages, they allow to link texts in different languages to a common representation. Such a representation allows for new cross-lingual applications such as e.g. cross-lingual information extraction (Ji & Grishman, 2011; Ji et al., 2011). In addition, thanks to its multilinguality, Wikipedia allows to study how information can be shared across languages. This thesis contributes to this line of research by studying the effect of sharing concepts co-occurrences across languages.

### 10.6 Summary

In this chapter, we discussed how the proposed approach relates to other work by focusing on five aspects that are essential for this thesis: the task definitions, the methods, the context definitions, the modeled information and multi- and cross-linguality. Table 9.2 in Section 9.1.4 summarizes the closest approaches to which we directly compared to in the experiment section.

# Chapter 11

## Conclusions and Future Work

We started this thesis with a quote from Weaver (1955) in which the question is asked how big the context window – or with Weaver (1955) how big “the slit in the opaque mask” – needs to be that a word can be disambiguated. This thesis of course is far from answering this question, but it at least contributes towards a better context understanding. According to the insights gained in this thesis, different mentions require different context definitions dependent on their embedding into discourse and their denoted concepts. Hence, one opaque mask is not sufficient, but different opaque masks with different slits are required.

In the following, we summarize the contributions of this thesis (Section 11.1), discuss some limitations of the proposed approach (Section 11.2) and propose some future research directions (Section 11.3).

### 11.1 Contributions

In the introduction, we asked four research questions related to (1) the interrelations between disambiguation and clustering, (2) the context definition, (3) the modeling of the linguistic insights using machine learning approaches and (4) the multi- and cross-lingual scalability. In the following, we summarize the contribution of this thesis to each of these questions.

#### 11.1.1 Joint Concept Disambiguation and Clustering

**Question 1:** What is the relation between concept disambiguation and clustering? Can concept clustering be used to improve concept disambiguation and vice versa?

**Linguistic Analysis.** In this thesis, we first analyzed the relationship between concept disambiguation and concept clustering from a linguistic perspective and concluded that the two

tasks mutually depend on each other. If two mentions are in the same cluster, they must denote the same concept and if two mentions denote the same concept, they must be in the same cluster. Hence, concept disambiguation can benefit from concept clustering and vice versa. More specifically, if it is known to the disambiguation component that two mentions are in the same cluster, it can exploit the context of both mentions. In particular, if one mention in the clustering relation is easier to disambiguate than the other mention (Figure 2.9), clustering is beneficial. On the other hand, if the clustering component knows that two mentions denote different concepts, they are not supposed to be clustered and if two mentions denote the same concept, they must be clustered. This also affects the clustering of the NILs, as if one mention is a NIL according to the disambiguation component and the other mentions denotes a concept in the inventory, the two should not be clustered. However, it is important to notice that concept disambiguation and clustering are only mutually dependent given the assumption that the granularity of the concept clusters corresponds to the granularity of the concepts in the inventory.

**Method.** We proposed a joint approach for concept disambiguation and clustering in Markov logic. Markov logic combines first order logic with probabilities and allows to formalize and model interdependencies concisely. The interdependencies between concept disambiguation and concept clustering are modeled via hard constraints. To our knowledge, we are the first to perform concept disambiguation and clustering jointly in one single step.

**Results.** Our experiments show that the clustering significantly improves the disambiguation and vice versa. We compared our joint disambiguation and clustering approach in Markov logic to a cascaded disambiguation and clustering approach that use exactly the same features. Our joint approach leads consistently to higher results than the cascaded approach on different data sets. These findings are in line with the results obtained by other approaches. All top performing systems in the recent editions of the TAC entity linking task exploit clustering information during the disambiguation, although using a cascaded approach.

### 11.1.2 Discourse-aware Concept Disambiguation and Clustering

**Question 2:** What is the relevant context to disambiguate or cluster a mention? Which factors determine which context is relevant to disambiguate or cluster a mention?

**Linguistic Analysis.** The context definition is one of the most determining factors in concept disambiguation and clustering. In this thesis, we investigated the textual context from a discourse perspective. We argue that the context relevant to disambiguate a mention consists of the syntactic context and the mentions with which it shares concept-level cohesive

ties. However, identifying concept-level cohesive ties is challenging, as they depend on the denoted concepts of the mentions. Hence, resolving ambiguities requires that the concept-level cohesive ties are known, but to identify concept-level cohesive ties, the mentions should already be disambiguated. We can summarize these insights as follows: different mentions require different notions of contexts. Which context is relevant to disambiguate or cluster a mention depends on its concept-level cohesive ties or – more generally – on its embedding into discourse and its denoted concept.

**Method.** To account for the fact that different mentions require different context definitions, we propose a latent variable model for discourse-aware concept disambiguation using Markov logic. Each mention is assigned a cohesive scope. Dependent on its cohesive scope, the mention is disambiguated differently. We distinguish three broad categories of cohesive scopes: (1) mentions with local cohesive scope exhibit cohesive ties within the same sentence; (2) mentions with intermediate cohesive scope show cohesive ties both within the sentence and beyond; (3) mentions with global cohesive scope form cohesive ties with mentions across sentence boundaries. The scope assignment is performed jointly with the disambiguation and clustering. This has the advantage that the learning of the weights for the scope assignment task is guided by the annotations available for the target prediction task (i.e. the disambiguation). Hence, while annotated data are required for the concept disambiguation and clustering task, no annotated data are necessary for the scope assignment task. We thus model the scopes as latent variables and adapt the approach proposed by Poon & Domingos (2008) in the context of coreference resolution to learn the weights.

**Results.** We showed that our scope-aware system significantly improves a scope-unaware system *ceteris paribus*. In particular, we observed gains for mentions that are assigned local or global scope. Hence, using different models for different mentions significantly improves the results.

### 11.1.3 Modeling Interdependencies with Markov Logic

**Question 3:** How can we model concept disambiguation and clustering in accordance to our linguistic findings using state-of-the-art machine learning techniques?

From a machine learning perspective, all implications for the modeling of concept disambiguation and clustering we draw from our linguistic analysis deal with label interdependencies. We have mostly summarized how we model these interdependencies in the context of the previous two contributions (Section 11.1.1 and 11.1.2). In the following, we therefore only summarize the differences between the modeling of these interdependencies.

We identified three different types of interdependencies: (1) interdependencies between concepts of different mentions, (2) interdependencies between concept disambiguation and clustering, (3) interdependencies between the context relevant to disambiguate a mention and its concept. Ideally, we would model all these interdependencies jointly. However, this is not tractable. We thus use two main approximations: context bins instead of individual concept definitions and an aggregative approach to account for cohesive ties of type relatedness. Our main contribution in terms of modeling is thus the identification of a balance between joint and approximate techniques, so that we have a tractable model that still reflects our linguistic insights.

#### 11.1.4 High Scalability across Languages

**Question 4:** How can we port our system to other languages than English? Can we share information across languages to boost the performance?

To decide how we can port our system to other languages than English, we first analyze which parts of the system are likely to scale across languages and which parts may require language-specific adaptations. This analysis revealed that the core of our approach (Chapter 4) is language-independent, while the features may require some language-specific adaptations. The features used in our joint concept disambiguation and clustering approach are likely to scale across languages and only need few adaptations. All language-specific information necessary to calculate these features can be extracted from the Wikipedia version in the respective language. However, the features for the scope assignment task are much more language-dependent, as they make use of language-specific information (e.g. regarding the sentence structure). To port these features many adaptations are required and may even other features are necessary for languages other than English. Our multi- and cross-lingual system thus corresponds to the English system without scope information.

Our results for cross-lingual Spanish and Chinese concept disambiguation and clustering are comparable to our English monolingual results which indicates that our joint scope-unaware concept disambiguation and clustering system highly scales across languages requiring only a few adaptations. It achieves state-of-the-art results for Spanish and Chinese, even without retraining. Sharing concept-level co-occurrence information across languages leads to significantly higher results for Spanish and Chinese, although the gains for Spanish are higher. This difference might be due to the fact that the coverage of the cross-lingual concept mapping is 10% higher (absolute) for Spanish than for Chinese.

## 11.2 Limitations

Our proposed approach has several limitations and can be extended and improved in different ways. In this section, we name three main limitations and discuss how they can be addressed.

### 11.2.1 Modeling of NILs

As the results in Section 9.3 show, the performance on the NILs is low. One of the reasons might be that we only model them using negative evidence. We only model that a mention does not denote a certain known concept. We lack any features indicating that a mention should denote another concept which is not part of the inventory. Recently, some approaches have been proposed that go into this direction and actually use positive evidence to recognize NILs (Han et al., 2011; Hoffart et al., 2014). However, their performance is still low and more research is required.

### 11.2.2 Restriction to Common Nouns and Proper Names

Our current approach is restricted to common nouns and proper names. To extend our approach for words of other part-of-speech tags, an additional inventory would be necessary, as Wikipedia does not cover concepts for e.g. verbs and adjectives. Moro et al. (2014b) use BabelNet which combines Wikipedia with WordNet. It is questionable if the combination of Wikipedia and WordNet is a good choice, as their design and thus their granularity do not match (Chapter 3). Given our concept definition, Wiktionary might be a better choice as an additional inventory, as it is built collaboratively as Wikipedia. However, Wiktionary is not designed in a concept-centric way. We thus argue that more investigations are required to identify a suitable inventory for words of other parts of speech than nouns.

### 11.2.3 Speed and Scalability

We evaluated our approach on relatively small data sets consisting of a few hundred documents. To scale up our approach to thousands or even millions of documents, further research is required. The main part that slows down its speed is the cross-document clustering. The bigger the document collection to process is, the bigger the candidate clusters become and the more mentions need to be processed together. Hence, a strategy is required to decouple big mention groups into smaller groups and first process these smaller groups, before these groups are clustered. Research in this direction has been done by Singh et al. (2011).

## 11.3 Future Research Directions

While in the last section we discussed the main limitations of our current approach and suggested how they can be addressed, we indicate in this section broader future research directions.

### 11.3.1 Evaluation Settings and Metrics

Although the evaluation settings and metrics are quite standardized, they still need to be improved. In particular, it is not clear how to evaluate an end-to-end system. Following previous work (Resnik & Yarowsky, 1999; Ratinov et al., 2011), we considered the annotated mentions as a representative sample of all mentions and only evaluate on the annotated mentions. It needs to be investigated if there is a better way to evaluate an end-to-end system without requiring human annotators to annotate all mentions in a text, as this may lead to a low inter-annotator agreement (Resnik & Yarowsky, 1999).

In addition, the evaluation of clustering results based on a gold standard is an open research problem. Current clustering metrics such as  $B^3$  show some irregularities. It is not clear how strongly these irregularities bias the overall evaluation results.

An extrinsic evaluation of our approach within a specific application such as summarization or textual entailment could help to better obtain more insights in the performance of our system.

