Inducing Implicit Arguments via Cross-document Alignment

A Framework and its Applications

Juni 2014

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

vorgelegt
von

Michael Roth

Betreuer und Erstgutachter:	Prof. Dr. Anette Frank Institut für Computerlinguistik Ruprecht-Karls-Universität Heidelberg
Zweitgutachter:	Prof. Dr. Michael Strube Institut für Computerlinguistik Ruprecht-Karls-Universität Heidelberg
Datum der Einreichung:	29. August 2013
Datum der Disputation:	3. Dezember 2013

Abstract

Natural language texts frequently contain related information in different positions in discourse. As human readers, we can recognize such information across sentence boundaries and correctly infer relations between them. Given this inference capability, we understand texts that describe complex dependencies even if central aspects are not repeated in every sentence. In linguistics, certain omissions of redundant information are known under the term *ellipsis* and have been studied as cohesive devices in discourse (Halliday and Hasan, 1976). For computational approaches to semantic processing, such cohesive devices are problematic because methods are traditionally applied on the sentence level and barely take surrounding context into account.

In this dissertation, we investigate omission phenomena on the level of predicate-argument structures. In particular, we examine instances of structures involving arguments that are not locally realized but inferable from context. The goal of this work is to automatically acquire and process such instances, which we also refer to as *implicit arguments*, to improve natural language processing applications. Our main contribution is a framework that identifies implicit arguments by aligning and comparing predicate-argument structures across pairs of comparable texts. As part of this framework, we develop a novel graph-based clustering approach, which detects corresponding predicate-argument structures using pairwise similarity metrics. To find discourse antecedents of implicit arguments, we further design a heuristic method that utilizes automatic annotations from various linguistic pre-processing tools.

We empirically validate the utility of automatically *induced* instances of implicit arguments and discourse antecedents in three extrinsic evaluation scenarios. In the first scenario, we show that our induced pairs of arguments and antecedents can successfully be applied to improve a pre-existing model for linking implicit arguments in discourse. In two further evaluation settings, we show that induced instances of implicit arguments, together with their aligned explicit counterparts, can be used as training material for a novel model of local coherence. Given discourse-level and semantic features, this model can predict whether a specific argument should be explicitly realized to establish local coherence or whether it is inferable and hence redundant in context.

Kurzfassung

Texte in natürlicher Sprache enthalten häufig Informationen, die miteinander in Beziehung stehen, gleichzeitig aber über den gesamten Text verteilt sind. Als Menschen können wir solche Informationen auch über Sätze hinweg erkennen und deren Zusammenhänge inferieren. Aufgrund dieser Auffassungsgabe verstehen wir Texte über komplexe Sachverhalte auch ohne dass wesentliche Aspekte in jedem Satz wiederholt werden müssen. In der Linguistik sind bestimmte Vorkommen solcher Auslassungen auch als *Ellipse* bekannt und werden als Mittel zur Textverknüpfung untersucht (Halliday and Hasan, 1976). Für maschinelle Ansätze der semantischen Sprachverarbeitung sind solche Kohäsionsmittel jedoch problematisch, da Methoden traditionell auf der Satzebene arbeiten und Kontext kaum berücksichtigen.

In dieser Dissertation beschäftigen wir uns mit dem Fall von semantischer Sprachverarbeitung auf der Ebene von Prädikat-Argument-Strukturen. Wir betrachten dabei insbesondere Instanzen solcher Strukturen, in denen nicht alle Argumente innerhalb eines Satzes realisiert wurden, diese aber durch den Kontext inferiert werden können. Das Ziel der Arbeit ist die automatische Gewinnung und Verarbeitung solcher Argumente, welche wir im Folgenden auch *implizite Argumente* nennen. Zur automatischen Gewinnung impliziter Argumente entwickeln wir ein Framework, in dem Paare von Prädikat-Argument-Strukturen über monolinguale Vergleichstexte hinweg aligniert und fehlende Argumente durch einen Abgleich zweier Strukturen erkannt werden. Wir erarbeiten speziell für diese Anwendung ein graph-basiertes Clustering-Verfahren, welches korrespondierende Strukturen in Paaren von Texten erkennt und miteinander verbindet. Darüber hinaus entwerfen wir einen heuristischen Ansatz, um auf Basis verschiedener Vorverarbeitungsschritte automatisch Diskurs-Antezedenten für implizite Argumente zu finden.

Wir validieren den Nutzen automatisch *induzierter* Instanzen impliziter Argumente und Antezedenten empirisch in drei Evaluations-Szenarien: in der ersten Evaluation zeigen wir, dass automatisch gefundene Argument-Antezedent-Paare aus Vergleichstexten genutzt werden können, um ein bestehendes System zum Verlinken impliziter Argumente zu verbessern; in zwei weiteren Evaluationen zeigen wir, dass auf Basis des automatisch erzeugten Datensets ein Kohärenzmodell gelernt werden kann, das auf Basis von semantischen und kontextuellen Faktoren voraussagt, ob ein bestimmtes Argument an einer gegebenen Textstelle realisiert werden sollte, um Kohärenz zu erzeugen, oder ob es aus dem Kontext bereits inferierbar ist.

Contents

Part I.

Introduction and background

1. Introduction

The goal of semantic parsing is to automatically process natural language text and map the underlying meaning of text to appropriate meaning representations. Towards this goal, *semantic role labeling* aims at inducing shallow semantic representations, so-called predicate-argument structures, by processing sentences and mapping (sequences of) words that they contain to predicates and associated arguments. Applying a semantic role labeling system on the sentence "Nicaragua withdrew its troops last month" would, for example, result in a semantic representation that consists of the predicate "withdraw", a temporal modifier ("last month") and two associated arguments: the entity withdrawing ("Nicaragua") and the thing being withdrawn ("its troops"). Semantic role labeling systems process text on the sentence level and hence only induce local structures that represent meaning aspects of the sentence. Information relevant to these structures, however, can be non-local in natural language texts. For instance, consider the two sentences in Example (1):

(1) a. El Salvador is now the only Latin American country which has troops in [Iraq].

b. Nicaragua withdrew its troops last month.

Sentence (1b) is the same as discussed in the example before. When looking at the discourse context, however, we can now see that there is one more argument realized: namely Iraq, the source from which Nicaragua withdrew its troops. We refer to such arguments, which do not occur within the local structure of the predicate itself, as being *implicit* or *non-local*. In the given example, a human reader can easily infer that "Iraq", from Sentence (1a), is an implicit argument of the predicate in Sentence (1b). In contrast, computationally modeling this inference step is difficult as it involves an interplay of two challenging sub-tasks: a semantic parser has to determine that an argument is not locally realized (but inferable), and a suitable reference (antecedent) has to be found within the discourse context. We refer to these steps as *identifying* and *linking* implicit arguments to discourse antecedents. We particularly address the second step in this thesis and describe ways to improve statistical models for it by automatically inducing suitable training data. As illustrated by Example (1), models for linking implicit arguments in discourse are essential to understand the full meaning of natural language texts computationally. Successful approaches to this task could hence be useful to improve the performance of natural language processing systems in applications such as question answering and information extraction. We discuss semantic role labeling, as a basis for this endeavor, in more detail in Section 1.1.

One of the reasons why implicit arguments occur in natural language text is that each sentence in a discourse focuses only on a set of entities that are *salient* at the

specific point. When viewing the salience of an entity as measurements on a scale, non-realizations can be explained by both possible extremes: at one end of the scale, entities can be non-salient at the current point in discourse. For example, the sentence "John is still eating" expresses that someone is eating something. The type of food being eaten, however, might be irrelevant in context and, in fact, there might be no explicit reference to it in discourse at all. At the other end of the scale, entities can be highly salient and hence be understood implicitly, for example, because they are directly inferable from the immediate context. Example (1) illustrated such a case. Mentioning "Iraq" in the first and second sentence here is simply not necessary (for a human being) to understand the meaning of the text. In contrast, making both references explicit, as shown in Example (2), would be redundant and could lead to the perception that the text is merely a concatenation of two independent sentences – rather than a set of adjacent sentences that form a meaningful, or *coherent*, discourse.

(2) a. El Salvador is now the only Latin American country which has troops in [Iraq].

b. Nicaragua withdrew its troops from [Iraq] last month.

In natural language generation (NLG), the task of *referring expression generation* is to generate appropriate descriptions for entities in a way such that a reader can identify them in context. Traditionally, the task is formulated as a decision process between proper name use, pronouns and definite descriptions (including the choice of distinct or salient properties). As seen in Example (1), however, entities can also be understood without making use of explicit mentions. This aspect has mostly been ignored in previous work on NLG, even though it could be beneficial to improve the quality of automatically generated texts. In this thesis, we address this shortcoming by developing a model for predicting whether an explicit entity reference would contribute to the coherence of a text, or whether it would be redundant because the entity can be inferred from context. We discuss referring expression generation in more detail in Section 1.2.

As indicated in the previous paragraphs, the phenomenon of implicit arguments has neither been extensively studied in context of semantic parsing nor in text generation. One of the main reasons for this lies in the fact that models for these tasks are typically developed on the basis of annotated corpora. In contrast, there are only few and small data sets available, in which implicit arguments and their antecedents are explicitly marked. In this thesis, we present a novel approach to inducing training data that contains automatic annotations of implicit arguments and their respective antecedents in discourse. To achieve this goal, our framework exploits pairs of *comparable texts*, which convey information about the same events, states and entities. The methods developed in this thesis are inspired by previous work on inducing semantic resources using cross-document methods (cf. Section 1.3). In our work, we show that instances of implicit arguments and discourse antecedents can be induced from comparable texts, given automatic annotations on local semantic role labels and entity coreference chains. We empirically validate the utility of the induced data set in extrinsic evaluations, where we show how the data can be used to train new models that can enhance semantic role labeling and coherence assessment. The resulting models and research insights will be of

importance for any application that involves the understanding or generation of natural language text beyond the sentence level. An overview of this thesis and details regarding our contributions are described in Section 1.4.

1.1. Semantic Role Labeling

The goal of semantic role labeling is to automatically induce predicate-argument structures that represent a shallow analysis of the meaning of a natural language sentence. In the context of this thesis, we define a predicate as a lexical item, which can be a single word or a phrase, that expresses a relation between entities or properties, the so-called arguments of the predicate. Each predicate naturally comes with a predefined set of participants (or properties), its *semantic roles*. Lexicons that define the *role sets* of a predicate have become increasingly available in the past two decades. The two most prominent examples for English are developed as part of the *FrameNet* project at the International Computer Science Institute (ICSI) in Berkeley (Ruppenhofer et al., 2010a) and the *PropBank* project at the University of Colorado at Boulder (Palmer et al., 2005). We briefly describe the general idea behind both projects in the next paragraphs. A more comprehensive overview of semantic role labeling approaches can be found in (Palmer et al., 2010).