### 11.3.2 A Multi-layer Model for Concept Disambiguation and Clustering

The proposed scope-aware approach has the potential to be expanded in two dimensions. Currently, our approach consists of a target prediction task (i.e. the disambiguation and clustering) and a supportive classification task (scopes). However, the scopes may itself depend on another supportive classification task, which we could model as an additional latent layer. For instance, if a mention is a relational noun such as e.g. *rest* (Löbner, 1985), it is more likely to be of local scope, as if it is a sortal noun. The distinction between sortal and relational nouns can be considered by itself as a classification task that could support the scope assignment task. In addition, other supportive classification tasks may exist that influence the disambiguation and clustering task directly. For instance, the domain of a text or the communication context (see next section) could be modeled in this way. Figure 11.1 illustrates these different extensions. On the top left, our current proposed approach is sketched. The supportive classification task is illustrated by an empty circle, while the target prediction task is illustrated by a filled rectangle. On the top right, multiple supportive task on the same level are depicted that directly influence the target prediction task and vice versa. On the bottom (left-hand side)

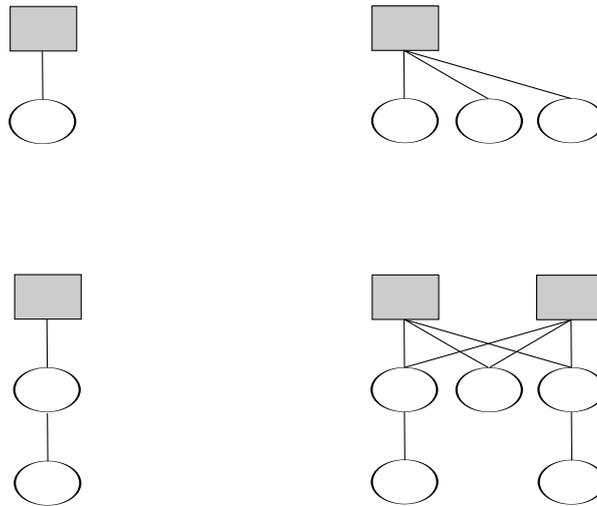


Figure 11.1: Towards a multi-layer model for concept disambiguation and clustering. The filled rectangle illustrates the target prediction task (e.g. the concept disambiguation and clustering task), while the empty circles represent supportive classification tasks (e.g. the scope assignment task). On the top left our proposed approach is represented with one supportive classification task. On the top left and the bottom, different extensions are sketched.

the supportive classification task is itself supported by a supportive classification task. On the bottom (right-hand side), the two extensions are combined.

### 11.3.3 Beyond the Textual Context

In this thesis, we focused on the textual context. However, the textual context is only one part of the overall communication context. A text is a piece of communication and involves a whole communication situation. It would be interesting to model concept disambiguation and clustering by taking into account aspects of the whole communication situation. For instance, the time the text has been written (Fang & Chang, 2014), the source where it has been published and the net it builds together with other texts could be exploited for disambiguation and clustering. Stoyanov et al. (2012) proposes some approach in this direction, but this proposed approach is only partially implemented and needs further investigations.



# Bibliography

- Abel, F., Gao, Q., Houben, G.-J., & Tao, K. (2011). Semantic enrichment of twitter posts for user profile construction on the social web. In *Proceedings of the 8th Extended Semantic Web Conference*, Heraklion, Crete, Greece, 29 May – 2 June 2011, pages 375–389.
- Agirre, E., Ansa, O., Martinez, D., & Hovy, E. (2000). Enriching very large ontologies with topic signatures. In *Proceedings of the 14th European Conference on Artificial Intelligence*, Berlin, Germany, 20–25 August 2000.
- Agirre, E., Arregi, X., & Otegi, A. (2010a). Document expansion based on WordNet for robust IR. In *Proceedings of Coling 2010: Poster Volume*, Beijing, China, 23–27 August 2010, pages 9–17.
- Agirre, E. & Edmonds, P. G. (Eds.) (2006). *Word Sense Disambiguation: Algorithms and Applications*. Heidelberg, Germany: Springer.
- Agirre, E., López de Lacalle, O., Fellbaum, C., Hsieh, S.-K., Tesconi, M., Monachini, M., Vossen, P., & Segers, R. (2010b). SemEval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2)*, Uppsala, Sweden, 15–16 July 2010, pages 75–80.
- Agirre, E. & Martínez, D. (2004). Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 25–26 July 2004, pages 25–32.
- Agirre, E., Martínez, D., de Lacalle, O. L., & Soroa, A. (2006). Two graph-based algorithms for state-of-the-art WSD. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 22–23 July 2006, pages 585–593.
- Agirre, E. & Soroa, A. (2007). SemEval-2007 Task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-1)*, Prague, Czech Republic, 23–24 June 2007, pages 7–12.

- Agirre, E. & Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, 30 March – 3 April 2009, pages 33–41.
- Agirre, E. & Stevenson, M. (2006). Knowledge sources for WSD. In Agirre, E. & Edmonds, P. (Eds.), *Word Sense Disambiguation: Algorithms and Applications*, pages 217–251. Heidelberg, Germany: Springer.
- Andrews, N., Eisner, J., & Dredze, M. (2014). Robust entity clustering via phylogenetic inference. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Md., 22–27 June 2014, pages 775–785.
- Artiles, J., Borthwick, A., Gonzalo, J., Sekine, S., & Amigó, E. (2010). WePS-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks. In *CLEF 2010 LABs and Workshops, Notebook Papers*, Padua Italy, 22-23 September 2010.
- Artiles, J., Gonzalo, J., & Sekine, S. (2007). The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-1)*, Prague, Czech Republic, 23–24 June 2007, pages 64–69.
- Artiles, J., Gonzalo, J., & Sekine, S. (2009). WePS 2 evaluation campaign: Overview of the web people search clustering task. In *Proceedings of the 2nd Web People Search Evaluation Workshop (WePS 2009)*, Madrid, Spain, 21 April 2009.
- Artiles, J., Sekine, S., & Gonzalo, J. (2008). Web people search: Results of the first evaluation and the plan for the second. In *Proceedings of the 17th World Wide Web Conference*, Beijing, China, 21–25 April, 2008, pages 1071–1072.
- Asher, N. & Lascarides, A. (1995). Lexical disambiguation in a discourse context. *Journal of Semantics*, 12(1):69–108.
- Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a Web of open data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, Busan, Korea, 11-15 November 2007, pages 722–735.
- Bagga, A. & Baldwin, B. (1998a). Algorithms for scoring coreference chains. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, Granada, Spain, 28–30 May 1998, pages 563–566.

- Bagga, A. & Baldwin, B. (1998b). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montréal, Québec, Canada, 10–14 August 1998, pages 79–85.
- Bakir, G. H., Hofmann, T., Schölkopf, B., Smola, A. J., Taskar, B., & Vishwanathan, S. V. N. (2007). *Predicting Structured Data*. Cambridge, Massachusetts, London, United Kingdom: MIT Press (Neural Information Processing).
- Banea, C. & Mihalcea, R. (2011). Word sense disambiguation with multilingual features. In *Proceedings of the 9th International Conference on Computational Semantics*, Oxford, U.K., 12–14 January 2011, pages 25–34.
- Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 6–12 January 2007, pages 2670–2676.
- Bar-Hillel, Y. (1960). Automatic translation of languages. In Booth, D. & Meagher, R. (Eds.), *Advances of Computers*. New York, N.Y.: Academic Press.
- Baron, A. & Freedman, M. (2008). Who is Who and What is What: Experiments in cross-document co-reference. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 274–283.
- Barzilay, R. & Elhadad, M. (1997). Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent and Scalable Text Summarization, Madrid, Spain, July 1997*, pages 10–17.
- Barzilay, R. & Elhadad, M. (1999). Using lexical chains for text summarization. In Mani, I. & Maybury, M. T. (Eds.), *Advances in Automatic Text Summarization*, pages 111–121. Cambridge, Mass.: MIT Press.
- Barzilay, R. & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology. Special Issue on Interaction of Linguistics and Computational Linguistics*, pages 1–26.

- Bentivogli, L. & Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor corpus. *Natural Language Engineering*, 11(3):247–261.
- Bentivogli, L., Clark, P., Dagan, I., & Giampiccolo, D. (2011). The seventh PASCAL recognizing textual entailment challenge. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.
- Bentivogli, L., Forner, P., Giuliano, C., Marchetti, A., Pianta, E., & Tymoshenko, K. (2010). Extending English ACE 2005 corpus annotation with ground-truth links to Wikipedia. In *Proceedings of the 2nd Workshop on The People's Web: Collaboratively Constructed Semantic Resources*, Beijing, China, 28 August 2010, pages 19–27.
- Bergsma, S. & Lin, D. (2006). Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 17–21 July 2006, pages 33–40.
- Bernhard, D. & Gurevych, I. (2009). Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing*, Singapore, 2–7 August 2009, pages 728–736.
- Bhagwani, S., Satapathy, S., & Karnick, H. (2013). Merging word senses. In *Proceedings of TextGraphs-8: Graph-based Algorithms for Natural Language Processing, Workshop at EMNLP 2013*, Seattle, Wash., 18 October 2013, pages 11–19.
- Blunsom, P., Cohn, T., & Osborne, M. (2008). A discriminative latent variable model for statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio, 15–20 June 2008, pages 200–208.
- Bögel, T. & Frank, A. (2013). A joint inference architecture for global coreference clustering with anaphoricity. In Gurevych, I., Biemann, C., & Zesch, T. (Eds.), *Language Processing and Knowledge in the Web*, pages 35–46. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science, 8105).
- Bronner, A., Negri, M., Mehdad, Y., Fahrni, A., & Monz, C. (2012). CoSyne: Synchronizing multilingual wiki content. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, Linz, Austria, 27–29 August 2012, pages 33:1–33:4.

- Brown, P. F., Della Pietra, V. J., & Mercer, R. L. (1991). Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, Cal., 18–21 June 1991, pages 264–270.
- Bruce, R. & Wiebe, J. (1994). Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, N.M., 27–30 June 1994, pages 139–146.
- Bryl, V., Giuliano, C., Serafini, L., & Tymoshenko, K. (2010). Supporting natural language processing with background knowledge: Coreference resolution case. In *Proceedings of the 9th International Semantic Web Conference, Revised Selected Papers, Part I*, Shanghai, China, 7-11 November 2010, pages 80–95.
- Buitelaar, P., Magnini, B., Strapparava, C., & Vossen, P. (2006). Domain-specific wsd. In Agirre, E. & Edmonds, P. (Eds.), *Word Sense Disambiguation*, pages 275–298. Heidelberg, Germany: Springer.
- Bunescu, R. & Paşca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 3–7 April 2006, pages 9–16.
- Bußmann, H. (2002). *Lexikon der Sprachwissenschaft*. Stuttgart: Kröner.
- Cai, J. & Strube, M. (2010). End-to-end coreference resolution via hypergraph partitioning. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 23–27 August 2010, pages 143–151.
- Cai, J. F., Lee, W. S., & Teh, Y. W. (2007). NUS-ML: Improving word sense disambiguation using topic features. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-1)*, Prague, Czech Republic, 23–24 June 2007, pages 249–252.
- Cano, A. E., Rizzo, G., Varga, A., Rowe, M., Stankovic Milan, & Dadzie, A.-S. (2014). Making sense of microposts (#Microposts2014) named entity extraction & linking challenge. In *Proceedings of the 23rd World Wide Web Conference, Making Sense of Microposts'14*, Seoul, Korea, 7–11 April, 2011.
- Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., & Li, H. (2007). Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, Oreg., 20–24 June 2007, pages 129–136.
- Carpuat, M. (2013). NRC: A machine translation approach to cross-lingual word sense disambiguation (SemEval-2013 Task 10). In *Proceedings of STARSEM 2013: The 2nd Joint*

- Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, 14–15 June 2013, pages 188–192.
- Cassidy, T., Chen, Z., Artiles, J., Ji, H., Deng, H., Ratinov, L.-A., Zheng, J., Han, J., & Roth, D. (2011). CUNY-UIUC-SRI TAC-KBP 2011 entity linking system description. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.
- Cassidy, T., Ji, H., Ratinov, L.-A., Zubiaga, A., & Huang, H. (2012). Analysis and enhancement of Wikification for microblogs with context expansion. In *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India, 8–15 December 2012, pages 441–456.
- Castillo, J. J. (2011). A WordNet-based semantic approach to textual entailment and cross-lingual textual entailment. *International Journal of Machine Learning and Cybernetics*, 2(3):177–189.
- Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., & Trani, S. (2013). Learning relatedness measures for entity linking. In *Proceedings of the ACM 22nd Conference on Information and Knowledge Management (CIKM 2013)*, San Francisco, California, USA, 26–30 October 2010, pages 139–148.
- Chakrabarti, S., Dom, B., & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, Seattle, Washington, USA, 1–4 June 1998, pages 307–318.
- Chan, Y. S., Ng, H. T., & Zhong, Z. (2007). NUS-PT: Exploiting parallel texts for word sense disambiguation in the English all-words tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-1)*, Prague, Czech Republic, 23–24 June 2007, pages 253–256.
- Chang, A. X., Spitkovsky, V. I., Agirre, E., & Manning, C. D. (2011). Stanford-UBC entity linking at TAC-KBP, again. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.
- Chang, M.-W., Hsu, B.-J., Ma, H., Loynd, R., & Wang, K. (2014). E2E: An end-to-end entity linking system for short and noisy text. In *Proceedings of the 4th Workshop on Making Sense of Microposts*, Seoul, Korea, 7 April 2014.

- Chang, M.-W., Ratinov, L., & Roth, D. (2012). Structured learning with constrained conditional models. *Machine Learning*, 88(3):399–431.
- Charton, E., Meurs, M.-J., Jean-Louis, L., & Gagnon, M. (2014). Mutual disambiguation for entity linking. In *Proceedings of the ACL 2014 Conference Short Papers*, Baltimore, Md., 22–27 June 2014, pages 476–481.
- Che, W. & Liu, T. (2010). Jointly modeling WSD and SRL with Markov logic. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 23–27 August 2010, pages 161–169.
- Chen, L., Feng, Y., Zou, L., & Zhao, D. (2012). Explore person specific evidence in web person name disambiguation. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, pages 832–842.
- Chen, Y. & Martin, J. (2007). Towards robust unsupervised personal name disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning*, Prague, Czech Republic, 28–30 June 2007, pages 190–198.
- Chen, Z. & Ji, H. (2011). Collaborative ranking: A case study on entity linking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, U.K., 27–29 July 2011, pages 771–781.
- Cheng, X., Chen, B., Samdani, R., Chang, K.-W., Fei, Z., Sammons, M., Wieting, J., Subhro Roy, S., Chizheng Wang, C., & Roth, D. (2014). Illinois Cognitive Computation Group UI-CCG TAC 2013 entity linking and slot filler validation systems. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 18–19 November 2013.
- Cheng, X. & Roth, D. (2013). Relational inference for Wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Wash., 18–21 October 2013, pages 1787–1796.
- Chiang, D. & Bikel, D. M. (2002). Recovering latent information in treebanks. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, 24 August – 1 September 2002, pages 183–189.
- Chklovski, T., Mihalcea, R., Pedersen, T., & Purandare, A. (2004). The SensEval-3 multilingual English-Hindi lexical sample task. In *Proceedings of the 3rd International Workshop*

- on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3) at ACL-04*, Barcelona, Spain, 25–26 July 2004, pages 5–8.
- Chlovski, T. & Mihalcea, R. (2002). Building a sense tagged corpus with Open Mind Word Expert. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, Penn., July 2002, pages 116–122.
- Chu-Carroll, J. & Fan, J. (2011). Leveraging Wikipedia characteristics for search and candidate generation in question answering. In *Proceedings of the 25th Conference on the Advancement of Artificial Intelligence*, San Francisco, Cal., 7–11 August 2011, pages 872–877.
- Chugur, I., Gonzalo, J., & Verdejo, F. (2002). Polysemy and sense proximity in the Senseval-2 test suite. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, Penn., July 2002, pages 32–39.
- Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
- Cilibrasi, R. L. & Vitányi, P. M. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383.
- Cimiano, P., Schultz, A., Sizov, S., Sorg, P., & Staab, S. (2009). Explicit vs. latent concept models for cross-language information retrieval. In *Proceedings of the 21th International Joint Conference on Artificial Intelligence*, Pasadena, Cal., 14–17 July 2009, pages 1513–1518.
- Clarke, J., Merhav, Y., Suleiman, G., Zheng, S., & Murgatroyd, D. (2012). Basis Technology at TAC 2012 entity linking. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 5–6 November 2012.
- Collins, M. (2002). Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, Penn., 6–7 July 2002.
- Cook, P., Lau, J. H., Rundell, M., McCarthy, D., & Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word senses. In *Proceedings of eLex*, Tallinn, Estonia, 17–19 October 2013.
- Cornolti, M., Ferragina, P., & Ciaramita, M. (2013). A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd World Wide Web Conference*, Rio de Janeiro, Brazil, 28 March – 1 April, 2013, pages 249–260.

- Coursey, K. & Mihalcea, R. (2009). Topic identification using Wikipedia graph centrality. In *Companion Volume to the Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Col., 31 May – 5 June 2009, pages 117–120.
- Cowie, J., Guthrie, J., & Guthrie, L. (1992). Lexical disambiguation using simulated annealing. In *Proceedings of the 15th International Conference on Computational Linguistics*, Nantes, France, 23–28 August 1992, pages 359–365.
- Cox, C. (2005). Assessing the utility of ResearchCyc in recognizing textual entailment. Technical report, Stanford University.
- Csomai, A. & Mihalcea, R. (2008). Linking documents to encyclopedic knowledge. *IEEE Intelligent Systems*, 23(5):34–41.
- Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning*, Prague, Czech Republic, 28–30 June 2007, pages 708–716.
- Cucerzan, S. (2012). TAC entity linking by performing full-document entity extraction and disambiguation. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.
- Cucerzan, S. (2013). The MSR system for entity linking at TAC 2012. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 5–6 November 2012.
- Cucerzan, S. & Sil, A. (2014). The MSR systems for entity linking and temporal slot filling at TAC 2013. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 18–19 November 2013.
- Cumming, S. (2008). Discourse content. In Sherman, A. B. . B. (Ed.), *Metasemantics*. Oxford University Press.
- Dagan, I. & Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.
- Dai, H.-J., Tsai, R. T.-H., & Hsu, W.-L. (2011). Entity disambiguation using a Markov-Logic network. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, 8–13 November 2011, pages 846–855.

- Dalton, J. & Dietz, L. (2013). A neighborhood relevance model for entity linking. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, Lisbon, Portugal, 22–25 August 2013, pages 149–156.
- Dalton, J. & Dietz, L. (2014). UMass CIIR at TAC KBP 2013 entity linking: query expansion using Urban Dictionary. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 18–19 November 2013.
- Dandala, B., Mihalcea, R., & Bunescu, R. (2013). Multilingual word sense disambiguation using Wikipedia. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, Nagoya, Japan, 14–18 October 2013, pages 498–506.
- Daneš, F. (1974). Functional sentence perspective and the organization of the text. In Daneš, F. (Ed.), *Papers on Functional Sentence Perspective*, pages 106–128. Prague: Academia.
- Dantzig, G., Fulkerson, R., & Johnson, S. (1954). Solution of a large-scale traveling-salesman problem. *Operations Research*, 2:393–410.
- de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 22–28 May 2006, pages 449–454.
- de Melo, G. & Weikum, G. (2009). Towards a universal Wordnet by learning from combined evidence. In *Proceedings of the ACM 18th Conference on Information and Knowledge Management (CIKM 2009)*, Hong Kong, China, 2–6 November 2009, pages 513–522.
- De Vault, D., Oved, I., & Stone, M. (2006). Societal grounding is essential to meaningful language use. In *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, Mass., 16–20 July 2006, pages 747–754.
- Decadt, B., Hoste, V., Daelemans, W., & van den Bosch, A. (2004). GAMBL, genetic algorithm optimization of memory-based WSD. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3) at ACL-04*, Barcelona, Spain, 25–26 July 2004, pages 108–112.
- Demetriou, G. C. (1993). Lexical disambiguation using constraint handling in Prolog (CHIP). In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands, 21–23 April 1993, pages 431–436.
- Devitt, M. & Sterelny, K. (1999). *Language and Reality: An Introduction to the Philosophy of Language*. Cambridge, Mass.: MIT Press (A Bradford book).

- Dhillon, P. S. & Ungar, L. H. (2009). Transfer learning, feature selection and word sense disambiguation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Singapore, 2–7 August 2009, pages 257–260.
- Diab, M. (2004). Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 21–26 July 2004, pages 303–310.
- Diab, M. & Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Penn., 7–12 July 2002, pages 255–262.
- Dietz, L. & Dalton, J. (2013). Across-document neighborhood expansion: UMass at TAC KBP 2012 entity linking. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 5–6 November 2012.
- Do, Q. X., Lu, W., & Roth, D. (2012). Joint inference for event timeline construction. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, pages 677–687.
- Domingos, P. & Lowd, D. (2009). *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan Claypool Publishers.
- Domingos, P. & Richardson, M. (2007). Markov Logic: A unifying framework for statistical relational learning. In Getoor, L. & Taskar, B. (Eds.), *Introduction to Statistical Relational Learning*, pages 339–371. Cambridge, Mass.: MIT Press.
- Dredze, M., McNamee, P., Rao, D., Gerber, A., & Finin, T. (2010). Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 23–27 August 2010, pages 277–285.
- Drenner, S., Harper, M., Frankowski, D., Riedl, J., & Terveen, L. (2006). Insert movie reference here: A system to bridge conversation and item-oriented web sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada, 22–27 April 2006, pages 951–954.
- Dyvik, H. (2004). Translations as semantic mirrors: From parallel corpus to WordNet. *Language and Computers*, 49:311–326.
- Eco, U. (2002). *Einführung in die Semiotik*. Autorisierte dt. Ausgabe, 9., unveränderte Auflage von Jürgen Trabant. München: Wilhelm Fink Verlag (UTB für Wissenschaft: Uni-Taschenbücher; 105).

- Edmonds, P. & Cotton, S. (2001). SENSEVAL-2: Overview. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2) at ACL-01*, Toulouse, France, 5–6 July 2001, pages 1–5.
- Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using Explicit Semantic Analysis. *ACM Transactions on Information Systems*, 29:8:1–34.
- Ellis, J., Getman, J., Mott, J., Li, X., Griffitt, K., Strassel, S. M., & Wright, J. (2013). Linguistic resources for 2013 knowledge base population evaluations. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 18–19 November 2013.
- Ellis, J., Li, X., Griffitt, K., Strassel, S. M., & Wright, J. (2012). Linguistic resources for 2012 knowledge base population evaluations. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 5–6 November 2012.
- Erk, K. (2006). Unknown word sense detection as outlier detection. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, New York, N.Y., 4–9 June 2006, pages 128–135.
- Fader, A., Soderland, S., & Etzioni, O. (2009). Scaling Wikipedia-based named entity disambiguation to arbitrary web text. In *Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at IJCAI-09*, Pasadena, Cal., July 2009.
- Fahrni, A., Göckel, T., & Strube, M. (2013). HITS' monolingual and cross-lingual entity linking system at TAC 2012: A joint approach. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 5–6 November 2012.
- Fahrni, A., Heinzerling, B., Göckel, T., & Strube, M. (2014). HITS' monolingual and cross-lingual entity linking system at TAC 2013. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 18–19 November 2013.
- Fahrni, A., Nastase, V., & Strube, M. (2011a). Deliverable 4.1: Hyperlink identification. Technical report, HITS gGmbH.
- Fahrni, A., Nastase, V., & Strube, M. (2011b). HITS' graph-based system at the NTCIR-9 cross-lingual link discovery task. In *Proceedings of the 9th NTCIR Workshop Meeting*, Tokyo, Japan, 6-9 December 2011, pages 473–480.