FrameNet. Following the project description by Ruppenhofer et al. (2010a), FrameNet is a lexicon resource for English, based on *frame semantics* and supported by corpus evidence. Frame semantics, following Fillmore (1976), is a theory of meaning that emphasizes the close relation between language and experience. According to Fillmore, "particular words or speech formulas ... are associated in memory with particular frames", which can be defined as schematic representations of events, relations and entities. Each *semantic frame* in FrameNet describes such a scheme, together with its *frame elements*, the participants and properties of a frame. For example, the frame TEXT_CREATION represents a prototypical situation, in which an `author` creates a `text`. In addition to essential participants in a frame, the so-called *core roles*, additional properties can be expressed using *non-core roles*, for example, the `instrument` used for text creation. Linguistically, instances of the TEXT_CREATION frame can be expressed, or *evoked*, by words and phrases such as "text" or "type in", which are called the *lexical units* of a frame. While lexical units can correspond to any part of speech, most units in the current version of FrameNet[1] are nouns (5,177; 40.7%) and verbs (4,877; 38.4%). Using frames and frame elements for representing aspects of the meaning of a sentence, we can illustrate such structures in text as shown in Example (3):

(3) "[John]$_{\texttt{author}}$ drafted$_{\text{TEXT_CREATION}}$ [his thesis]$_{\texttt{text}}$ [with pen and paper]$_{\texttt{instrument}}$."

PropBank. In contrast to FrameNet, predicates and semantic roles in PropBank (Palmer et al., 2005) are represented on a level of abstraction that is close to their syntactic re-

[1] cf. `http://framenet.icsi.berkeley.edu/fndrupal/current_status` (updated Aug 22, 2013)

A0	`agent`
A1	`patient`
A2	`beneficiary` / `instrument` / `attribute` / `end state`
A3	`start point` / `beneficiary` / `instrument` / `attribute`
A4	`end point`

Table 1.1.: Annotation scheme for mapping argument labels to semantic roles, according to the "Guidelines for Propbank framers".[2]

alization in text. Instead of grouping related concepts in frames, each predicate in PropBank is defined by its respective word senses. Each predicate sense, in turn, comes with its own role set that defines the *arguments* of a predicate. The role set of a predicate depends on its usage in natural language and is based on syntactic constituents. For example, the predicate TYPE (in the sense of typing up) typically occurs with two arguments: a writer (subject) and the text being written (object). In PropBank, all arguments are labeled in a numerical order, starting with zero. The first two arguments, A0 and A1, are reserved for the proto-agent and proto-patient of a predicate (Dowty, 1991), respectively. Other numbered arguments vary across semantic classes, following the scheme in Table 1.1, and are assigned to all syntactic constituents that are "required for the valency of a predicate ... or that occur with high-frequency in actual usage" (Bonial et al., 2012). In contrast to FrameNet, where roles are defined on the basis of frames, the role sets of related predicates in PropBank might or might not overlap. For example, the arguments of BUY and SELL do not correspond to each other due to a different mapping to semantic roles. That is, the `buyer` corresponds to the agent (A0) in one case but to the patient (A1) in the other case. Following the **NomBank** extension to PropBank (Meyers et al., 2008), role sets of verbal predicates are also used to label arguments of nouns that are derived from a verb. For example, the predicate SALE uses the same role set as the predicate SELL. In addition to arguments, PropBank defines a fixed set of *modifiers* that capture properties, such as location (`LOC`), time (`TMP`) and manner (`MNR`), that are not specific to a single predicate. For example, in the text "he wrote the letter with a pen", "with a pen" would not be treated as a argument of WRITE but rather as a modifier that expresses how the writing is being done. When using the notation of PropBank, we illustrate predicate-argument structures as shown in Example (4):

(4) "[John]$_{\text{A0}}$ <u>typed up</u> [his entire dissertation]$_{\text{A1}}$ [in one weekend]$_{\text{TMP}}$."

In this thesis, we view FrameNet and PropBank as complementary paradigms and make use of advantages of both resources: in particular, we adopt the PropBank paradigm for semantic role labeling and we make use of FrameNet frames to relate arguments of different predicates to one another. The first decision is based on the pragmatic insight that PropBank-based parsers achieve a higher precision and better coverage than state-of-the-art frame-semantic analysers. Though a direct comparison of parsers from

[2] cf. `http://verbs.colorado.edu/~mpalmer/projects/ace/FramingGuidelines.pdf`

both paradigms is difficult (due to the use of different corpora), the overall difference in numbers for their performance on full-text labeling of predicates and arguments is substantial: while current PropBank-based parsers (Choi and Palmer, 2011; Zhao et al., 2009) achieve a precision and recall of up to 87% and 84%, respectively, the precision and recall figures for the best performing FrameNet-based parser (Das and Smith, 2011) lie around 71% and 66%, respectively.

Current state-of-the-art systems do, however, barely take into account arguments that are realized beyond the sentence level. One of the key problems is that annotated data for this task is scarce. In this thesis, we propose to bridge this gap by inducing instances of implicit arguments, with links to their discourse antecedents, from pairs of comparable texts. We do so by first applying a PropBank-based system on sentences from the two texts and, secondly, by merging corresponding predicate-argument structures, which can be partially overlapping, across documents. We demonstrate that instances of implicit arguments, induced by this kind of approach, can be mapped to FrameNet and applied to improve an existing model for the task of "linking events and their participants in discourse". We discuss this task in more detail, together with the role of implicit arguments in previous work, in Chapter 2.

1.2. Referring Expression Generation

Systems for natural language generation (NLG) are traditionally designed as pipeline architectures (Reiter and Dale, 2000). Given units of information to be realized in a text, the goal of referring expression generation is to produce references to entities that will occur in the generated text. The produced expressions should make it possible for the reader to easily understand who or what a specific phrase refers to. Many algorithms in NLG focus on generating expressions that make the referred entity distinguishable from other entities in a specific context, for example, by listing salient or distinct properties (Dale, 1992; Dale and Reiter, 1995; Krahmer et al., 2003, inter alia). As pointed out by Viethen and Dale (2006), however, all of these algorithms produce just one deterministic description of an entity while a single person may use various different expressions to refer to the same entity within one discourse. Taking this observation into account, Belz and Varges (2007) ask the question of how multiple references to the same entity should be realized in context. This question formed the basis for a NLG task on discourse-based referring expression generation, which had been organized as an annual challenge in the following years (Belz et al., 2008, 2009; Belz and Kow, 2010b). Given a piece of text, the task is to improve "referential clarity and coherence" by postprocessing referring expressions in context. In all challenges, the only type of implicit reference considered is that of referring expressions in subject positions, where the referring expression is the subject of multiple coordinated verb phrases, as illustrated in Example (5). Apart from this exception, only explicit references are subject to evaluation.

(5) a. [He] stated the first version of the Law of conservation of mass, [∅] introduced the Metric system, and [∅] helped to reform chemical nomenclature.

As shown in Example (5), the antecedent of implicit references in coordinated verb phrases can still be retrieved within the sentence. In contrast, we discussed in the beginning of this thesis that explicit mentions are not always necessary for a reader to understand the entities that are referred to in a sentence. To the best of our knowledge, the only previous study that took this factor into consideration was carried out by Zarrieß and Kuhn (2013). In their work, they investigate NLG architectures that perform referring expression generation and surface realization. More specifically, they examine architectures that are able to produce syntactic structures, such as passives and coordinations, in which generated references can either be inserted or omitted. Within the limits of their study, they show that realization decisions related to implicit references can be modeled with up to 85% accuracy using only few contextual features: the last mention of the affected entity, its realization in the header of the text, and the role and realization of the closest preceding reference in text. Their study relies, however, on manual annotations of referents and their evaluation is restricted to annotated instances of core roles of the FrameNet ROBBERY frame in a German newspaper corpus.

In this thesis, we propose an unrestricted setting, in which we examine the question of whether an entity needs to be explicitly mentioned at a specific point in discourse or whether it can also be inferred from context. We argue that redundant entity mentions can have a negative impact on the perceived coherence of a discourse, whereas the realization of non-redundant mentions is necessary to establish coherence. Our model for this task is hence related to previous approaches to entity-based coherence modeling. Following the evaluation scheme of previous NLG challenges, we design our model and the evaluation setting independent of a particular NLG system or application. That is, we present to the model multiple possibilities of what a potentially coherent text may look like, let the model compute scores for each candidate and select the one that is predicted to maximize the coherence of the discourse. In our evaluation, we contrast this approach to previous models of local coherence. We introduce these models, together with theoretical background of entity-based coherence modeling, in Chapter 2.

1.3. Semantic Resource Induction

The goal of this thesis is to improve models for natural language processing by taking into account entity references in the form of implicit arguments. As discussed in the previous sections, affected tasks can be found both in the area of semantic parsing and in the field of text generation. Both lines of research suffer, however, from the scarce amount of available data to extend existing models to implicit arguments: semantic role labeling models are traditionally trained and evaluated on a sentence-by-sentence basis, making it impossible for them to cover arguments that are non-local and only inferable from context; similarly, NLG systems are able to generate various kinds of explicit referring expressions but they typically do not take into consideration that realizations can be redundant, and hence be omitted, in contexts where a reference is understood implicitly. In this thesis, we aim to overcome the lack of suitable training resources by automatically inducing a data set that contains annotations of implicit arguments.

Data for leveraging the potential of implicit arguments in the two aforementioned tasks has to meet certain requirements: to identify parts of a predicate-argument structure that are realized non-locally, we need training instances that involve implicit arguments and their discourse antecedents; to determine whether references can be understood implicitly, we need training data that provides contexts for instances of local and non-local realizations alike. In this thesis, we aim to induce a semantic resource that provides such contexts. We propose an approach that exploits pairs of texts that describe the same events but are published by different news agencies, thus providing differing contexts. Inducing data from pairs of texts has previously been shown useful for creating other kinds of semantic resources. For example, texts with comparable content have been used to detect lexical synonyms and paraphrases (Cohn et al., 2008, inter alia), and to bootstrap tools for semantic role labeling (Titov and Kozhevnikov, 2010). Texts that are available in multiple languages have further been used to learn potential translations of words and phrases (Kay and Röscheisen, 1993; DeNero et al., 2008, inter alia) and to create semantic lexicons in one language – given that such a resource already exists in another language, from which it can be "transferred" or "projected" (Padó and Lapata, 2009; Kozhevnikov and Titov, 2013, inter alia). In Chapter 3, we provide a more detailed overview of methods used in previous work and discuss a range of their applications.

1.4. Thesis Overview and Contributions

As discussed in the beginning of this introduction, implicit arguments are an important aspect for full natural language understanding, yet they are not covered by traditional semantic role labeling systems. Similarly, referring expression generation is crucial for generating coherent texts in natural language but the possibility of choosing implicit references has widely been ignored in previous work. Two potential reasons why implicit arguments received little attention in previous research lie in the scarcity of data annotated with non-local arguments and in the inherent difficulty of inferring antecedents of implicit arguments automatically. We address the two problems in this thesis by proposing an induction approach that automatically identifies implicit arguments together with their discourse antecedents from comparable texts. We then show how induced implicit arguments can be utilized as training data for semantic parsing and coherence modeling. The thesis is divided into three parts: in the first part, we discuss related work on semantic parsing, coherence modeling and cross-document methods; in the second part, we develop methods for the automatic induction of implicit arguments and discourse antecedents; in the third part, we demonstrate the utility of automatically induced data for semantic parsing and coherence modeling.

Background. In Chapter 2, we discuss previous research on implicit arguments in semantic parsing and on entity-based coherence modeling. Both lines of research represent the starting point for our work: research in semantic role labeling has shown that linking implicit arguments is a challenging task for current state-of-the-art systems; coherence models, based on entity realizations, have been successfully applied to rank alternative

versions of generated texts, yet they do not take into account implicit entity references. The goals of this thesis are to provide data and methods to better capture the phenomenon of implicit arguments in both tasks. To accomplish this goal, we propose to automatically induce training data from pairs of texts. This kind of training data shall contain explicit links between implicit arguments and their respective discourse antecedents (to enhance models for linking implicit arguments); furthermore, the comparable texts shall provide discourse contexts for explicit and implicit references to the same entity (for training a suitable coherence model). The idea of automatically constructing such a resource builds on previous work on parallel and comparable texts, which we outline in Chapter 3.