- Fahrni, A., Nastase, V., & Strube, M. (2012). HITS' cross-lingual entity linking system at TAC 2011: One model for all languages. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.
- Fahrni, A. & Strube, M. (2012). Jointly disambiguating and clustering concepts and entities with Markov logic. In *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India, 8–15 December 2012, pages 815–832.
- Fahrni, A. & Strube, M. (2014). A latent variable model for discourse-aware concept and entity disambiguation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 26–30 April 2014, pages 491–500.
- Fang, Y. & Chang, M.-W. (2014). Entity linking in microblogs with spatial and temporal signals. *Transactions of the Association for Computational Linguistics*, 2:259–272.
- Fellbaum, C. (2012). WordNet. In *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd.
- Fellegi, I. P. & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210.
- Fernández García, N., Arias Fisteus, J., & Sánchez Fernández, L. (2014). Comparative evaluation of link-based approaches for candidate ranking in link-to-Wikipedia systems. *Journal of Artificial Intelligence Research*, 49:733–773.
- Ferragina, P. & Scaiella, U. (2012). Fast and accurate annotation of short texts with Wikipedia pages. *IEEE Software*, 29(1):70–75.
- Ferrández, S., Toral, A., Ferrández, Ó., Ferrández, A., & Muñoz, R. (2007). Applying Wikipedia's multilingual knowledge to cross-lingual question answering. In *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems*, Paris, France, 27–29 June 2007, pages 352–363.
- Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *SIGMOD Record*, 41(2):15–26.
- Finin, T., Syed, Z., Mayfield, J., McNamee, P., & Piatko, C. (2009). Using Wikitology for cross-document entity coreference resolution. In *AAAI 2009 Spring Symposium on Learning by Reading and Learning to Read*, Stanford, Cal., 23–25 March 2009, pages 29–35.

- Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Mich., 25–30 June 2005, pages 363–370.
- Finlayson, M. & Kulkarni, N. (2011). Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, Portland, Oregon, USA, 23 June 2011, pages 20–24.
- Fleischman, M. & Hovy, E. (2002). Fine grained classification of named entities. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, 24 August – 1 September 2002.
- Fleischman, M. & Hovy, E. (2004). Multi-document person name resolution. In *Proceedings of the Workshop on Reference Resolution and Its Applications*, Barcelona, Spain, 25–26 July 2004, pages 1–8.
- Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift fr Philosophie und philosophische Kritik, NF 100*, pages 25–50.
- Friedman, N., Getoor, L., Koller, D., & Pfeffer, A. (1999). Learning probabilistic relational models. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, 1–6 August 1999, volume 2, pages 1300–1307.
- Gabrilovich, E. & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, Mass., 16–20 July 2006, pages 1301–1306.
- Gabrilovich, E. & Markovitch, S. (2007a). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 6–12 January 2007, pages 1606–1611.
- Gabrilovich, E. & Markovitch, S. (2007b). Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. *Journal of Machine Learning Research*, 8:2297–2345.
- Gabrilovich, E., Ringgaard, M., & Subramanya, A. (June 2013). FACC1: Freebase annotation of ClueWeb corpora, version 1.

- Gale, W., Church, K. W., & Yarowsky, D. (1992a). Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Del., 28 June – 2 July 1992, pages 249–256.
- Gale, W. A., Church, K. W., & Yarowsky, D. (1992b). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5-6):415–439.
- Gale, W. A., Church, K. W., & Yarowsky, D. (1992c). One sense per discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop*, New York, N.Y., 23-26 February 1992, pages 233–237.
- Gale, W. A., Church, K. W., & Yarowsky, D. (1992d). Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, Quebec, Canada, pages 101–112.
- Gelbukh, A. & Kolesnikova, O. (2013). Multiword expressions in NLP. In Bandyopadhyay, S., Naskar, S. K., & Ekbal, A. (Eds.), *Emerging Applications of Natural Language Processing: Concepts and New Research*, pages 1–21. IGI Global.
- Getoor, L. & Taskar, B. (2007a). Introduction. In Getoor, L. & Taskar, B. (Eds.), *Introduction to Statistical Relational Learning*, pages 1–11. Cambridge, Mass.: MIT Press.
- Getoor, L. & Taskar, B. (Eds.) (2007b). *Introduction to Statistical Relational Learning*. Cambridge, Mass.: MIT Press.
- Gildea, D. & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.
- Gliozzo, A., Strapparava, C., & Dagan, I. (2004). Unsupervised and supervised exploitation of semantic domains in lexical disambiguation. *Computer Speech & Language*, 18(3):275–299.
- Gonzalo, J., Peas, A., & Verdejo, F. (1999). Lexical ambiguity and information retrieval revisited. In *Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, Md., 21–22 June 1999.
- Gooi, C. H. & Allen, J. (2004). Cross-document coreference resolution on a large scale corpus. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, Mass., 2–7 May 2004, pages 9–16.

- Gottipati, S. & Jiang, J. (2011). Linking entities to a knowledge base with query expansion. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, U.K., 27–29 July 2011, pages 804–813.
- Granger, R. H. (1982). Scruffy text understanding: Design and implementation of 'tolerant' understanders. In *Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics*, Toronto, Ontario, Canada, 16–18 June 1982, pages 157–160.
- Green, S., Andrews, N., Gormley, M. R., Dredze, M., & Manning, C. D. (2012). Entity clustering across languages. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Québec, Canada, 3–8 June 2012, pages 60–69.
- Guiliano, C., Gliozzo, A., & Strapparava, C. (2009). Kernel methods for minimally supervised WSD. *Computational Linguistics*, 35(4):513–528.
- Guo, S., Chang, M.-W., & Kiciman, E. (2013a). To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 9–14 June 2013, pages 1020–1030.
- Guo, W. & Diab, M. (2010). COLEPL and COLSLM: An unsupervised WSD approach to multilingual lexical substitution, Tasks 2 and 3 SemEval 2010. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2)*, Uppsala, Sweden, 15–16 July 2010, pages 129–133.
- Guo, W., Li, H., Ji, H., & Diab, M. (2013b). Linking tweets to news: A framework to enrich short text data in social media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 4–9 August 2013, pages 239–249.
- Guo, Y., Che, W., Liu, T., & Li, S. (2011). A graph-based method for entity linking. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, 8–13 November 2011, pages 1010–1018.
- Guo, Y., Qin, B., Liu, T., & Li, S. (2013c). Microblog entity linking by leveraging extra posts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Wash., 18–21 October 2013, pages 863–868.
- Guo, Z., Xu, Y., Mesquita, F., Barbosa, D., & Kondrak, G. (2012). UAlberta at TAC-KBP 2012. English and cross-lingual entity linking. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 5–6 November 2012.

- Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., & Wirth, C. (2012). UBY – A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 23–27 April 2012, pages 580–590.
- Gurobi Optimization, I. (2014). Gurobi optimizer reference manual.
- Gutiérrez, Y., Castañeda, Y., González, A., Estrada, R., Piug, D. D., Abreu, J. I., Pérez, R., Fernández Orquín, A., Montoyo, A., Muñoz, R., & Camara, F. (2013). UMCC.DLSI: Reinforcing a ranking algorithm with sense frequencies and multidimensional semantic resources to solve multilingual word sense disambiguation. In *Proceedings of STARSEM 2013: The 2nd Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, 14–15 June 2013, pages 241–249.
- Habash, N. & Dorr, B. (2003). A categorial variation database for English. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Alberta, Canada, 27 May –1 June 2003, pages 17–23.
- Habib, M. B., van Keule, M., & Zhu, Z. (2014). Named entity extraction and linking challenge: University of Twente at #microposts2014. In *Proceedings of the 4th Workshop on Making Sense of Microposts*, Seoul, Korea, 7 April 2014.
- Hachey, B., Nothman, J., & Radford, W. (2014). Cheap and easy entity evaluation. In *Proceedings of the ACL 2014 Conference Short Papers*, Baltimore, Md., 22–27 June 2014, pages 464–469.
- Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating entity linking with Wikipedia. *Artificial Intelligence*, 194:130–150.
- Hajishirzi, H., Zilles, L., Weld, D. S., & Zettlemoyer, L. (2013). Joint coreference resolution and named-entity linking with multi-pass sieves. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Wash., 18–21 October 2013, pages 289–299.
- Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. London, U.K.: Longman.
- Han, X. & Sun, L. (2011). A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oreg., 19–24 June 2011, pages 945–954.

- Han, X. & Sun, L. (2012). An entity-topic model for entity linking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea, 8–14 July 2012, pages 105–115.
- Han, X., Sun, L., & Zhao, J. (2011). Collective entity linking in web text: A graph-based method. In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* Beijing, China, 25–29 July 2011, pages 765–774.
- Han, X. & Zhao, J. (2009). Named entity disambiguation by leveraging Wikipedia semantic knowledge. In *Proceedings of the ACM 18th Conference on Information and Knowledge Management (CIKM 2009)*, Hong Kong, China, 2–6 November 2009, pages 215–224.
- Han, X. & Zhao, J. (2010). Structural semantic relatedness: A knowledge-based method to named entity recognition. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11–16 July 2010, pages 50–59.
- Harris, Z. (1968). *Mathematical Structures of Language*. New York: Interscience Publishers John Wiley & Sons (Interscience Tracts in Pure and Applied Mathematics; 21).
- He, Z., Liu, S., Li, M., Zhou, M., Zhang, L., & Wang, H. (2013). Learning entity representation for entity disambiguation. In *Proceedings of the ACL 2013 Conference Short Papers*, Sofia, Bulgaria, 4–9 August 2013, pages 30–34.
- Henrich, V., Hinrichs, E., & Vodolazova, T. (2012). WebCAGe – a web-harvested corpus annotated with GermaNet senses. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 23–27 April 2012, pages 387–396.
- Hirst, G. & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum, C. (Ed.), *WordNet: An Electronic Lexical Database*, pages 305–332. Cambridge, Mass.: MIT Press.
- Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.
- Hoffart, J., Altun, Y., & Weikum, G. (2014). Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd World Wide Web Conference*, Seoul, Korea, 7–11 April, 2011, pages 385–395.
- Hoffart, J., Seufert, S., Nguyen, D. B., Theobald, M., & Weikum, G. (2012). KORE: Keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the ACM 21st Conference on Information and Knowledge Management (CIKM 2012)*, Maui, Hawaii, USA, 26–30 October 2010, pages 545–554.

- Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-Kelham, E., de Melo, G., & Weikum, G. (2011a). YAGO2: Exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th World Wide Web Conference*, Hyderabad, India, 28 March – 1 April, 2011, pages 229–232.
- Hoffart, J., Yosef, M. A., Bordino, I., Fürstenauf, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., & Weikum, G. (2011b). Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, U.K., 27–29 July 2011, pages 782–792.
- Hong, K. & Nenkova, A. (2014). Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 26–30 April 2014, pages 712–721.
- Hoste, V., Kool, A., & Daelemans, W. (2001). Classifier optimization and combination in the English all words task. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2) at ACL-01*, Toulouse, France, 5–6 July 2001, pages 83–86.
- Hou, Y., Markert, K., & Strube, M. (2013). Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 9–14 June 2013, pages 907–917.
- Hovy, E., Navigli, R., & Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.
- Hu, J., Fanj, L., Cao, Y., Zeng, H.-J., Li, H., Yang, Q., & Chen, Z. (2008). Enhancing text clustering by leveraging Wikipedia semantics. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* Singapore, 20–24 July 2008, pages 179–186.
- Huang, D., Geva, S., & Trotman, A. (2010). Overview of the INEX 2009 link the Wiki track. In Geva, S., Kamps, J., & Trotman, A. (Eds.), *Focused Retrieval and Evaluation, 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009, Brisbane, Australia, 7–9 December 2009, Revised and Selected Papers*, pages 312–323. Heidelberg, Germany: Springer.
- Huang, D. W., Xu, Y., Trotman, A., & Geva, S. (2008). Overview of INEX 2007 link the Wiki track. In Fuhr, N., Kamps, J., Lalmas, M., & Trotman, A. (Eds.), *Focused Access*

- to XML Documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007, Dagstuhl Castle, Germany, 17-19 December 2007. Selected Papers*, pages 373–387. Berlin, Heidelberg, German: Springer.
- Huynh, T. N. & Mooney, R. J. (2009). Max-margin weight learning for Markov logic networks. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Bled, Slovenia, 7–11 September 2009, pages 248–263.
- Huynh, T. N. & Mooney, R. J. (2011). Online max-margin weight learning for Markov Logic Networks. In *Proceedings of Eleventh SIAM International Conference on Data Mining*, Phoenix, Arizona, USA, 28–30 April 2011, pages 642–651.
- Ide, N. (2000). Cross-lingual sense determination: Can it work? *Computers and the Humanities*, 34(1-2):223–234.
- Ide, N., Erjavec, T., & Tufis, D. (2002). Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, Penn., July 2002, pages 61–66.
- Ide, N. & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):2–40.
- Jarmasz, M. & Szpakowicz, S. (2003). Not as easy as it seems: Automating the construction of lexical chains using Roget's Thesaurus. In *Proceedings of the Canadian Conference on Artificial Intelligence*, Halifax, Canada, 11–13 June 2003, pages 544–549.
- Jensen, D., Neville, J., & Gallagher, B. (2004). Why collective inference improves relational classification. In *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Seattle, Wash., 22–25 August 2004, pages 593–598.
- Ji, H. & Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oreg., 19–24 June 2011, pages 1148–1158.
- Ji, H., Grishman, R., & Dang, H. (2011). Overview of the TAC 2011 knowledge base population track. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.
- Ji, H., Grishman, R., Dang, H., Griffitt, K., & Ellis, J. (2010). Overview of the TAC 2010 knowledge base population track. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 15–16 November 2010.

- Jin, P., Wu, Y., & Yu, S. (2007). SemEval-2007 Task 05: Multilingual Chinese-English lexical sample. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-1)*, Prague, Czech Republic, 23–24 June 2007, pages 19–23.
- Jin, Y., Kıcıman, E., Wang, K., & Loynd, R. (2014). Entity linking at the tail: Sparse signals, unknown entities, and phrase models. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, New York, N.Y., 24–28 February 2014, pages 453–462.
- Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, 23–26 July 2002, pages 133–142.
- Kahusk, N., Orav, H., & Oim, H. . (2001). Sensiting inflectionality: Estonian task for SENSEVAL-2. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2) at ACL-01*, Toulouse, France, 5–6 July 2001, pages 25–28.
- Kaplan, A. (1955). An experimental study of ambiguity and context. *Mechanical Translation*, 2:39–46.
- Karttunen, L. (1976). Discourse referents. *Syntax and Semantics*, pages 363–386.
- Kautz, H., Selman, B., & Jiang, Y. (1996). A general stochastic approach to solving problems with hard and soft constraints. In Gu, D., Du, J., & Pardalos, P. (Eds.), *The Satisfiability Problem: Theory and Applications*, pages 573–586. New York, NY: American Mathematical Society.
- Kehagias, A., Petridis, V., Kaburlasos, V. G., & Fragkou, P. (2001). A comparison of word- and sense-based text categorization using several classification algorithms. *Journal of Intelligent Information Systems*, 21:227–247.
- Kennedy, A. & Szpakowicz, S. (2008). Evaluating Roget’s thesauri. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio, 15–20 June 2008, pages 416–424.
- Khapra, M., Sohoney, S., Kulkarni, A., & Bhattacharyya, P. (2010). Value for money: Balancing annotation effort, lexicon building and accuracy for multilingual WSD. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 23–27 August 2010, pages 555–563.

- Khapra, M. M., Joshi, S., Chatterjee, A., & Bhattacharyya, P. (2011). Together we can: Bilingual bootstrapping for WSD. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oreg., 19–24 June 2011, pages 561–569.
- Khapra, M. M., Shah, S., Kedia, P., & Bhattacharyya, P. (2009). Projecting parameters for multilingual word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 6–7 August 2009, pages 459–467.
- Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.
- Kilgarriff, A. (2001). English lexical sample task description. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2) at ACL-01*, Toulouse, France, 5–6 July 2001, pages 17–20.
- Kilgarriff, A. & Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computer and the Humanities*, 34(1-2):15–48.
- Kim, J. & Gurevych, I. (2013). UKP at CrossLink2: CJK-to-English subtasks. In *Proceedings of the 10th NTCIR Workshop Meeting*, Tokyo, Japan, 18-21 June 2013, pages 57–61.
- Kim, J. & Gurevych, I. (2011). UKP at CrossLink: Anchor text translation for cross-lingual link discovery. In *Proceedings of the 9th NTCIR Workshop Meeting*, Tokyo, Japan, 6-9 December 2011, pages 487–494.
- Knoth, P. & Herrmannova, D. (2013). Simple yet effective methods for cross-lingual link discovery (CLLD) – KMI at NTCIR-10 CrossLink-2. In *Proceedings of the 10th NTCIR Workshop Meeting*, Tokyo, Japan, 18–21 June 2013, pages 39–46.
- Knoth, P., Zilka, L., & Zdrahal, Z. (2011a). KMI, The Open University at NTCIR-9 CrossLink: Cross-lingual link discovery in Wikipedia using Explicit Semantic Analysis. In *Proceedings of the 9th NTCIR Workshop Meeting*, Tokyo, Japan, 6-9 December 2011.
- Knoth, P., Zilka, L., & Zdrahal, Z. (2011b). Using explicit semantic analysis for cross-lingual link discovery. In *Proceedings of the 5th International Workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies*, Chiang Mai, Thailand, 13 November 2011, pages 2–10.

- Koeling, R., McCarthy, D., & Carroll, J. (2005). Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing*, Vancouver, B.C., Canada, 6–8 October 2005, pages 419–426.
- Kok, S., Singla, P., Richardson, M., & Domingos, P. (2005). The Alchemy system for statistical relational AI. Technical report, University of Washington, Seattle, WA.
- Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In Getoor, L. & Taskar, B. (Eds.), *Introduction to Statistical Relational Learning*, pages 13–55. MIT Press, Cambridge, Mass.
- Koo, T. & Collins, M. (2005). Hidden-variable models for discriminative reranking. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing*, Vancouver, B.C., Canada, 6–8 October 2005, pages 507–514.
- Kozareva, Z. & Ravi, S. (2011). Unsupervised name ambiguity resolution using a generative model. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, U.K., 27–29 July 2011, pages 105–112.
- Kripke, S. (1980). *Naming and Necessity*. Basil Blackwell, Oxford, U.K.
- Krovetz, R. & Croft, W. B. (1989). Word sense disambiguation using machine-readable dictionaries. In *Proceedings of the 12th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Cambridge, Massachusetts, USA, 25–28 June 1989, pages 127–136.
- Krovetz, R. (1998). More than one sense per discourse. Technical report, NEC Research Institute.
- Kulkarni, S., Singh, A., Ramakrishnan, G., & Chakrabarti, S. (2009). Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Paris, France, 28 June – 1 July 2009, pages 457–466.
- Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato’s problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

- Lau, H. J., Cook, P., McCarthy, D., Gella, S., & Baldwin, T. (2014). Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Md., 22–27 June 2014, pages 259–270.
- Leacock, C., Miller, G. A., & Chodorow, M. (1998). Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.
- Leacock, C., Towell, G., & Voorhees, E. (1993). Corpus-based statistical sense resolution. In *Proceedings of a Workshop on Human Language Technology*, Plainsboro, N.J., 21–24 March 1993, pages 260–265.
- Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.
- Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Shared Task of the 15th Conference on Computational Natural Language Learning*, Portland, Oreg., 23–24 June 2011, pages 28–34.
- Lee, H., Recasens, M., Chang, A., Surdeanu, M., & Jurafsky, D. (2012). Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, pages 489–500.
- Lefever, E. & Hoste, V. (2010). SemEval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, Uppsala, Sweden, 15–16 July 2010, pages 15–20.
- Lefever, E. & Hoste, V. (2013). SemEval-2013 Task 10: Cross-lingual word sense disambiguation. In *Proceedings of STARSEM 2013: The 2nd Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, 14–15 June 2013, pages 158–166.
- Lefever, E., Hoste, V., & De Cock, M. (2011). ParaSense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oreg., 19–24 June 2011, pages 317–322.

- Lefever, E., Hoste, V., & Fayruzov, T. (2007). AUG: A combined classification and clustering approach for web people disambiguation. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-1)*, Prague, Czech Republic, 23–24 June 2007, pages 105–108.
- Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. M., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., & Bizer, C. (2014). DBpedia – A large scale, multi-lingual knowledge base extracted from Wikipedia. *Semantic Web Journal*. To appear.
- Lenat, D. B. (1995). Cyc: A large scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual Conference on Systems Documentation*, Toronto, Ontario, Canada, pages 24–26.
- Li, C. & Li, H. (2002). Word translation disambiguation using bilingual bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Penn., 7–12 July 2002, pages 343–351.
- Li, X., Ellis, J., Griffit, K., Strassel, S. M., Parker, R., & Wright, J. (2011). Linguistic resources for 2011 knowledge base population evaluation. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.
- Li, Y., Wang, C., Han, F., Han, J., Roth, D., & Yan, X. (2013). Mining evidences for named entity disambiguation. In *Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Chicago, Ill., 11–14 August 2013, pages 1070–1078.
- Liddy, E. D. & Paik, W. (1992). Statistically-guided word sense disambiguation. In *Proceedings of the AAAI Fall Symposium*, Cambridge, Massachusetts, 23–25 October 1992, pages 98–107.
- Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montréal, Québec, Canada, 10–14 August 1998, pages 768–774.
- Lin, T., Mausam, & Etzioni, O. (2012). No noun phrase left behind: Detecting and typing unlinkable entities. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, pages 893–903.

- Litkowski, K. (2004). Senseval-3 task: Word sense disambiguation of WordNet glosses. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3) at ACL-04*, Barcelona, Spain, 25–26 July 2004, pages 13–16.
- Liu, S., Liu, F., Yu, C., & Meng, W. (2004). An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* New York, N.Y., 25–29 July 2004, pages 266–272. ACM Press.
- Löbner, S. (1985). Definites. *Journal of Semantics*, 4:279–326.
- Lowd, D. & Domingos, P. (2007). Efficient weight learning for Markov logic networks. In *Proceedings of the 11th European Conference on Principles and Practices of Knowledge Discovery in Databases*, Warsaw, Poland, 17–21 September 2007, pages 200–211.
- Lu, Z., Kao, H.-Y., Wei, C.-H., Huang, M., Liu, J., Kuo, C.-J., Hsu, C.-N., Tsai, R. T.-H., Dai, H.-J., Okazaki, N., Cho, H.-C., Gerner, M., Solt, I., Agarwal, S., Liu, F., Vishnyakova, D., Ruch, P., Romacker, M., Rinaldi, F., Bhattacharya, S., Srinivasan, P., Liu, H., Torii, M., Matos, S., Campos, D., Verspoor, K., Livingston, K. M., & Wilbur, W. J. (2011). The gene normalization task in BioCreative III. *BMC Bioinformatics*, 12(Suppl.8):S2.
- Macskassy, S. A. & Provost, F. (2003). A simple relational classifier. In *Proceedings of the Second Workshop on Multi-Relational Data Mining (MRDM-2003) at KDD-2003*, Washington, D.C., 27 August 2003, pages 64–76.
- Madhu, S. & Lytel, D. (1965). A figure of merit technique for the resolution of non-grammatical ambiguity. *Mechanical Translation*, 8(2):9–13.
- Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. (2002). The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373.
- Mahesh, K., Nirenburg, S., Cowie, J., & Farwell, D. (1996). An assessment of Cyc for natural language processing. Technical Report MCCA-96-302, CRL, New Mexico State University, Las Cruces, New Mexico.
- Manandhar, S., Klapaftis, I., Dligach, D., & Pradhan, S. (2010). SemEval-2010 Task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2)*, Uppsala, Sweden, 15–16 July 2010, pages 63–68.
- Mann, G. S. & Yarowsky, D. (2003). Unsupervised personal name disambiguation. In *Proceedings of the 7th Conference on Computational Natural Language Learning*, Edmonton, Alberta, Canada, 31 May – 1 June 2003, pages 33–40.

- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the ACL 2014 System Demonstrations*, Baltimore, Md., 22–27 June 2014, pages 55–60.
- Markert, K., Hou, Y., & Strube, M. (2012). Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea, 8–14 July 2012, pages 795–804.
- Màrquez, L., Escudero, G., Martínez, D., & Rigau, G. (2006). Supervised corpus-based methods for WSD. In Agirre, E. & Edmonds, P. (Eds.), *Word Sense Disambiguation: Algorithms and Applications*, pages 167–216. Heidelberg, Germany: Springer.
- Martin, J. R. (1992). *English Text: System and Structure*. Amsterdam and Philadelphia: John Benjamins.
- Martínez, D. & Agirre, E. (2000). One sense per collocation and genre/topic variations. *CoRR*.
- Matsuzaki, T., Miyao, Y., & Tsujii, J. (2005). Probabilistic CFG with latent annotations. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Mich., 25–30 June 2005, pages 75–82.
- Mayfield, J., Lawrie, D., McNamee, P., & Oard, D. W. (2011). Building a cross-language entity linking collection in 21 languages. In *Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation*, Amsterdam, The Netherlands, 19–22 September 2011.
- McCallum, A. (2009). Joint inference for natural language processing. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, Boulder, Colorado, 15–16 July 2010.
- McCarthy, D. (2006a). Relating WordNet senses for word sense disambiguation. In *Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, Trento, Italy, 6 April 2006, pages 17–24.
- McCarthy, D. (2006b). Relating wordnet senses for word sense disambiguation. In *Proceedings of the ACL Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, Trento, Italy, 4 April 2006, pages 17–24.
- McCarthy, D., Koeling, R., Weeds, J., & Carroll, J. (2007). Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.

- McNamee, P. (2010). HLTCOE efforts in entity linking at TAC 2010. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 15–16 November 2010.
- McNamee, P. & Dang, H. T. (2009). Overview of the TAC 2009 knowledge base population track. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 16–18 November 2009.
- McNamee, P., Mayfield, J., Lawrie, D., Oard, D., & Doermann, D. (2011). Cross-language entity linking. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, 8–13 November 2011, pages 255–263.
- Medelyan, O. & Legg, C. (2008). Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense. In *Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08*, Chicago, Ill., 13 July 2008, pages 13–18.
- Mehdad, Y., Negri, M., & Federico, M. (2011). Using bilingual parallel corpora for cross-lingual textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, 19–24 June 2011, pages 1336–1345.
- Mehdad, Y., Negri, M., & Federico, M. (2012). Detecting semantic equivalence and information disparity in cross-lingual documents. In *Proceedings of the ACL 2012 Conference Short Papers*, Jeju Island, Korea, 8–14 July 2012, pages 120–124.
- Meij, E., Balog, K., & Odijk, D. (2013a). Entity linking and retrieval. In *Proceedings of the 22nd World Wide Web Conference*, Rio de Janeiro, Brazil, 28 March – 1 April, 2013.
- Meij, E., Balog, K., & Odijk, D. (2013b). Entity linking and retrieval. In *Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* Dublin, Ireland, 28 July – 1 August 2013, pages 1127–1127.
- Meij, E., Balog, K., & Odijk, D. (2014). Entity linking and retrieval for semantic search. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, New York, N.Y., 24–28 February 2014, pages 683–684.
- Meij, E., Weerkamp, W., & de Rijke, M. (2012). Adding semantics to microblog posts. In *Proceedings of the 5th International Conference on Web Search and Web Data Mining*, Seattle, Wash., 8–12 February 2012, pages 563–572.
- Merhav, Y., Barry, J., Clarke, J., & Murgatroyd, D. (2013). Basis Technology at TAC 2013 entity linking. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 18–19 November 2013.

- Meyer, C. M. & Gurevych, I. (2010). How web communities analyze human language: Word senses in Wiktionary. In *Proceedings of the Second Web Science Conference*, Raleigh, NC, USA, 26–27 April 2010.
- Meyer, C. M. & Gurevych, I. (2012). To exhibit is not to loiter: A multilingual, sense-disambiguated Wiktionary for measuring verb similarity. In *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India, 8–15 December 2012, pages 1763–1780.
- Meza-Ruiz, I. & Riedel, S. (2009). Jointly identifying predicates, arguments and senses using Markov logic. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Col., 31 May – 5 June 2009, pages 155–163.
- Michelson, M. & Macskassy, S. A. (2010). Discovering users' topics of interest on Twitter: A first look. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, Toronto, ON, Canada, 26–30 October 2010, pages 73–80.
- Mihalcea, R. (2002). Bootstrapping large sense tagged corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain, 29–31 May 2002, pages 1407–1411.
- Mihalcea, R. (2006). Knowledge-based methods for WSD. In Agirre, E. & Edmonds, P. (Eds.), *Word Sense Disambiguation: Algorithms and Applications*, pages 107–131. Heidelberg, Germany: Springer.
- Mihalcea, R. (2007). Using Wikipedia for automatic word sense disambiguation. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, N.Y., 22–27 April 2007, pages 196–203.
- Mihalcea, R., Chklovski, T., & Kilgarriff, A. (2004a). The Senseval-3 English lexical sample task. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3) at ACL-04*, Barcelona, Spain, 25–26 July 2004, pages 25–28.
- Mihalcea, R. & Csomai, A. (2005). SenseLearner: Word sense disambiguation for all words in unrestricted text. In *Proceedings of the Interactive Poster and Demonstrations Sessions at the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Mich., 25–30 June 2005, pages 53–56.

- Mihalcea, R. & Csomai, A. (2007). Linking documents to encyclopedic knowledge. In *Proceedings of the ACM 16th Conference on Information and Knowledge Management (CIKM 2007)*, Lisbon, Portugal, 6–9 November 2007, pages 233–242.
- Mihalcea, R. & Edmonds, P. G. (Eds.) (2004). *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3) at ACL-04*, Barcelona, Spain, 25–26 July 2004.
- Mihalcea, R. & Moldovan, D. (2001a). Automatic generation of a coarse grained WordNet. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA, June 2001, page ??
- Mihalcea, R. & Moldovan, D. I. (1999). An automatic method for generating sense tagged corpora. In *Proceedings of the 16th National Conference on Artificial Intelligence*, Orlando, Flo., 18–22 July 1999, pages 461–466.
- Mihalcea, R., Tarau, P., & Figa, E. (2004b). PageRank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, 23–27 August 2004, pages 1126–1132.
- Mihalcea, R. F. & Moldovan, D. I. (2001b). Pattern learning and active feature selection for word sense disambiguation. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2) at ACL-01*, Toulouse, France, 5–6 July 2001, pages 127–130.
- Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.
- Miller, G. A., Leacock, C., Teng, R., & Bunker, R. (1993). A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, Plainsboro, N.J., 21–24 March 1993, pages 303–308.
- Miller, T., Biemann, C., Zesch, T., & Gurevych, I. (2012). Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India, 8–15 December 2012, pages 1781–1796.

- Miller, T., Erbs, N., Zorn, H.-P., Zesch, T., & Gurevych, I. (2013). DKPro WSD: A generalized UIMA-based framework for word sense disambiguation. In *Proceedings of the ACL 2013 System Demonstrations*, Sofia, Bulgaria, 4–9 August 2013, pages 37–42.
- Milne, D. & Witten, I. H. (2008a). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08*, Chicago, Ill., 13 July 2008, pages 25–30.
- Milne, D. & Witten, I. H. (2008b). Learning to link with Wikipedia. In *Proceedings of the ACM 17th Conference on Information and Knowledge Management (CIKM 2008)*, Napa Valley, Cal., USA, 26–30 October 2008, pages 1046–1055.
- Moldovan, D., Paşca, M., Harabagiu, S., & Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems*, 21(2):133–154.
- Moldovan, D. & Rus, V. (2001). Logic form transformation of WordNet and its applicability to question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, pages 394–401.
- Moldovan, D. I. & Mihalcea, R. (2000). Using WordNet and lexical operators to improve internet searches. *IEEE Internet Computing*, 4(1):34–43.
- Moldovan, D. I. & Novischi, A. (2004). Word sense disambiguation of WordNet glosses. *Computer Speech & Language*, 18(3):301–317.
- Monahan, S. & Carpenter, D. (2012). Lorify: A knowledge base from scratch. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 5–6 November 2012.
- Monahan, S., Lehmann, J., Nyberg, T., Plymale, J., & Jung, A. (2011). Cross-lingual cross-document coreference with entity linking. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.
- Monz, C., Nastase, V., Negri, M., Fahrni, A., Mehdad, Y., & Strube, M. (2011). CoSyne: A framework for multilingual content synchronization of Wikis. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, Mountain View, Calif., 3–5 October 2011, pages 217–218.
- Morato, J., Marzal, M. A., Lloréns, J., & Moreiro, J. (2004). WordNet applications. In *Proceedings of the Second Global WordNet Conference* Brno, Czech Republic, 20–23 January 2004.

- Moro, A., Navigli, R., Tucci, F. M., & Passonneau, R. J. (2014a). Annotating the MASC corpus with BabelNet. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 26–31 May 2014.
- Moro, A., Raganato, A., & Navigli, R. (2014b). Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.
- Morris, J. & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.
- Moschitti, R. & Basili, R. (2004). Complex linguistic features for text classification: a comprehensive study. In *Proceedings of the 26th European Conference on Advances in Information Retrieval*, Glasgow, U.K., 30 March – 3 April 2004, pages 181–196. Springer Verlag.
- Muggleton, S. (1995). Stochastic logic programs. In Raedt, L. D. (Ed.), *Advances in Inductive Logic Programming*. IOS Press, Ohmsha.
- Müller, C. & Strube, M. (2001). MMAX: A tool for the annotation of multi-modal corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Seattle, Wash., 5 August 2001, pages 45–50.
- Nadeau, D. & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.
- Nakashole, N., Tylenda, T., & Weikum, G. (2013). Fine-grained semantic typing of emerging entities. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 4–9 August 2013, pages 1488–1497.
- Nastase, V. (2008). Topic-driven multi-document summarization with encyclopedic knowledge and activation spreading. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 763–772.
- Nastase, V., Fahrni, A., & Strube, M. (2011). Deliverable 4.3: Text segmentation and coarse grained alignment. Technical report, HITS gGmbH.
- Nastase, V., Judea, A., Markert, K., & Strube, M. (2012). Local and global context for supervised and unsupervised metonymy resolution. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, pages 183–193.