Induction framework. In the second part of this thesis, we develop a novel approach to inducing training data with automatic implicit argument annotations. In this approach, we make use of comparable texts. That is, pairs of texts that convey information about the same events, states and entities. The question underlying this approach is:

Q1: How can implicit arguments be induced from comparable texts?

The motivation for asking this question is that comparable texts can contain the same information expressed in various different ways. We expect that these differences also affect the use of implicit and explicit arguments. That is, a reference to a specific entity might be understood implicitly in one text (because it can be inferred from context), while an explicit reference to the same entity might be necessary in a comparable text (given the different context). Based on this assumption, our method aims at finding complementary (explicit) information in pairs of texts, which can be aligned and merged to detect missing (implicit) pieces in one another. Our approach, which we outline in more detail in Chapter 4, can be summarized as a multi-step framework, in which we break down Q1 into three intermediate questions, starting with:

q_i: How can we automatically identify pairs of comparable texts?

To this end, we construct a large corpus of comparable texts from a resource that contains billions of newswire articles. We present our approach to this problem and the resulting data set in Chapter 5. A challenging aspect of comparable texts is that while they overlap in information, they can significantly differ in their perspective on the described information and in the amount of details that they convey. Hence our next intermediate question is:

q_{ii}: How can we detect information that is shared across two texts?

Based on the constructed corpus, we propose a new task to answering this question: "aligning predicate-argument structures across comparable texts". We outline this task as well as a novel graph-based clustering technique to tackle it in Chapter 6. We show that by working on the level of predicate-argument structures, our model performs well both on parallel and comparable texts. Based on aligned structures from pairs of text, we

finally seek to answer question Q1 by distinguishing between two types of arguments that are realized in one text but not in the other: the first type corresponds to entities that are completely missing from one text (non-realized); the other type covers all entities that are implicit at a specific position in discourse but inferable from the textual context (locally unrealized). In other words, the two types differ in that there exists a discourse antecedent in one case but not in the other. So the last question to answer Q1 becomes:

q_{iii}: How can we find antecedents for implicit arguments within their discourse context?

To answer this question, we propose an induction approach that detects implicit arguments together with their discourse antecedents. The approach relies on information that can automatically be extracted from pairs of comparable texts: aligned predicate-argument structures and entity coreference chains. More precisely, we make use of aligned predicate-argument structures (PAS) and look for entities, for which we can detect mentions in one PAS (explicit argument) but not in the aligned structure (implicit argument). Given such cases, we apply a cross-document coreference resolution technique to also find co-referring entity mentions in the document in which the entity is implicit in the aligned structure. We describe how to perform this task computationally, together with an intrinsic evaluation of its performance, in Chapter 7.

Applications and further directions. In the third part of this thesis, we demonstrate the utility of automatically induced implicit arguments for semantic parsing and coherence modeling. We ask the following two questions:

Q2: How can we employ induced implicit arguments to improve existing SRL models?

and

Q3: How can we predict coherent realizations of arguments in discourse context?

We address both questions in task-based settings in which we make use of our automatically induced implicit arguments as training data for statistical learning models. For the first task, we apply our data to enhance training of an existing system that tries to identify and link implicit arguments in discourse. To evaluate the impact of our induced data set on this task, we test the modified model on a standard evaluation data set, on which we can compare our results with those of previous work. For the second task, we develop a new coherence model that predicts whether an argument realization or non-realization in context would improve the perceived coherence of the affected segment in discourse. We evaluate this coherence model in two tasks: the first is an intrinsic evaluation scenario, in which we compare model predictions to human judgments on argument use; the second task is an extrinsic evaluation scenario, in which we apply our new model to post-processing automatically generated summaries. All tasks, data sets and results are described in detail in Chapter 8.

We summarize our contributions in Chapter 9, together with a discussion on potential benefits for other lines of research. Finally, we conclude this thesis in Chapter 10 with some final remarks on the phenomenon of implicit arguments and how, in our view, it should be treated in future NLP applications.

2. Implicit Arguments

As outlined in Chapter 1, semantic role labeling systems traditionally process texts in a sentence-by-sentence fashion, inducing local semantic structures that represent aspects of the meaning of a sentence. We discussed that arguments, however, can be non-local in natural language texts. Covering this phenomenon is hence necessary for full natural language understanding. In Section 2.1, we describe previous work on the role of non-local arguments in shallow semantic parsing. As discussed in Section 1.2, omitting entity references that are redundant in context is also essential for generating coherent and natural sounding texts. In Section 2.2, we discuss the effect of entity references on local coherence and how previous work addressed this phenomenon in context of entity-based coherence modeling.

2.1. Implicit Arguments in Semantic Parsing

In this section, we discuss the role of implicit arguments in shallow semantic parsing. We focus our discussion on semantic role labeling (SRL) approaches as defined in Section 1.1. In Section 2.1.1, we introduce first analyses that represent the ground work for computational models of implicit arguments. In Section 2.1.2, we describe a systematic effort to include implicit arguments in noun-based SRL. We discuss an annotated data set that was released in context of a shared task in 2010 in Section 2.1.3. This data set has been the basis for a range of recent approaches to model the identification and linking of implicit arguments. We describe one of these approaches in detail in Section 2.1.4 and briefly discuss other developments and current directions in Section 2.1.5.

2.1.1. Early Work

Two of the most prominent projects that develop lexicon resources for semantic role labeling are FrameNet and PropBank (cf. Section 1.1). In both projects, instances of predicates and their associated argument structure are being annotated in English text corpora. In PropBank, only arguments are annotated that are part of the same sentence that contains the considered predicate instance. In FrameNet, omitted arguments that correspond to a core role are explicitly marked as missing, or *null-instantiated*, following early work by Fillmore (1986). The discourse antecedents, to which missing but inferable arguments refer, are not consistently annotated though. Based on frame-semantic analysis, Burchardt et al. (2005) perform a small-scale study, in which they show that some of these implicit arguments can be inferred through relations between frames and frame elements that are instantiated in a text. Some alternative approaches to resolve

implicit arguments have been proposed by Palmer et al. (1986), Whittemore et al. (1991) and Tetreault (2002). We describe all of these approaches in more detail below.

Fillmore (1986). In his 1986 paper, Fillmore builds upon previous work in linguistics in which aspects of implicit arguments have been studied under various different names, including "Unspecified NP Deletion" (Fraser and Ross, 1970), "Definite Object Deletion" (Mittwoch, 1971), "Latent Object" (Matthews, 1981), and "Contextual Deletion/Suppression" (Allerton, 1982). In contrast to previous studies, Fillmore's analysis takes into account aspects on the levels of pragmatics as well as lexical semantics and is not restricted to specific grammatical functions or categories. In his work, Fillmore refers to missing elements in text as *null complements* and distinguishes between two kinds of instances. The first kind, *indefinite null complements* (INC), comprises cases in which a missing element is "unknown or a matter of indifference". In the second case, that of *definite null complements* (DNC), the missing element must be retrievable in the given context. We present one instance for each type of null complement in Example (6):

(6) [I]$_{\texttt{donor}}$ already contributed$_{\text{GIVING}}$ [INC]$_{\texttt{theme}}$ [DNC]$_{\texttt{recipient}}$.

When reading the sentence in (6), we would generally assume that there exists both a gift (`theme`) and a receiver (`recipient`) of the contribution action that is being referred to. The difference between the two missing elements is, according to Fillmore, that the receiver should be inferable from the context, in which the sentence was uttered. In contrast, Fillmore argues that it is not necessary to have a "shared advance understanding" of the identity or nature of the gift itself. Without context, it would hence sound odd to hear or read the sentence in Example (7), in which the `recipient` of a GIVING event is not specified.

(7) [I]$_{\texttt{donor}}$ contributed$_{\text{GIVING}}$ [5 Dollars]$_{\texttt{theme}}$ [DNC]$_{\texttt{recipient}}$.

Conversely, the sentence in Example (8), in which only the `theme` is unspecified, would not sound odd:

(8) [I]$_{\texttt{donor}}$ contributed$_{\text{GIVING}}$ [to the British Heart Foundation]$_{\texttt{recipient}}$ [INC]$_{\texttt{theme}}$.

As indicated by Fillmore's definition, DNCs can only be used when the affected argument can potentially be inferred from the given context. In this thesis, we focus on a special case of these DNCs, namely instances, in which the affected argument is not only inferable but also explicitly realized elsewhere in context. To avoid confusion with other types of instances, we use the term "implicit argument" instead of adopting the terminology by Fillmore.

Palmer et al. (1986). To the best of our knowledge, the first approach to (automatically) identify discourse antecedents of implicit arguments is that used in the text understanding system PUNDIT (Palmer et al., 1986). In PUNDIT, Palmer et al. propose

a combination of syntactic, semantic and pragmatic modules for this task. Following earlier suggestions by Fillmore (1969, 1986), Palmer et al. use syntactic and semantic information to identify implicit arguments. These include, for example, missing subjects and objects of verbs that can be transitive and intransitive. Given an identified implicit argument, a suitable antecedent is determined in the second step via reference resolution. Candidate antecedents are selected from a "focusing list" that contains discourse entities from the previous context. The order of the focusing list is inspired by previous work on *focusing* and *centering* (cf. Section 2.2) and takes into account factors such as pronoun references and syntactic constituent types (Dahl, 1986). Based on these factors, PUNDIT employs heuristics that reflect how likely an entity is to become the focus of the following sentence or, in other words, in which order the entities should be considered as antecedents of an implicit argument.

Whittemore et al. (1991). A few years later, Whittemore et al. (1991) proposed to use Discourse Representation Theory (DRT; Kamp, 1981) to build meaning representations of events incrementally. While they work out a full framework within DRT, they neither provide a computational account for resolving implicit arguments nor do the perform any sort of evaluation. Like Palmer et al. (1986), they do note, however, that possible referents should be ranked according to their "forward focusing character".

Burchardt et al. (2005). In their work, Burchardt et al. present a small case study, in which they apply an event building paradigm to a short text from Wikipedia. Instead of using Discourse Representation Theory, however, they make use of frame semantics, as proposed in earlier work by Fillmore and Baker (2001). More precisely, Burchardt et al. suggest to link implicit arguments across frame instances by performing inference over semantic roles that are co-referential or semantically related. Text (9) shows an example fragment from Burchardt et al., which we slightly modified to be consistent with the current version of FrameNet:

(9) a. (...) the [Higher Regional Court of Hamburg]$_{\texttt{court}}$ has passed down the [maximum sentence$_{\text{SENTENCING}}$]$_{\texttt{type}}$.

b. [Mounir al Motassadeq]$_{\texttt{prisoner}}$ will serve$_{\text{BEING_INCARCERATED}}$ [15 years]$_{\texttt{duration}}$ [in prison]$_{\texttt{prison}}$.

In (9a), the lexical unit "sentence" evokes a SENTENCING frame with the core frame elements `convict`, `court`, `sentence`, `term_of_sentence` and `offense`. Only `court`, however, is realized in the shown fragment. By noting the close relationship between the SENTENCING frame in the first sentence and BEING_INCARCERATED in the following sentence, two non-local arguments can be inferred: firstly, the `convict` in (9a) should be coreferent with the `prisoner` in (9b); and secondly, the `term_of_sentence` should be identical to the `duration` for which the prisoner will remain in jail. While Burchardt et al. outline general ideas of how this inference process can be automated, they perform no evaluation of such a model themselves. Some of their suggestions are, however,

implemented in the form of features in the models of Gerber and Chai (2012) and Silberer and Frank (2012). We describe these two models in more detail in Section 2.1.2 and Section 2.1.4, respectively.

Tetreault (2002). A first small-scale data set, on which automatic techniques to link implicit arguments can be evaluated, was released by Tetreault (2002). The data set is a transcription of one dialog and comprises 86 sentences. The annotated transcription contains a total of 62 annotated instances of implicit arguments, distributed over 14 different verb types related to "moving and loading of food and trains". Tetreault develops and evaluates an approach to resolve implicit arguments, making use of a focusing algorithm similar to the one proposed by Palmer et al. (1986). Instead of assessing for each candidate antecedent whether it would be a suitable argument, Tetreault uses multiple focusing lists, each for one out of four different types of roles: `instrument`, `theme`, `from-loc` and `to-loc`. While Tetreault discusses that his algorithm seems to work well for three out of the four roles, he concludes that "a more extensive corpus is needed to confirm this claim". We discuss two notable attempts to annotate implicit arguments in larger corpora and richer domains in the next two sections.

2.1.2. Gerber & Chai (2009–2012)

As part of the CoNLL Shared Task in 2008, Surdeanu et al. (2008) organized an evaluation of semantic parsers following the PropBank/NomBank annotation paradigm (Palmer et al., 2005; Meyers et al., 2008). In contrast to previous evaluations, the training and test data contains not only predicates with argument structures but also annotations for (nominal) predicates that occurred without any local arguments. The predicate DISTRIBUTION in Example (10) represents one such instance:

(10) The distribution represents available cash flow between Aug. 1 and Oct. 31.

Gerber et al. (2009) found that by taking gold annotations on such predicates into account during evaluation, performance of SRL systems decreases by more than 9 percentage points in F_1-score. This outcome shows that predicates without local arguments seem to be more difficult to recognize than others. Following these observations, Gerber and Chai (2010) annotate non-local arguments and take them into account in their model. The annotation effort in their work, however, is restricted to 10 nominal predicate types in the Penn TreeBank (Marcus et al., 1993). In total, they annotate 246 instances of implicit arguments. While this number seems fairly low, Gerber and Chai extend their original annotation in follow-up work (Gerber and Chai, 2012), increasing the number of instances of implicit arguments for the 10 nominal predicates up to 966. Based on this data set, they develop a log-linear model for linking implicit arguments in discourse. Their model makes use of a broad range of features, which are grouped into six categories: 1. argument labels, 2. features from a manually constructed ontology, 3. properties of the missing argument, 4. corpus statistics on predicates and arguments, 5. discourse relations, and 6. "other" features, including for example, the distance between the affected predicate and candidate antecedent.

The evaluation of the model is performed using gold standard annotation from the Penn TreeBank (syntactic parses), PropBank and NomBank (local semantic arguments). Given three-tuples of predicate, missing argument and candidate antecedent, the evaluation task is to determine whether the candidate is a correct antecedent for the argument that is missing. In their experiments, Gerber and Chai show that the log-linear model can predict the correct antecedent in 44.5% of all cases, with an average precision of 57.9%, resulting in an overall F_1-score of 50.3%. They perform additional experiments to assess the impact of training data size and each feature group. In the first set of these experiments, Gerber and Chai observed improvements for increasing training data size but the performance seems to stagnate at around 80% of the overall data available. In the latter experiments, they found significant losses in performance when excluding features based on argument labels, including valuable indicators expressing information such as "the A0 of LOSE is the A0 of INVEST". The importance of this feature group is particularly interesting as it closely resembles the idea of co-referring and related semantic roles by Burchardt et al. (2005) (cf. Section 2.1.1).

2.1.3. SemEval 2010 Task 10

Another data set of implicit arguments was released as part of the SemEval 2010 shared task on "Linking Events and Participants in Discourse" (Ruppenhofer et al., 2010b). In contrast to the data set by Gerber and Chai, annotation of semantic roles is based on frame-semantic theory and utilizes the FrameNet lexicon (cf. Section 1.1). The SemEval data set further differs from other data sets in that annotation is not restricted to predicates with specific lemmas or parts-of-speech. In other words, all predicates and their arguments in discourse are annotated. In total, the released training and test data sets contain 580 and 710 annotated "null instantiations" (NI), respectively. According to the organizer's own analysis (Ruppenhofer et al., 2012), however, only 245 and 259 annotated arguments, respectively, are linked to discourse antecedents. These linked cases, which Ruppenhofer et al. also call *resolvable*, correspond to the kind of implicit arguments that we focus on in this thesis.

The part of the SemEval shared task that we describe here is called the "NI only" task. In this task, participating systems have to identify and link implicit arguments in discourse. Given manual annotations of local predicate-argument structures, Ruppenhofer et al. (2010b) describe the task as three steps: firstly, NIs have to be identified; secondly, they have to be classified as "being accessible to the speaker" (definite null instantiation, DNI) or as being "only existentially bound within discourse" (indefinite null instantiations, INI); finally, all resolvable null instantiations have to be linked to discourse antecedents. In 2010, three teams participated in the "NI only" task. Only two of the participating systems did, however, resolve any NIs. We briefly describe these systems in the next paragraph .

Participating systems. The first of these two systems was developed by Chen et al. (2010) and is an extended version of the semantic role labeling (SRL) system SEMAFOR (Das et al., 2010). For linking implicit arguments, the system chooses the highest ranked

candidate from all noun phrases that occur within the previous three sentences. Ranking is performed using standard SRL features (part-of-speech tags, passive voice, etc.) with two modifications: the number of sentences between the predicate and candidate is used instead of traditional features of word ordering and distance; and distributional similarity is used as an additional feature to determine whether a candidate is appropriate, given a locally unrealized semantic role. The second system was developed by Tonelli and Delmonte (2010) and is an extension of a semantic processing and textual entailment engine called VENSES (Delmonte, 2005). The adapted system, which they call VENSES++, combines several linguistic analysis modules, including a syntactic-semantic parser (based on LFG; Bresnan, 1982), anaphora resolution (using a "topic hierarchy"; Delmonte, 2006) and an additional step to identify and link implicit arguments. This last step is based on a restricted set of heuristics and hand-crafted knowledge sources such as WordNet (Fellbaum, 1998) and ConceptNet (Liu and Singh, 2004). The results of both systems, the extended version of SEMAFOR and VENSES++, for linking implicit arguments lie slightly above 1% in terms of F_1-score.

2.1.4. Silberer & Frank (2012)

Following the shared task in 2010, more work has been carried out to better handle implicit arguments and achieve better results on the released data set. We describe a range of recent approaches in Section 2.1.5. In this section, we describe one particular system in more detail as we make use of it in our own experiments to train and evaluate new models. This system was developed by Silberer and Frank (2012).

As the training data in the SemEval task only comprises 245 resolvable instances of implicit arguments, Silberer and Frank propose to heuristically acquire additional data by treating anaphoric pronoun mentions as being implicit. As a proxy for anaphoricity, Silberer and Frank consider pronouns that occur in a manually or automatically annotated coreference chain. For each coreference chain that contains a pronoun, one (positive) training data point is created that consists of the pronoun – treated as an implicit argument – and the closest previous mention as the correct antecedent. To create negative training data, all entity mentions that occur between the artificial implicit argument and its correct antecedent are selected as incorrect antecedents.

The system is able to tackle all three steps in the "NI only" setting of the SemEval task: (1) identifying missing arguments, (2) classifying arguments as DNIs and INIs, and (3) linking resolvable arguments to antecedents in discourse. In step (1), the system by Silberer and Frank identifies unfilled FrameNet core roles as implicit arguments, given that the roles do not compete with already filled roles; in step (2), a SVM classifier is used to predict whether implicit arguments are resolvable based on a small amount of features: the semantic type of the affected Frame Element, the relative frequency of its realization type in the SemEval training corpus, and a Boolean feature that indicates whether the affected sentence is in passive voice and does not contain a (deep) subject. In step (3), a BayesNet classifier is used to find appropriate antecedents for arguments that are predicted to be resolvable. The classifier is learned using training data from the SemEval shared task and heuristically acquired data (based on anaphoric pronouns).

In contrast to Chen et al. (2010), features for classification go beyond typical semantic role labeling approaches and also take into account coreference information. As no coreference annotations were provided during evaluation in the SemEval shared task, automatic coreference chains were computed using a coreference resolution system (Cai and Strube, 2010). The ten highest weighted features in Silberer and Frank's approach comprise the following factors:

1. the number of times a candidate entity has previously been mentioned,
2. the part of speech or phrase type of the most recent mention,
3. the number of different entity mentions (if any) between the missing argument and the candidate antecedent,
4. the distance between the missing argument and candidate antecedent in sentences,

5.-7. the VerbNet roles (Kipper et al., 2008), Frame Element names and their semantic types (in FrameNet) of the missing argument and all roles that the candidate entity fills according to its coreference chain and local role annotations,

8. the average distance between all mentions of the candidate entity and the missing argument,
9. agreement of the semantic type of the missing argument (according to FrameNet) and the supersense of the candidate entity, according to its hyperonyms in WordNet (Fellbaum, 1998), and
10. the grammatical function of the candidate antecedent.

As can be seen in the above list, most of the highest weighted factors concern coreference, including the number of and distance between mentions. In contrast, typical semantic role labeling features, such as grammatical function and type agreement, are lower ranked. Without additional training data, the model of Silberer and Frank achieves a precision and recall of 6.0% and 8.9%, respectively, outperforming the best system of the shared task by a margin of 6 percentage points in F_1-score (from around 1% to 7.1%). To create additional training data with the pronoun heuristics, Silberer and Frank make use of three different corpora that are annotated with local semantic role information. For each corpus, they perform feature selection and train a separate linking model. Feature selection is performed using 10-fold cross-validation on the SemEval training data plus additional data from the given corpus. Their best performing model achieves a precision and recall of 9.2% and 11.2%, respectively, resulting in a F_1-score of 10.1%, a further improvement of 3 percentage points.

2.1.5. Recent Developments

In this section, we give a brief overview of other recent developments on the task of linking implicit arguments. To the best of our knowledge, only two recent approaches used a different data set than the one provided by SemEval: the work by Gerber and Chai (cf. Section 2.1.2) and a recent follow-up study by Laparra and Rigau (2013a). As most previous work focused on the more diverse data set provided by the SemEval shared task, we restrict our discussion accordingly.

Laparra and Rigau (2012–2013). Laparra and Rigau (2012) propose to use statistics computed over the annotated FrameNet corpus to identify resolvable implicit arguments: given a specific frame and realized frame elements, they look up the most common combination of frame elements that contains those that are realized; based on the looked up list, they classify all elements as implicit that are not realized in the given context. The performance on this step has an immediate influence on the overall results. Laparra and Rigau only link arguments to discourse antecedents that have been classified as implicit. To link arguments in discourse, they train a simple probabilistic model based on the SemEval training data. This model consists of two features: the first is the part-of-speech tag of a candidate word, and the second is its semantic type according to a concept ontology (Álvez et al., 2008). In follow-up work, Laparra and Rigau (2013b) add more linguistically motivated features to their original approach. These include the syntactic relationship between the predicate, which is affected by a missing argument, and the candidate antecedent, a feature that indicates whether mentions occur within dialogues or monologues, and discourse-level features that take into account focus and centering relations.