- Nastase, V. & Strube, M. (2013). Transforming Wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85.
- Navigli, R. & Ponzetto, S. P. (2012a). BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.
- Navigli, R. (2006). Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 17–21 July 2006, pages 105–112. Association for Computational Linguistics.
- Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.
- Navigli, R., Jurgens, D., & Vannella, D. (2013). SemEval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of STARSEM 2013: The 2nd Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, 14–15 June 2013, pages 222–231.
- Navigli, R. & Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692.
- Navigli, R., Litkowski, K. C., & Hargraves, O. (2007). SemEval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-1)*, Prague, Czech Republic, 23–24 June 2007, pages 30–35.
- Navigli, R. & Ponzetto, S. P. (2012b). Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, pages 1399–1410.
- Navigli, R. & Vannella, D. (2013). SemEval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Proceedings of STARSEM 2013: The 2nd Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, 14–15 June 2013, pages 193–201.

- Nelken, R. & Shieber, S. (2006). Lexical chaining and word-sense-disambiguation. Technical Report TR-06-07, Computer Science Group, Harvard University, Cambridge, Mass.
- Newman, M. E. (2010). *Networks: An Introduction*. Oxford University Press, New York, N.Y.
- Ng, H. T. & Lee, H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, Cal., 24–27 June 1996, pages 40–47.
- Ng, H. T., Lim, C. Y., & Foo, S. K. (1999). A case study on inter-annotator agreement for word sense disambiguation. In *Proceedings of SIGLEX99: Standardizing Lexical Resources*, College Park, Maryland, 21–22 June 1999, pages 9–13.
- Ng, H. T., Wang, B., & Chan, Y. S. (2003). Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 7–12 July 2003, pages 455–462.
- Niu, C., Li, W., & Srihari, R. K. (2004). Weakly supervised learning for cross-document person name disambiguation supported by information extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 21–26 July 2004, pages 597–604.
- Niu, C., Li, W., Srihari, R. K., & Li, H. (2005). Word independent context pair classification model for word sense disambiguation. In *Proceedings of the 9th Conference on Computational Natural Language Learning*, Ann Arbor, Mich., USA, 29–30 June 2005, pages 33–39.
- Niu, F., Ré, C., Doan, A., & Shavlik, J. (2011). Tuffy: Scaling up statistical inference in markov logic networks using an RDBMS. *Proc. VLDB Endow.*, 4(6):373–384.
- Noessner, J., Niepert, M., & Stuckenschmidt, H. (2013). RockIt: Exploiting parallelism and symmetry for MAP inference in statistical relational models. *CoRR*.
- Noreen, E. (1989). *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. New York: Interscience Publishers John Wiley & Sons.
- Ogden, C. K. & Richards, I. A. (1923). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Kegan Paul, London, U.K.

- Okumura, M. & Honda, T. (1994). Word sense disambiguation and text segmentation based on lexical cohesion. In *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan, 5–9 August 1994, volume 2, pages 755–761.
- Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: towards a wider multilinguality. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 21–27 May 2012.
- Palmer, M., Dang, H. T., & Fellbaum, C. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, pages 137–163.
- Palmer, M., Fellbaum, C., Cotton, S., Delfs, L., & Dang, H. T. (2001). English tasks: All-words and verb lexical sample. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2) at ACL-01*, Toulouse, France, 5–6 July 2001.
- Palmer, M., Ng, H., & Dang, H. (2006). Evaluation of WSD systems. In Agirre, E. & Edmonds, P. (Eds.), *Word Sense Disambiguation: Algorithms and Applications*, pages 75–106. Springer, Heidelberg, Germany.
- Parker, R., Graff, D., Kong, J., Chen, K., & Maeda, K. (2011). English Gigaword Fifth Edition. LDC2011T07.
- Passonneau, R., Salieb-Aouissi, A., & Ide, N. (2009). Making sense of word sense variation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, Boulder, Colorado, 4 June 2009, pages 2–9.
- Pedersen, T. (2006). Unsupervised corpus-based methods for WSD. In Agirre, E. & Edmonds, P. (Eds.), *Word Sense Disambiguation: Algorithms and Applications*, pages 133–166. Heidelberg, Germany: Springer.
- Pedersen, T., Banerjee, S., & Patwardhan, S. (2005). Maximizing semantic relatedness to perform word sense disambiguation. Research Report UMSI 2005/25, University of Minnesota Supercomputing Institute.
- Pedersen, T., Kulkarni, A., Angheluta, R., Kozareva, Z., & Solorio, T. (2006). An unsupervised language independent method of name discrimination using second order co-occurrence features. In Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing*, pages 208–222. Springer, Heidelberg, Germany.

- Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::Similarity – Measuring the relatedness of concepts. In *Companion Volume to the Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, Mass., 2–7 May 2004, pages 267–270.
- Peters, J. D. (1999). *Speaking into the air: a history of the idea of communication*. Chicago and London: The University of Chicago Press.
- Petrov, S., Barrett, L., Thibaux, R., & Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 17–21 July 2006, pages 433–440.
- Pianta, E., Bentivogli, L., & Girardi, C. (2002). MultiWordNet: developing an aligned multi-lingual database. In *Proceedings of the First International Conference on Global Word-Net*, Mysore, India, 21–25 January 2002, pages 93–302.
- Ponzetto, S. P. & Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11–16 July 2010, pages 1522–1531.
- Ponzetto, S. P. & Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, New York, N.Y., 4–9 June 2006, pages 192–199.
- Ponzetto, S. P. & Strube, M. (2007). Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212.
- Ponzetto, S. P. & Strube, M. (2011). Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175(9/10):1737–1756.
- Poon, H. & Domingos, P. (2008). Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 650–659.
- Popescu, O. (2009). Name perplexity. In *Companion Volume to the Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Col., 31 May – 5 June 2009, pages 153–156.
- Popescu, O. (2010). Dynamic parameters for cross document coreference [sic]. In *Proceedings of Coling 2010: Poster Volume*, Beijing, China, 23–27 August 2010, pages 988–996.

- Pradhan, S., Loper, E., Dligach, D., & Palmer, M. (2007). SemEval-2007 Task-17: English lexical sample, SRL and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-1)*, Prague, Czech Republic, 23–24 June 2007, pages 87–92.
- Pradhan, S., Moschitti, A., & Xue, N. (2012). CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Shared Task of the 16th Conference on Computational Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, pages 1–40.
- Prescher, D. (2005). Head-driven PCFGs with latent-head statistics. In *Proceedings of the Ninth International Workshop on Parsing Technology*, Vancouver, B.C., Canada, 9–10 October 2005, pages 115–124.
- Purandare, A. & Pedersen, T. (2004). Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the 8th Conference on Computational Natural Language Learning*, Boston, Mass., USA, 6–7 May 2004, pages 41–48.
- Putnam, H. (1975). *Mind, Language, and Reality*. Cambridge, United Kingdom: Cambridge University Press.
- Quattoni, A., Wang, S. B., Morency, L.-P., Collins, M., & Darrell, T. (2007). Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1848–1852.
- Radev, D. R., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H., Celebi, A., Liu, D., & Drabek, E. (2003). Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 7–12 July 2003, pages 375–382.
- Rahman, A. & Ng, V. (2011). Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oreg., 19–24 June 2011, pages 814–824.
- Rao, D., McNamee, P., & Dredze, M. (2010). Streaming cross document entity coreference resolution. In *Proceedings of Coling 2010: Poster Volume*, Beijing, China, 23–27 August 2010, pages 1050–1058.
- Rao, D., Paul, M., Fink, C., Yarowsky, D., Oates, T., & Coppersmith, G. (2011). Hierarchical Bayesian models for latent attribute detection in social networks. In *Proceedings of the 5th International Conference on Weblogs and Social Media*, Barcelona, Spain, 17–21 July 2011.

- Ratinov, L. & Roth, D. (2011). GLOW TAC-KBP2011 entity linking system. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.
- Ratinov, L. & Roth, D. (2012). Learning-based multi-sieve co-reference resolution with knowledge. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, pages 1234–1244.
- Ratinov, L., Roth, D., Downey, D., & Anderson, M. (2011). Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oreg., 19–24 June 2011, pages 1375–1384.
- Remus, S. & Biemann, C. (2013). Three knowledge-free methods for automatic lexical chain extraction. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 9–14 June 2013, pages 989–999.
- Resnik, P. (1993). *Selection and Information: A Class-based Approach to Lexical Relationships*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Penn.
- Resnik, P. & Yarowsky, D. (1999). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.
- Richardson, M. & Domingos, P. (2006). Markov logic networks. *Machine Learning*, 62(1-2):107–136.
- Riedel, S. (2008). Improving the accuracy and efficiency of MAP inference for Markov logic. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, Helsinki, Finland, 9–12 July 2008, pages 468–475.
- Riedel, S. (2009). *Efficient Prediction of Relational Structure and its Application to Natural Language Processing*. PhD thesis, University of Edinburgh.
- Rizzo, G., Troncy, R., Hellmann, S., & Brümmer, M. (2012). NERD meets NIF: lifting NLP extraction results to the linked data cloud. In *Proceedings of the 5th Workshop on Linked Data on the Web*, Lyon, France, 16 April 2012, Lyon, FRANCE.
- Rizzolo, N. & Roth, D. (2010). Learning based Java for rapid development of NLP systems. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, La Valetta, Malta, 17–23 May 2010, Valletta, Malta.

- Robkop, K., Thoongsup, S., Charoenporn, T., Sornlertlamvanich, V., & Isahara, H. (2010). WNMS: Connecting the distributed WordNet in the case of Asian WordNet. In *Proceedings of the 5th International Conference of the Global WordNet Association*, Mumbai, India, 31 January – 4 February 2010.
- Roget, P. M. (1964). *Roget's Thesaurus of English Words and Phrases*. Boston, Mass.: Gould and Lincoln.
- Roth, D. (1996). On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1-2):273–302.
- Roth, D., Ji, H., Chang, M.-W., & Cassidy, T. (2014). Wikification and beyond: The challenges of entity and concept grounding. In *Proceedings of the ACL 2014 Conference Tutorials*, Baltimore, Md., 22–27 June 2014, page 7.
- Roth, D. & Yih, W.-t. (2004). A linear programming formulation for global inference in natural language tasks. In *Proceedings of the 8th Conference on Computational Natural Language Learning*, Boston, Mass., USA, 6–7 May 2004, pages 1–8.
- Rozovskaya, A. & Sproat, R. (2007). Multilingual word sense discrimination: A comparative cross-linguistic study. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007 Special Theme: Information Extraction and Enabling Technologies*, Prague, Czech Republic, 29 June 2007, pages 82–87.
- Russell, S. J. & Norvig, P. (1995). *Artificial Intelligence. A Modern Approach*. Prentice Hall, Englewood Cliffs, N.J.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing, 3rd International Conference*, Mexico City, Mexico, 16–22 February 2002, pages 1–15.
- Salton, G. & Lesk, M. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36.
- Salton, G. & McGill, M. (1983). *Introduction to Modern Information Retrieval*. New York, N.Y.: McGraw-Hill.
- Sanderson, M. (1994). Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 3–6 July 1994, pages 142–151.