Gorinski et al. (2013). A different approach is taken by Gorinski et al. (2013) who model implicit argument linking using majority voting. Overall, they use four modules (or votes) to find appropriate antecedents for a missing argument. Two of the four modules select antecedents as potential candidates that fill a frame element with the same name or semantic type, respectively, in the previous discourse context. The third module selects candidates from the preceding context that involve the same types of frame elements (role names) as the frame that involves the missing element. Finally, the fourth module selects the candidate with the highest distributional similarity to a centroid representation of all explicit instances of the specific frame element in the SemEval training data. If at least two modules "vote" for the same discourse antecedent, the missing argument is classified as implicit and linked to the elected antecedent.

Results and further directions. While the approaches by Laparra and Rigau (2012), Laparra and Rigau (2013b) and Gorinski et al. (2013) all outperform the participating systems from the SemEval shared task in terms of F_1-score (from around 1% up to 18%), the differences in precision and recall are mixed. All three approaches are able to improve recall (from around 1% up to 25%) but none of them achieves the precision of the system by Chen et al. (2010). In fact, all precision values lie between 13% and 15%, whereas the system by Chen et al. achieves 25%. One problem for all current systems seems to lie in the sparse training data. This has been pointed out as a main error source by task participants (Chen et al., 2010; Laparra and Rigau, 2013b) and in an analysis by the task organizers (Ruppenhofer et al., 2012). Follow-up work by Moor et al. (2013) proposes to alleviate this issue by using additional training data that is created by manual annotation. They show that more data can successfully be used to improve the precision of models for the sub-task of linking resolvable implicit arguments to discourse antecedents (from 25.6% up to 34.3%). Manual annotation, however, is

costly and hence does not scale well to different and more diverse domains. A partial solution to the problem of scarce training data has been proposed by Silberer and Frank (2012, cf. Section 2.1.4). They show that by treating anaphoric pronouns as instances of implicit arguments, training can successfully be extended to improve the performance of linking discourse antecedents. Pronouns, however, do not necessarily reflect the same properties as implicit arguments. For example, anaphoric pronouns typically refer to entities that are salient in discourse; in contrast, arguments can also be omitted because they are irrelevant in context. Vice versa, not every anaphoric pronoun can be omitted in practice, leading to an incorrect overgeneralization.

In this thesis, we propose an alternative approach to creating additional training data for the task of linking implicit arguments. In contrast to previous approaches, we neither rely on costly annotation nor on instances that are artificially created based on a related but different linguistic phenomenon. Instead, we propose to identify implicit arguments by comparing argument structures in pairs of comparable texts. We present our overall framework for this task in Chapter 4. An experimental evaluation of the impact of automatically induced data in context of the SemEval shared task is presented in Chapter 8.

2.2. Entity-based Coherence Modeling

In this section, we describe previous work on entity-based coherence modeling. We give an overview of such models, which relate the (local) coherence of a text to entity realizations in discourse, and discuss potential improvements. As indicated in Chapter 1, whether an entity needs to be explicitly realized in text depends, among other factors, on its discourse salience. In the previous sections, we have already seen that salience is an essential factor for linking implicit arguments: Silberer and Frank (cf. Section 2.1.4) found the number of preceding entity mentions to be the strongest feature in their model; earlier approaches to linking implicit argument (cf. Section 2.1.1) relied on variants of the so-called focusing algorithm to determine the salience of an entity, and hence to decide whether it is suitable antecedent for implicit argument linking. The idea of the focusing algorithm, as originally proposed by Sidner (1979, 1981), is to provide rules and inference tools for tracking the focus of attention in discourse. While the original motivation for Sidner's approach lies in the interpretation of pronominal anaphora, work by her and others (Joshi and Weinstein, 1981; Grosz, 1977, inter alia) have spurred further research on discourse structure and led to more general studies on the interaction between local coherence and (choice of) referring expressions. This aspect is also of particular importance in natural language generation. In the following sections, we discuss previous work in this direction in more detail. In Section 2.2.1, we introduce *Centering*, an entity-based framework for modeling local coherence in discourse. In Section 2.2.2, we describe the *entity grid approach*, a computational model that implements some of the ideas from Centering. In Section 2.2.3, we outline other entity-based coherence models proposed in the literature. Finally, we discuss the role of implicit arguments in coherence modeling and give an outlook on our approach in Section 2.2.4.

2.2.1. Centering

As described in the seminal work by Grosz et al. (1995), constituents of a discourse, its *discourse segments*, exhibit local and global coherence: on the level of *local coherence*, utterances cohere within a discourse segment; on the level of *global coherence*, a segment coheres with other segments in discourse. Grosz et al. propose *Centering* as a framework to modeling local coherence. According to this framework, utterances in a discourse segment are linked by so-called *centers*, which are "semantic objects" in the discourse. For each utterance, Grosz et al. propose to model these objects using two representations: a (partially) ordered list of *forward-looking centers* and a unique *backward-looking center*. The list of forward-looking centers reflects the semantic objects that are "realized" within an utterance.[1] The backward-looking center is a single semantic object that links the current utterance to a forward-looking center from the previous utterance. The higher a forward-looking center is "ranked" in the previous utterance, the more likely it is predicted to be the backward-looking center of the current utterance. Based on the two structures, Grosz et al. claim that local coherence is affected by how centers change from one utterance to another. More specifically, they define three types of center transitions:

> CONTINUATION – the backward-looking center of the previous and current utterance is the same, and the highest ranked forward-looking center of the current utterance is the same as its backward-looking center,
>
> RETAINING – the backward-looking center of the previous and current utterance is the same, but the highest ranked forward-looking center of the current utterance is not the same as its backward-looking center,
>
> SHIFTING – the backward-looking centers of the previous and current utterance are not the same.

According to one of the rules in the Centering framework, "sequences of continuation are preferred over sequences of retaining; and sequences of retaining are to be preferred over sequences of shifting" (Grosz et al., 1995). Following this rule, Grosz et al. characterize a locally coherent segment by the observation that SHIFTING should typically be followed by a sequence of CONTINUATION transitions.

The interpretation of several concepts in Centering remains open in the framework put forward by Grosz et al. In this thesis, we interpret "utterances" to be sentences and "semantic objects" to be the entities that are explicitly referred to in a text. Based on these two interpretations, we answer the question of what is "realized" in a sentence on the basis of semantic analysis. In particular, we take into account that entities, in the form of semantic arguments, can contribute to the meaning of a sentence without being mentioned in a specific sentence but by being understood implicitly (cf. Chapter 1). This interpretation is in line with the definitions by Grosz et al., who explicitly outline the possibility that a center "of an utterance is realized but not directly realized in that utterance."

[1] According to Grosz et al., "the precise definition of [whether utterance] U realizes [a center] c depends on the semantic theory one adopts"

An open question that remains at this point is how forward-looking centers are ranked. The Centering framework only discusses two factors explicitly: pronominalization ("lower-ranked elements ... cannot be pronominalized unless higher-ranked ones are") and the grammatical role of the expression that realizes a center: `subject` (highest rank), `object(s)` (lower rank) or `other` (lowest rank). Various other criteria have been discussed in related work: for example, Gordon et al. (1993; 1995) examine the effect of word order, grammatical roles, passivization and pronominalization; in contrast, other researchers argue that grammatical indicators should be replaced by functional role patterns (Strube and Hahn, 1996) and that a number of reference types, including deixis and event reference, are problematic for a purely syntactic approach (Cote, 1998). The consideration of additional factors is even more essential in languages in which pronoun references can be omitted in specific syntactic positions. In the context of Centering, this phenomenon of *pronoun-dropping* has been studied, for example, in Japanese (Walker et al., 1990), Turkish (Turan, 1995) and Italian (Di Eugenio, 1990).

While various different models of local coherence based on Centering have been suggested in the literature, only few of them have been implemented and evaluated empirically. In the following sections, we review such models from previous work, which can be used to predict the local coherence of a text. To ensure that models are applicable in our evaluation setting, we restrict our discussion on approaches that are not bound to a specific domain or rely on manual annotation. In particular, we do skip models in our discussion that were developed on the manually annotated GNOME corpus (Poesio, 2004).

2.2.2. Entity Grid Model

Inspired by the Centering framework, Barzilay and Lapata (2005) propose an entity-based model to automatically assess the local coherence of a natural language text. The model builds on the assumption by Grosz et al. (1995) that specific centering transitions should be preferred over others in a locally coherent text. In contrast to Grosz et al., however, Barzilay and Lapata abstract from a sole focus on centers and consider sentence-to-sentence transitions of references to all discourse entities in a text. More specifically, they represent a text by a so-called *entity grid*, which is a two-dimensional array that describes the references to each entity (represented by the columns of the grid) in each sentence (represented by the rows of the grid). Given the grid representation, patterns of local transitions can be learned from adjacent cells in each column of the grid, which represent how references to the same entity are realized in subsequent sentences. In the original proposal by Barzilay and Lapata, each cell only contains information on the grammatical role of a reference. An example is illustrated in Figure 2.1.

As observable in Figure 2.1, the transitions in this example grid reflect, to some extent, the preference of centering movements stipulated by Grosz et al.: entities realized in prominent syntactic positions in one sentence are more likely to be in prominent positions in the following sentence (e.g., *subject* $\rightarrow$ *subject*, *object* $\rightarrow$ *subject*) than entities realized in less prominent positions (e.g., *other* $\rightarrow$ –, – $\rightarrow$ *other*).

As a standalone model, the entity grid was proposed for and applied on three different

	John	Mike	a test
John has been acting quite odd.	*subject*	–	–
He called up Mike yesterday.	*subject*	*object*	–
Mike was studying for a test.	–	*subject*	*other*
He was annoyed by John's call.	*other*	*subject*	–

Figure 2.1.: Short text from Grosz et al. (1995) and its representation as an entity grid.

tasks: sentence ordering, summary coherence rating and readability assessment (Barzilay and Lapata, 2008). The information contained in the grid is limited, however, to the grammatical role of each reference. Hence, the model has commonly been applied in conjunction with additional features in other tasks, including for example, rating coherence and readability of news articles (Pitler and Nenkova, 2008) and essay responses (Burstein et al., 2010), story generation (McIntyre and Lapata, 2009, 2010), assigning texts to elementary school grade levels (Feng et al., 2010), and authorship attribution (Feng and Hirst, 2014). In the next section, we discuss two particular models that have been proposed to cover other factors related to entity references, which can be combined with features derived from Lapata and Barzilay's grid representation.

2.2.3. Other Approaches

The entity grid approach, as described in the previous section, only takes into account the grammatical role of an entity reference. As discussed in context of the Centering framework (cf. Section 2.2.1), other factors can further affect the ordering of forward-looking centers and hence influence the perceived coherence within a discourse segment. Two of these factors, namely information status and pronoun use, have been addressed in models complementary to the entity grid: for example, in the pronoun model by Charniak and Elsner (2009) and in the discourse-new model by Elsner and Charniak (2008). In this section, we briefly review both of these models and describe a recently proposed graph-based model that captures similar features as the entity grid.

Elsner and Charniak (2008). One factor not covered in the entity grid is whether a referring expression introduces an entity (*discourse-new*) or whether it refers to an entity that has been mentioned previously in discourse (*discourse-old*). Elsner and Charniak (2008) implement this factor in the form of a probabilistic model: given a (presumably coherent) text, they extract chains of coreferring entity mentions by looking for matching head words. Based on the acquired references, they learn to distinguish between the first mention and follow-up mentions using a maximum-entropy classifier (Daumé III, 2004) and syntactic features that have previously been applied to recognize discourse-new entities (Uryupina, 2003). This classifier can be used to predict whether entity references cohere by applying the same process in reverse: that is, it can be applied to classify whether an entity reference should be discourse-new or discourse-old and the

classification output can then be compared against the actual order of mentions in each coreference chain, extracted using the same heuristic as during training. Elsner and Charniak (2008) also introduce a pronoun-based model that predicts the probability of a pronoun realization given features such as the distance of the pronoun to its antecedent and the number of previous mentions. In contrast to their discourse-new model, this model is trained on manual annotations and hence requires additional training data.

Charniak and Elsner (2009). An unsupervised alternative to the pronoun-based model has been proposed in follow-up work by Charniak and Elsner (2009). In their work, they extend the idea of predicting the probability of a pronoun realization by relying on probability distributions over features that reflect the person, number, gender, and context (e.g., syntactic positions and part-of-speech information) of the pronoun and its potential antecedents. In contrast to the original model suggested in 2008, Charniak and Elsner do not rely on annotated data. Instead, they iteratively learn each probability distribution by maximizing the expected likelihood of the observed data (also called *expectation-maximization*; Baum, 1972). That is, given one pronoun in a text, their initial model (before the first iteration) views every discourse entity as an equally likely antecedent. In each learning iteration, this likelihood is then updated based on observed feature counts in the text: for example, if a text only contains the pronoun "he", the model will learn that all entities in this text are more likely to be antecedents of "he" than antecedents of "she"–because the pronoun "she" has not been observed in the text.

Guinaudeau and Strube (2013). As an alternative to the entity grid model, Guinaudeau and Strube recently proposed to represent entities in a graph. In this graph representation, entities and sentences are contained as nodes and edges between them indicate that an entity is mentioned in the respective sentence. Based on this graph, local coherence is modeled by computing a projection graph that represents how strongly sentence nodes are (indirectly) connected to each other (via entity nodes). One advantage of this representation is that it captures long-range "connections" between sentences. In contrast, transition patterns in the entity grid are typically restricted to two or three sentences as instances of longer patterns are sparse in data. A further difference to the entity grid is that the graph-based model does not contain explicit information on absent entities, meaning that this model does not make any (potentially incorrect) assumptions about entities that might be implicitly referred to in a sentence.

Results. Elsner and Charniak (2008, 2011a) and Guinaudeau and Strube (2013) perform several experiments to evaluate the performance of their models and of the entity grid model. In one specific task, called sentence ordering, the models have to distinguish between a text in its original (and presumably coherent) order and an alternative variant that contains the same sentences but in a randomly shuffled (hence presumably incoherent) order. Depending on the applied weighting scheme, the graph-based model by Guinaudeau and Strube outperforms Barzilay and Lapata's entity grid model on this task. While neither the pronoun-based model nor the discourse-new model is able to

outperform the entity grid, Elsner and Charniak (2011a) found that a combination of models performs better than the entity grid alone. This outcome demonstrates that the entity grid model is not sufficient on its own and that it can profit from complementary information covered by other models.

2.2.4. Local Coherence and Implicit Arguments

In the previous sections, we discussed several computational models for entity-based coherence modeling, each of which takes into account different factors addressed in the Centering framework and proposed extensions. The presented models are, with exception of the pronoun-based model by Elsner and Charniak (2008), trained on unannotated corpora, meaning that they only take into account entity references that are explicit in each sentence. In contrast, inferable entities, which can be understood from context and hence are implicit in text, are not explicitly considered. In the example of the entity grid, this leads to a crucial shortcoming, as illustrated in Figure 2.2: no matter how the two sentences in the example text are ordered, they do not share any references to the same entity.

	cigarettes	Le Havre	containers
27 tons of cigarettes were picked up in Le Havre.	*subject*	*other*	–
The containers had arrived yesterday.	–	–	*subject*

Figure 2.2.: Short text that involves only one explicit mention per entity.

In contrast, the Centering framework, as discussed in Section 2.2.1, explicitly accounts for the fact that entities can be centers even if they are not "directly realized" in an utterance. One way to accommodate for "indirectly" realized entities in the entity grid is by considering bridging relations between noun phrases (Hou et al., 2013). Following this argument, "the containers" in the example in Figure 2.2 could be understood as an anaphor that refers back to the "27 tons of cigarettes", making it the backward-looking center of this utterance. In this thesis, we argue that specific instances of bridging can also be interpreted as implicit arguments. For example, when viewing "container" as a nominalization of the verbal predicate CONTAIN, the "cigarettes" from the previous sentence can be understood as one of its arguments: namely the content (`co-theme`, A1). Yet, implicit arguments and bridging anaphora only partially overlap. We find a second instance of an implicit argument in the given example: namely "Le Havre", which is the (implicit) `destination` (A4) of the verbal predicate ARRIVE. As illustrated in Figure 2.3, the entity grid representation for the example text would be more dense when all implicit arguments were to be considered, reflecting the fact the two sentences actually do cohere.

Integrating implicitly understood entities into an entity-grid like model is challenging though as it requires computational methods that can reliably identify and link them. As discussed in Section 2.1, current state-of-the-art systems that process full text only achieve precision and recall figures up to 25% on this task.

	cigarettes	Le Havre	containers
27 tons of cigarettes were picked up in Le Havre.	*subject*	*other*	–
The containers had arrived yesterday.	*implicit*	*implicit*	*subject*

Figure 2.3.: Short text that involves only one explicit mention per entity.

In this thesis, we propose a new model to study the impact of implicit arguments on local coherence separately. That is, the idea of our approach is to model the relative effect of a single realization decision on perceived coherence, without modeling the coherence of a document in absolute terms. As an overall goal, we want to use the resulting model to predict whether an entity reference should be realized explicitly to establish (local) coherence – or whether the entity can already be understood from context. Based on such predictions, the model can be applied in text generation to ensure that necessary references are explicit and that redundant repetitions are avoided. To ensure that our model is applicable in different domains, we follow the unsupervised learning paradigm put forward by the models discussed in the previous sections. This means that, instead of collecting and manually annotating instances of explicit and implicit arguments in discourse as being coherent or incoherent, we want to determine such instances automatically. In Chapter 4, we present a novel framework to induce instances of implicit and explicit arguments from comparable texts by exploiting corresponding but only partially overlapping predicate-argument structures. Based on the induced instances, we propose to learn a coherence model that predicts whether a reference to an entity should be explicit or can be understood implicitly. We argue that the prediction by such a model reflects the extent to which the choice between an explicit and implicit argument does contribute to local coherence. We present and evaluate this model in Chapter 8. As this model only makes predictions on a specific phenomenon, it can be combined with models from previous work that cover complementary coherence-related factors. We discuss this possibility further in Chapter 9.

3. Cross-document methods

Pairs of texts that convey the same information have been an important resource in statistical machine translation for the past two decades. More specifically, texts that are available in multiple languages have been exploited to automatically induce translations of words and phrases. Methods for this task traditionally require text pairs as inputs that consist of sentences that are translations of one another. In this thesis, we refer to pairs of texts that correspond to each other on the sentence level as *parallel texts*. In contrast, we use the term *comparable texts* to refer to pairs of texts that convey the same information in essence but only correspond to each other on the document level. This is the case, for example, if two texts contain a different amount of details or if they present information from different perspectives. We introduce some example corpora that contain parallel and comparable texts in Section 3.1.

In statistical machine translation, parallel texts are typically used as input for statistical learning methods. These methods try to establish links, or *alignments*, between words and phrases in sentences of one language and words and phrases in corresponding sentences of another language. A simple method to establish such alignments is to consider the relative frequency with which words and phrases in the two languages occur in corresponding sentences. Similar methods can be applied in monolingual settings, where both texts are in the same language. We discuss some of these methods in more detail in Section 3.2. Alignments in multilingual parallel texts provide an ideal basis for the induction of dictionaries. That is, for a word in one language, aligned words in the other language can be viewed as specific translations. Similarly, aligned words in a monolingual setting can be viewed as synonymous in their respective contexts. By abstracting from the level of single words to semantic and syntactic structures, alignments can also serve as a basis for tasks such as annotation projection and paraphrase detection. We present an overview of such tasks that are related to our work in Section 3.3.

3.1. Parallel and Comparable Texts

In a cross-lingual setting, parallel texts are commonly used for inducing alignments and translations from one language to another. There exist a wide range of texts that are available in multiple languages, including translations of books and multilingual proceedings of parliament discussions. Two popular and freely available examples are the Bible, parts of which have been translated into more than 2000 languages (Resnik et al., 1999), and the Europarl corpus (Koehn, 2005), which contains the proceedings of the European Parliament in eleven languages. Examples of monolingual parallel texts include different editions of the same book and multiple translations of a text into one language (Barzilay and McKeown, 2001; Huang et al., 2002; Cohn et al., 2008). As the availability of such

resources is limited, there have been various attempts to extract parallel fragments from comparable texts (Barzilay and Elhadad, 2003; Regneri and Wang, 2012). For example, Barzilay and Elhadad (2003) propose to cluster corresponding paragraphs in comparable texts and extract sentence pairs based on word overlap. Monolingual comparable texts include, for example, news reports on the same event, each provided by a different news source. For the past two decades, such data sets have often been manually created as basis for tasks such as multi-document summarization (McKeown and Radev, 1995). Nowadays, there exist large amounts of data from which comparable texts can be extracted automatically: for example, they can be mined online from news web sites, or can be extracted from corpora of newswire articles (for example, the English Gigaword Corpus; Parker et al., 2011).

The focus of this thesis is on texts that describe the same events in a monolingual setting. We do not require texts to be parallel as our goal is to specifically identify corresponding predicate-argument structures that differ with respect to specific argument realizations. We argue that differences in content and presentation could actually be beneficial as they potentially represent factors that license the omission of an argument in one case that needs to be explicitly realized in another. As a basis for our approach, we construct our own corpus of comparable texts from newswire articles. In doing so, we address several shortcomings that we observed in corpora from previous work: manually compiled data sets, on the one hand, are too small to provide sufficient contexts for our work; on the other hand, automatically composed data sets are too noisy as they are either based on corpora derived from the web (Dolan et al., 2004; Wubben et al., 2009) or identified using methods that are not tuned for precision (Wang and Callison-Burch, 2011). In contrast, we refine an existing method of identifying comparable texts to achieve high precision and we apply it to one of the largest data sets of English newswire that is currently available. We describe this data set, our extraction procedure and results in more detail in Chapter 5.

3.2. Alignment

One of the first approaches to statistical word alignment was put forward by Brown et al. (1993). In their work, Brown et al. propose to build on the idea that frequently co-occurring words across languages are likely to be translations of one another. More recent developments on word alignment, for example, the alignment toolkit GIZA++ (Och and Ney, 2003) and the state-of-the-art Berkeley Aligner (Liang et al., 2006), extend this idea and provide readily available implementations. As we make use of these two toolkits in the evaluation of our own alignment model, we describe both of them in more detail below.