- Santamaría, C., Gonzalo, J., & Artiles, J. (2010). Wikipedia as sense inventory to improve diversity in web search results. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11–16 July 2010, pages 1357–1366.
- Schneider, D. M. (1968). *American Kinship: A cultural account*. New York: Prentice Hall.
- Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.
- Scott, S. & Matwin, S. (1998). Text classification using WordNet hypernyms. In *Proceedings of Usage of WordNet in natural language processing systems*, Montreal, Quebec, Canada, 16 August 1998, pages 45–52.
- Searle, J. R. (1958). *Proper Names*, chapter 2, pages 166–173. New Series, Vol. 67, No. 266. Oxford University Press.
- Sen, P. (2012). Collective context-aware topic models for entity disambiguation. In *Proceedings of the 21st World Wide Web Conference*, Lyon, France, 28 March – 1 April, 2012, pages 729–738.
- Sil, A., Cronin, E., Nie, P., Yang, Y., Popescu, A.-M., & Yates, A. (2012). Linking named entities to any database. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, pages 116–127.
- Silberer, C. & Ponzetto, S. P. (2010). UHD: Cross-lingual word sense disambiguation using multilingual co-occurrence graphs. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2)*, Uppsala, Sweden, 15–16 July 2010, pages 134–137.
- Singh, S., Subramanya, A., Pereira, F., & McCallum, A. (2011). Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oreg., 19–24 June 2011, pages 793–803.
- Singh, S., Subramanya, A., Pereira, F., & McCallum, A. (2012). Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical Report UM-CS-2012-015, University of Massachusetts, Amherst.
- Singla, P. & Domingos, P. (2005). Discriminative training of Markov Logic Networks. In *Proceedings of the 20th National Conference on Artificial Intelligence*, Pittsburgh, Penn., 9–13 July 2005, volume 2, pages 868–873.

- Smith, N. A. (2011). *Linguistic Structure Prediction*. Morgan & Claypool Publishers.
- Snow, R., Prakash, S., Jurafsky, D., & Ng, A. Y. (2007). Learning to merge word senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning*, Prague, Czech Republic, 28–30 June 2007, pages 1005–1014.
- Snyder, B. & Palmer, M. (2004). The English all-words task. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3) at ACL-04*, Barcelona, Spain, 25–26 July 2004, pages 41–43.
- Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.
- Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20. Reprinted in *Journal of Documentation*, 60(5), pp.493-502 (2004).
- Spitkovsky, V. I. & Chang, A. X. (2011). Strong baselines for cross-lingual entity linking. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 14–15 November 2011.
- Spitkovsky, V. I. & Chang, A. X. (2012). A cross-lingual dictionary for English Wikipedia concepts. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 21–27 May 2012, pages 3168–3175.
- Stoyanov, V., Mayfield, J., Xu, T., Oard, D. W., Lawrie, D., Oates, T., & Finin, T. (2012). A context-aware approach to entity linking. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, Montreal, Quebec, Canada, 22–25 August 2013, pages 62–67.
- Strapparava, C., GlioZZo, A., & Giuliano, C. (2004). Pattern abstraction and term similarity for word sense disambiguation. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3) at ACL-04*, Barcelona, Spain, 25–26 July 2004, pages 229–234.
- Strötgen, J., Gertz, M., & Popov, P. (2010). Extraction and exploration of spatio-temporal information in documents. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, Zurich, Switzerland, 18–19 February 2010, pages 16:1–16:8.
- Strube, M. & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, Mass., 16–20 July 2006, pages 1419–1424.

- Suchanek, F., Kasneci, G., & Weikum, G. (2008). YAGO: A large ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*, 6(3):203–217.
- Sutton, C., McCallum, A., & Rohanimanesh, K. (2007). Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8:693–723.
- Tamang, S., Chen, Z., & Ji, H. (2013). CUNY-BLENDER TAC-KBP2012 entity linking system and slot filling validation system. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 5–6 November 2012.
- Tang, L.-X., Geva, S., & Trotman, A. (2012a). An English-translated parallel corpus for the CJK Wikipedia collections. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, Dunedin, New Zealand, 5–6 December 2012, pages 104–110.
- Tang, L.-X., Geva, S., Trotman, A., Xu, Y., & Itakura, K. (2011). Overview of the NTCIR-9 crosslink task: Cross-lingual link discovery. In *Proceedings of the 9th NTCIR Workshop Meeting*, Tokyo, Japan, 6-9 December 2011, pages 437–463.
- Tang, L.-X., Geva, S., Trotman, A., Xu, Y., & Itakura, K. Y. (2014). An evaluation framework for cross-lingual link discovery. *Information Processing & Management*, 50(1):1–23.
- Tang, L.-X., Kang, I.-S., Kimura, F., Lee, Y.-H., Trotman, A., Geva, S., & Xu, Y. (2013). Overview of the NTCIR-10 cross-lingual link discovery task. In *Proceedings of the 10th NTCIR Workshop Meeting*, Tokyo, Japan, 18-21 June 2013, pages 8–38.
- Tang, L.-X., Trotman, A., Geva, S., & Xu, Y. (2012b). Cross-lingual knowledge discovery: Chinese-to-English article linking in Wikipedia. In *Proceedings of the 8th Asia Information Retrieval Societies Conference*, Tianjin, China, 17–19 December 2012, pages 286–295.
- Tanskanen, S.-K. (2006). *Collaborating Towards Coherence. Lexical Cohesion in English Discourse*. John Benjamins, Amsterdam, The Netherlands/Philadelphia, Penn.
- Tjong Kim Sang, E. F. & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the 7th Conference on Computational Natural Language Learning*, Edmonton, Alberta, Canada, 31 May – 1 June 2003, page ??

- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Alberta, Canada, 27 May –1 June 2003, pages 252–259.
- Trivedi, R. & Eisenstein, J. (2013). Discourse connectors for latent subjectivity in sentiment analysis. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 9–14 June 2013, pages 808–813.
- Tufiş, D., Cristea, D., & Stamou, S. (2004). BalkaNet: aims, methods, results and perspectives. a general overview. *Romanian Journal on Science and Technology of Information. Special Issue on BalkaNet*, 7(1-2):9–43.
- Tufiş, D., Ion, R., & Ide, N. (2004). Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, 23–27 August 2004, pages 1312–1318.
- Turdakov, D. Y. & Lizorkin, S. D. (2009). HMM expanded to multiple interleaved chains as a model for word sense disambiguation. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, China, 3–5 December 2009.
- Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.
- Turney, P. D. & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.
- Ullmann, S. (1962). *Semantics: An Introduction to the Science of Meaning*. Oxford: B. Blackwell.
- van Gompel, M. (2010). UvT-WSD1: A cross-lingual word sense disambiguation system. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2)*, Uppsala, Sweden, 15–16 July 2010, pages 238–241.
- Van Langendonck, W. (2008). *Theory and Typology of Proper Name*. Berlin, Boston: De Gruyter Mouton.
- Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G. M., & E., M. E. (2005). Semantic similarity methods in WordNet and their application to information retrieval on the web. In *Proceedings of the Seventh ACM International Workshop on Web Information and Data Management (WIDM)*, Bremen, Germany, pages 10–16.

- Varma, V., Bysani, P., Kranthi Reddy, V. B., Santosh GSK, K. K., Kovelamudi, S., Kiran Kumar, N., & Maganti, N. (2009). IIIT Hyderabad at TAC 2009. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 16–18 November 2009.
- Veenstra, J., van den Bosch, A., Buchholz, S., Daelemans, W., & Zavrel, J. (2000). Memory-based word sense disambiguation. *Computers and the Humanities*, 34(1-2):171–177.
- Véronis, J. (2004). HyperLex: Lexical cartography for information retrieval. *Computer Speech and Language*, 18(3):223–252.
- Vincze, V., T., I. N., & Berend, G. (2011). Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, 14–16 September 2011, pages 289–295.
- Voorhees, E. (1999). Natural language processing and information retrieval. In *Information Extraction: Towards Scalable, Adaptable Systems*, pages 32–48, London, UK. Springer.
- Vossen, P. (Ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer, Dordrecht, The Netherlands.
- Wang, M., Smith, N. A., & Mitamura, T. (2007). What is the Jeopardy model? A quasi-synchronous grammar for QA. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning*, Prague, Czech Republic, 28–30 June 2007, pages 22–32.
- Weaver, W. (1955). Translation. In Locke, W. & Boothe, A. (Eds.), *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, Mass. Reprinted from a memorandum written by Weaver in 1949.
- Wentland, W., Knopp, J., Silberer, C., & Hartung, M. (2008). Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 26 May – 1 June 2008.
- Winkler, W. E. (2006). Overview of record linkage and current research directions. Technical report, Statistical Research Division, U.S. Census Bureau.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, Cal., 3rd edition.

- Wittgenstein, L. (2001). *Philosophische Untersuchungen. Kritisch-genetische Edition*. Frankfurt: Wissenschaftliche Buchgesellschaft.
- Yarowsky, D. (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 15th International Conference on Computational Linguistics*, Nantes, France, 23-28 August 1992, pages 454–460.
- Yarowsky, D. (1993). One sense per collocation. In *Proceedings of a Workshop on Human Language Technology*, Plainsboro, N.J., 21–24 March 1993, pages 266–271.
- Yarowsky, D. (1995). Unsupervised word sense disambiguation rivalling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, Mass., 26–30 June 1995, pages 189–196.
- Yarowsky, D. (2000). Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, 34(1-2):179–186.
- Yarowsky, D., Cucerzan, S., Florian, R., Schafer, C., & Wicentowski, R. (2001). The John Hopkins SENSEVAL-2 system descriptions. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2) at ACL-01*, Toulouse, France, 5–6 July 2001, pages 163–166.
- Yessenalina, A., Choi, Y., & Cardie, C. (2010). Automatically generating annotator rationales to improve sentiment classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11–16 July 2010, pages 336–341.
- Yu, D., Li, H., Cassidy, T., Li, Q., Huang, H., Chen, Z., & Ji, H. (2014). RPI-BLENDER TAC-KBP2013 knowledge base population system. In *Proceedings of the Text Analysis Conference*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 18–19 November 2013.
- Zesch, T., Müller, C., & Gurevych, I. (2008). Using Wiktionary for computing semantic relatedness. In *Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence*, Chicago, Ill., 13–17 July 2008, pages 861–867.
- Zhang, T., Liu, K., & Zhao, J. (2013). Cross lingual entity linking with bilingual topic model. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, Beijing, China, 3–9 August 2013, pages 2218–2224.
- Zhang, W., Su, J., Tan, C. L., & Wang, W. T. (2010). Entity linking leveraging automatically generated annotation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 23–27 August 2010, pages 1290–1298.

- Zheng, Z., Li, F., Huang, M., & Zhu, X. (2010). Learning to link entities with knowledge base. In *Proceedings of Human Language Technologies 2010: The Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, Cal., 2–4 June 2010, pages 483–491.
- Zhong, Z. & Ng, H. T. (2009). Word sense disambiguation for all words without hard labor. In *Proceedings of the 21th International Joint Conference on Artificial Intelligence*, Pasadena, Cal., 14–17 July 2009, pages 1616–1621.
- Zhong, Z. & Ng, H. T. (2012). Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea, 8–14 July 2012, pages 273–282.
- Zhou, Y., Nie, L., Rouhani-Kalleh, O., Vasile, F., & Gaffney, S. (2010a). Resolving surface forms to Wikipedia topics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 23–27 August 2010, pages 1335–1343.
- Zhou, Y., Guo, Z., Ren, P., & Yu, Y. (2010b). Applying Wikipedia-based explicit semantic analysis for query-biased document summarization. In *Proceedings of the 6th International Conference on Advanced Intelligent Computing Theories and Applications: Intelligent Computing*, Changsha, China, 18–21 August 2010, pages 474–481.
- Zirn, C., Niepert, M., Stuckenschmidt, H., & Strube, M. (2011). Fine-grained sentiment analysis with structural features. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, 8–13 November 2011, pages 336–344.