GIZA++. The alignment toolkit GIZA++ provides implementations of various alignment models, including the "IBM models" by Brown et al. (1993) as well as several refinements and additional smoothing techniques. Given one sentence in two languages, the basic model by Brown et al., also called "Model 1", computes alignment probabili-

ties between any word in the sentence of one language and words in the corresponding sentence in another language. This probability can be estimated in two steps: in the first step, co-occurrences of word pairs are counted in all corresponding sentences; in the second step, probabilities are computed by normalizing the observed frequency counts. Brown et al. suggest several enhancements of their Model 1 to account for the possibility that translations can depend on contextual factors. In their "Model 2", for example, each alignment probability is calculated depending on the position of a word in the sentence. That is, the model can learn that words might be more likely to correspond when they occur in approximately the same position in a sentence. Other enhancements by Brown et al. ("Model 3" through "Model 5") cover the fact that one word in one language can correspond to multiple words in another language and that alignment probabilities can depend on previous words in context. As an additional refinement to the IBM models, GIZA++ provides several methods to combine alignments in both directions: that is, instead of just aligning words, for example, from an English sentence to French, the idea is to also align words from the French sentence to English. Based on these two sets of alignments, the intersection or union can be taken to obtain a more or less refined set of alignments. The GIZA++ alignments used in our experiments (cf. Chapter 6) are computed based on the intersection of two-way alignments in order to achieve high precision.

Berkeley Aligner. As an alternative to selecting the intersection of two-way alignments, Liang et al. (2006) propose to simply train one joint distribution that models alignments in both directions simultaneously. In their evaluation of this approach, they show that by optimizing the joint alignment probabilities, the rate of incorrect alignments can be reduced from 6.9% (using GIZA++) to 4.9%. Liang et al. freely distribute their implementation of this simple, yet very effective idea as an alignment toolkit called Berkeley Aligner. We make use of it in some of our experiments in Chapter 6.

One problem with alignments on the level of words is that phrases in two languages cannot always be translated literally. Hence, other approaches in related work suggest to extend alignment models from the word level to phrases and syntactic trees (Gildea, 2003, 2004; DeNero et al., 2008). Alignment tools developed for machine translation can also be applied to align words and phrases in monolingual parallel texts (Quirk et al., 2004; Cohn et al., 2008). In the case of comparable texts, however, the alignment task is more challenging. As documents can convey different pieces of information and describe events from different perspectives, some words and phrases in one text might not correspond to any word or phrase in the other text. Previous approaches to this problem can be categorized into two classes: one option is to use a pipeline, in which corresponding fragments and sentences are identified before applying alignment techniques on the word or phrase level (Barzilay and Elhadad, 2003; Wang and Callison-Burch, 2011); another option is to rely on more sophisticated methods that align words and phrases based on specific cues. For example, Shinyama et al. (2002) align phrases that contain references to the same named entity; and Shen et al. (2006) base alignment decisions on local

syntactic similarities.

In this thesis, we propose a novel clustering technique to perform alignments in pairs of comparable text. Instead of relying on specific semantic and syntactic cues, our method operates on the level of predicate-argument structures and makes use of various similarity measures. Each measure contributes a specific type of information: for example, lexical semantic relations between two predicates, similarity of their arguments, and commonalities in discourse context. By identifying predicate-argument structures (PAS) across texts that describe the same event, state or entity, we can examine which arguments are realized in each context to establish a coherent discourse. Different related aspects have been studied in previous work. For example, Filippova and Strube (2007) and Cahill and Riester (2009) examine factors that determine constituent order. Belz et al. (2009) study conditions for the use of different types of referring expressions (cf. Section 1.2). Identifying corresponding PAS in pairs of comparable texts allows us to further investigate the factors that govern the omission of an argument in a specific context, as a special form of coherence inducing element in discourse. We describe our clustering approach and similarity measures used to find corresponding predicate-argument structures in Chapter 6. For comparison with previous approaches to word alignment, we present evaluations on data sets of parallel and comparable texts.

3.3. Applications

As discussed in the beginning of this chapter, alignments in pairs of texts can be used as a basis for creating resources such as translation dictionaries and paraphrase databases. The simplest way to do so is to consider the relative frequency with which two words or phrases are aligned. By abstracting from specific textual realizations to higher levels of linguistic analysis, word alignments can furthermore be exploited to "project" annotations from one language to another language. Applications in monolingual settings, which are more closely related to ours, include textual inference and paraphrase extraction. We describe the different settings in more detail below.

Projection approach. The idea behind this approach is to induce annotated data in one language, given already annotated instances in another language. As an example, frame-semantic annotations of a text in English (cf. Section 1.1) can be transferred in a parallel text in order to induce annotated instances for a frame-semantic lexicon in another language (Padó and Lapata, 2009). In previous work, this *projection approach* has been applied on different levels of linguistic analysis: from syntactic information in the form of part-of-speech tags (Yarowsky and Ngai, 2001) and dependencies (Hwa et al., 2005), over annotations of temporal expressions (Spreyer and Frank, 2008) and semantic roles (Johansson and Nugues, 2006; van der Plas et al., 2011), to discourse-level phenomena such as coreference (Postolache et al., 2006) and relations between sentences (Versley, 2010). All of the aforementioned instances of the projection approach make use of the same kind of technique: firstly, words are aligned in a parallel corpus using statistical word alignment; secondly, annotations on a single word or between multiple

words in one text are transferred to the corresponding aligned word(s) in the parallel text. This approach is related to our work in that we are also interested in inducing annotation that is available in one text (explicit argument) but not in another (implicit argument). In contrast to previous applications of this approach, we are interested in the case of comparable texts, meaning that word alignments can be difficult to establish.

Textual inference. Another application of word alignments can be found in the area of textual inference. Since 2006, regular challenges have been organized on the task of Recognizing Textual Entailment (RTE). According to the task description by Dagan et al. (2006), *textual entailment* is defined as a relationship between pairs of text fragments, in which the meaning of one text, the entailed hypothesis H, can be inferred by interpreting the meaning of the other text, the entailing text T. Although the entailment relation does not necessarily require the presence of corresponding words, previous work by MacCartney et al. (2008) shows that word alignments are good indicators of entailment. In our work, we are interested in corresponding predicate-argument structures that differ with respect to specific argument realizations. If all realized arguments and the two predicates in a pair of PAS do correspond, we can also view predicate-argument structures as special cases of text and hypothesis pairs.

Paraphrase detection. The task of paraphrase detection is closely related to textual entailment. In fact, paraphrase detection can be defined as recognizing bi-directional, or symmetric, entailments. That is, each of two text snippets must entail the other. Wan et al. (2006) show that a simple approach solely based on word and lemmatized n-gram overlap can already achieve an F_1-score of up to 83% for detecting paraphrases in the Microsoft Research Paraphrase Corpus (MSRPC; Dolan and Brockett, 2005). This result lies just 0.6 percentage points below the state-of-the-art results reported by Socher et al. (2011).

Textual similarity. The goal of the Semantic Textual Similarity task (STS; Agirre et al., 2012) is to automatically rate the similarity of two sentences. The best performing system that participated in the STS task accomplished this by applying a combination of different similarity measures including features such as n-gram overlap and pairwise word similarities (Bär et al., 2012). In contrast to paraphrasing and entailment, sentence pairs in STS are not required to be in a specific relation to one another. We illustrate this case in the two Examples (11) and (12):

(11) John sold a car.

(12) Mike paid $3.000 for the car.

The sentences in Example (11) and (12) both describe aspects of an event, in which one person sold a car to another person. Yet, no sentence-level entailment or paraphrase relation can be observed as the buyer is unknown in (11) and the seller is unknown in (12). This assessment could change, however, if there was some additional context, from which we were to infer that the buyer and seller is the same in both sentences.

Discussion. Unfortunately, data sets in STS and other tasks, including the MSRPC and those of the first RTE challenges, only consist of isolated pairs of sentences. An exception to this is the "Search Task" that has been introduced in context of the Fifth PASCAL Recognizing Textual Entailment Challenge (Bentivogli et al., 2009). In this task, entailing sentences for a hypothesis have to be found in a set of full documents. This new task first opened the doors for assessing the role of discourse (Mirkin et al., 2010a,b) in RTE. This setting is still limited, however, as discourse contexts are only provided for the entailing part T of each text pair but not for the hypothesis H. In contrast, the corresponding predicate-argument structures, from which we want to induce implicit arguments, are both embedded in full discourse contexts.

In this thesis, we employ a technique that is similar to the projection approach but differs from it in two notable ways. Firstly, the data sets that we use comprise pairs of comparable texts. This means that sentences are not parallel and hence we cannot align each word or phrase in one text with a corresponding word or phrase in another text. To solve this problem, we build on pairs of aligned predicate-argument structures (cf. Section 3.2). For the alignment of PAS pairs across texts, we employ various similarity measures, some of which overlap with those implemented in systems that are also applied on paraphrase detection and rating textual similarity. The second notable difference is that the information we want to "project", namely explicit arguments that are otherwise implicit, is only present in one predicate-argument structure but not in the other. Hence we have to detect other mentions of the same entity in the text in which we found an argument to be missing. We outline our framework to deal with these challenges in Chapter 4. A computational implementation of our method to solve the actual task of identifying and linking implicit arguments in discourse is described in Chapter 7.

Part II.

Automatically inducing implicit arguments

4. A Framework for Implicit Argument Induction

***Q1**: "How can implicit arguments be induced from comparable texts?"*

As discussed in Chapter 2, implicit arguments are a frequent, yet difficult and understudied, phenomenon in semantic role labeling. Whether an argument is realized also plays a crucial role in establishing local coherence. Training data for computational models is, however, limited because there only exist few and small data sets that contain manual annotations of implicit arguments. In this chapter, we outline a novel approach to induce instances of implicit arguments together with their respective discourse antecedents. This approach exploits complementary information realized in monolingual comparable texts. The overall procedure involves three steps: (1) monolingual comparable texts are extracted from a large text corpus; (2) references to identical events, states and objects – in the form of predicate-argument structures (PAS) – are identified across comparable texts; and (3) pairs of PAS are compared to identify and link implicit arguments in discourse. We refer to the combination of these three steps as our "framework for implicit argument induction". An illustration of this framework is shown in Figure 4.1.

We previously outlined the induction method in the *SEM 2013 paper "Automatically identifying implicit arguments to improve argument linking and coherence modeling" (Roth and Frank, 2013). This chapter is based on parts of this paper, which have been revised and extended by additional details. The chapter is divided into three parts: in Section 4.1, we first describe a method to obtain a data set of comparable texts (Step 1); in Section 4.2, we address the task of aligning predicate-argument structures in comparable texts (Step 2); finally, we outline a heuristic approach to identify and link implicit arguments (Step 3) in Section 4.3. Note that this chapter only provides an overview of each of the three steps. More detail can be found in Chapter 5, 6 and 7, respectively.

4.1. Creating a Corpus of Comparable Texts

The goal of Step 1 of our implicit argument induction technique is to compile a data set of comparable texts. In Chapter 3, we discussed a range of applications that make use of parallel and comparable texts. Corpora containing parallel texts are, however, limited in monolingual settings. Furthermore, we are specifically interested in the impact of different discourse contexts on the realization of references to the same entity. For this task, we hence consider pairs of comparable texts, which convey information about the same events, states and entities. Some of the largest corpora for English, which at the

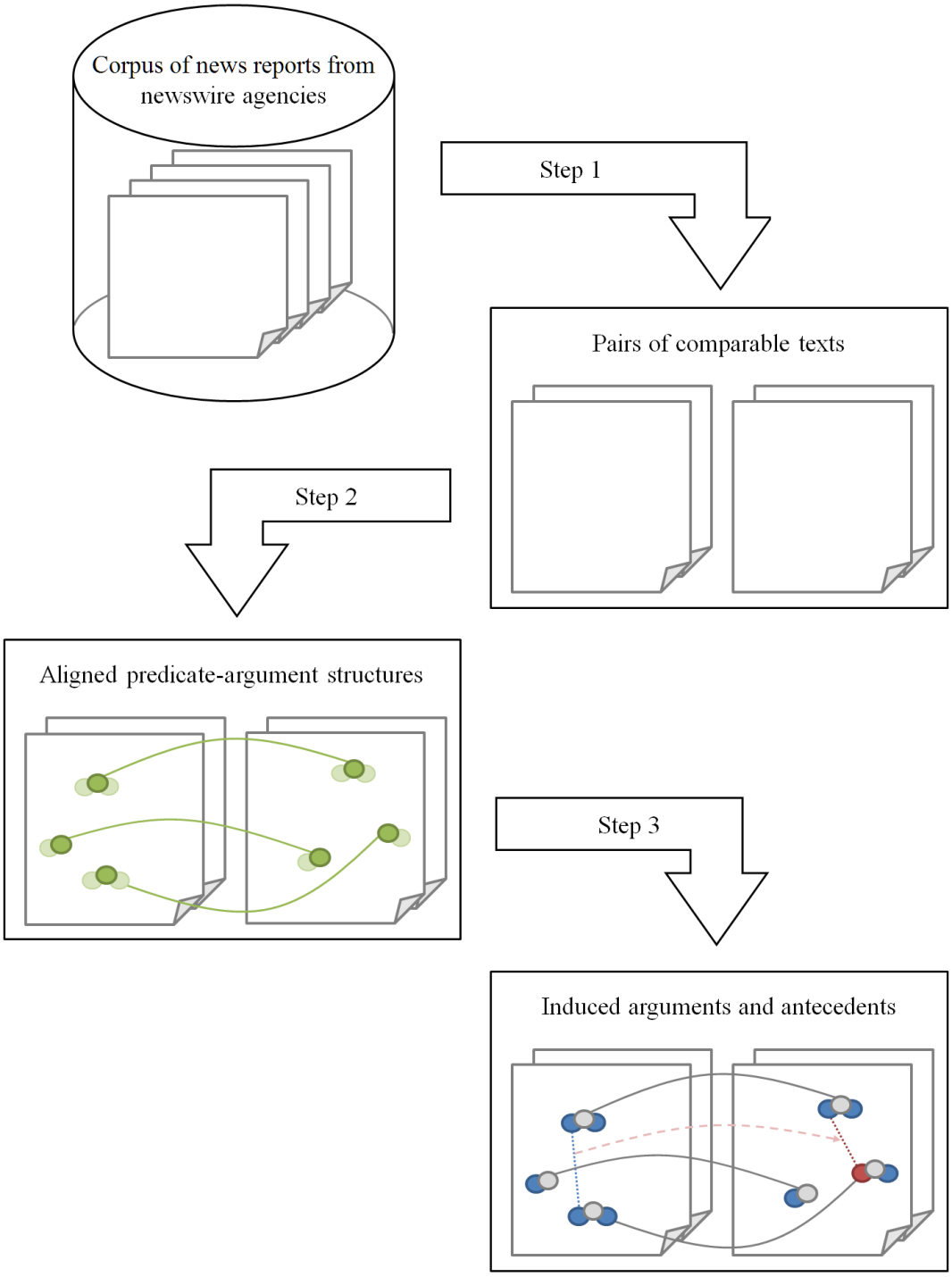

Figure 4.1.: Illustration of the three steps of our "framework for implicit argument induction": in Step 1, comparable texts are identified and extracted from a large corpus; in Step 2, predicate-argument structures are aligned across pairs of comparable texts; in Step 3, implicit arguments are identified and linked by exploiting complementary information from both documents.

same time fulfill this requirement, are available in the news domain. In this domain, various sources (typically newswire agencies) provide their own reports on the same real-world incidents. We treat such reports as comparable texts and identify them by comparing headlines and publication dates. A list of examples is displayed in (13):

(13) India fires tested anti-ship cruise missile
(Xinhua News Agency, 29 October 2003)

India tests supersonic cruise anti-ship missile
(Agence France Presse, 29 October 2003)

URGENT: India tests anti-ship cruise missile
(Associated Press Worldstream, 29 October 2003)

For our work, we make use of the English Gigaword Fifth Edition (Parker et al., 2011), which is one of the largest newswire corpora for English. We describe this corpus, our method to extract comparable texts from it, and resulting subcorpora in Chapter 5.

4.2. Aligning Predicate-Argument Structures

Based on the comparable texts extracted in Step 1, the goal of Step 2 of our induction method is to align predicate-argument structures (PAS) across documents. This step first requires PAS to be identified in each text. We discussed two prominent paradigms for semantic role labeling in Chapter 1. As we require accurate parses as input for alignment (Step 2) and implicit argument identification and linking (Step 3), we rely on a robust and highly precise preprocessing engine based on PropBank/NomBank (Palmer et al., 2005; Meyers et al., 2008). More specifically, we make use of MATE tools (Bohnet, 2010; Björkelund et al., 2010), a freely available pipeline of natural language processing modules that includes a state-of-the-art PropBank/NomBank semantic role labeler. In Figure 4.2, we illustrate the semantic structures obtained from applying this system on two sentences from a pair of comparable texts.

"It's a private visit by Bilal", the spokesman said.			Bilal is on a visit to India.		
visit	--A0--	Bilal	visit	--A0--	Bilal
	--MNR--	private		--A1--	India
say	--A0--	the spokesman			
	--A1--	``It's a private visit ...''			

Figure 4.2.: Two sentences and their respective predicate-argument structures. Each structure consists of a predicate (displayed on the left), associated arguments (on the right), and argument labels (in the center).

While both sentences in Figure 4.2 refer to a visit by a person named Bilal (A0), the first sentence contains the information that the visit is of private nature (MNR) and the

second sentence contains the location of the visit (A1), namely India. Each sentence is from one of two comparable texts and we hence assume that they refer to the same visit. Not all occurrences of the predicate VISIT in both texts refer, however, to the same event. For example, one text also contains the following sentence, referring to a different visit event: "Reports said Sonia Gandhi has also been invited to visit Pakistan".

As can be seen in the given example, the identification of corresponding predicate instances requires consideration of information that go beyond the realization of the predicates themselves. At the least, we also have to take into account information from the argument structure of specific realizations. In the remainder of this thesis, we hence refer to this task as "aligning predicate-argument structures across monolingual comparable texts". In Chapter 6, we develop a graph-based model for this task that takes into account information specific to predicates, associated arguments and their discourse contexts. The two latter groups of factors are not considered in simple word alignment methods as introduced in Section 3.2 as text fragments can be assumed to be about the same events, states and entities in parallel corpora. To evaluate different alignment methods and the impact of various features, we create a development and test set that contains manually annotated pairs of predicate-argument structures in a small corpus of comparable texts. The annotation process and evaluation of models is described in Chapter 6. For comparison with previous work, our evaluation includes various baselines as well as experiments on parallel and comparable texts.

4.3. Identifying and Linking Implicit Arguments

The final step of our implicit argument induction technique, Step 3, is to identify and link implicit arguments given pairs of aligned predicate-argument structures. In this section, we outline a heuristic method to perform this task (Section 4.3.1). We then discuss the potentials of applying this approach, based on manually aligned predicate-argument structures (Section 4.3.2). Further details regarding a computational implementation of the induction technique and its application on automatically aligned predicate-argument structures can be found in Chapter 7: Inducing Implicit Arguments.

4.3.1. Heuristic Approach

Given pairs of aligned predicates from comparable texts, we view the task of identifying and linking implicit arguments as two subsequent sub-tasks: firstly, implicit arguments have to be identified; and secondly, the identified implicit arguments have to be linked to suitable antecedents in discourse.

We tackle the first sub-task by examining the argument structures of two aligned predicates and determine which arguments are realized in each structure. We then compare the set of role labels assigned in both structures to determine whether one PAS contains an argument (explicit) that has not been realized in the other PAS. We treat all unrealized arguments in one PAS that are explicit in the aligned PAS as implicit arguments. Given an implicit argument and its explicitly realized counterpart, we tackle the second sub-task by determining references that denote the same entity as the explicit

argument. To ensure that implicit arguments are linked to antecedents *within discourse*, we restrict this procedure to realizations in the same text as the identified implicit argument.

4.3.2. Implicit Arguments from Manual Alignments

In this section, we discuss some cases of manually aligned predicate-argument structures. We are particularly interested in cases of implicit arguments and thus take a closer look at alignments involving arguments that are only realized in one of the aligned structures. The excerpts in Example (14) and (15) are from two comparable texts that describe a news report on two or more avalanches:

(14) a. Seven people, including a 12-year-old child, were killed in two [avalanches$_{\mathbf{i}}$] in southeastern Tajikistan at the weekend, an interior ministry official said Monday. (...)

b. The official said that [no bodies]$_{\text{A1}}$ had been recovered [from the avalanches$_{\mathbf{i}}$]$_{\text{A2}}$ (...).

(15) a. Six people were killed in [avalanches$_{\mathbf{i}}$] over the weekend in eastern Tajikistan, in the mountainous region bordering on Afghanistan and China, the Emergency Situations Ministry said Tuesday. (...)

b. [None of the victims' bodies]$_{\text{A1}}$ have been found [$\varnothing_{\mathbf{i}}$]$_{\text{A2}}$.

In both cases, (14b) and (15b), the theme (A1) of the predicate RECOVER and FIND is locally realized, the source (A2), however, has not been realized in (15b). Note that sentence (15b) is still part of a coherent discourse as a realization of the omitted argument (here: "avalanches") can be found in the preceding sentence (15a). An effective means to identify this discourse antecedent is to look for realizations that match the argument in the aligned PAS (here: marked by index **i**). Computationally, this step can be performed using coreference resolution techniques.

Examples (16) and (17) present another text pair, reporting on a Japanese travel alert, in which argument realizations differ for the same warning event.

(16) Japan's Foreign Ministry issued a travel alert on Monday for [Japanese nationals$_{\mathbf{i}}$] living or travelling in Europe, warning [$\varnothing_{\mathbf{i}}$]$_{\text{A2}}$ [of possible terrorist attacks by Al-Quaeda and affiliated groups]$_{\text{A1}}$. (...)

(17) Japan has issued a travel alert for [Japanese citizens$_{\mathbf{i}}$] living or traveling Europe, warning [them$_{\mathbf{i}}$]$_{\text{A2}}$ [of possible terrorist attack by al-Qaida or other groups]$_{\text{A1}}$. (...)

Here, the participle construction makes the recipient (A2) of WARN a (locally) optional argument. In the example pair (14) and (15), in contrast, omissability of the source argument (A2) is related to the lexical choice of the predicate: while RECOVER is defined as a predicate with three arguments (A0: `agent`, A1: `theme`, A2: `source`) in PropBank,

FIND only takes two arguments (i.e., A0 and A1). Syntactic-semantic information may give us insights as to why an argument is optional. There might be other factors, however, that led to the decision of realizing the given argument in one context but not in the other. In particular, discourse-level context could play a role here. While the two examples of implicit arguments above give rise to some speculative assumptions – for example, salience, sentence length and redundancy could be some factors –, more data points will be needed to estimate these factors and their influence reliably. With the goal of automatically extracting more data points, we discuss an implementation of the implicit argument induction technique on automatically aligned predicate-argument structures in Chapter 7: Inducing Implicit Arguments. We describe several approaches to make use of the induced implicit arguments in applications in Chapter 8: Applications.

4.4. Summary

This chapter introduced a framework to heuristically induce implicit arguments and their discourse antecedents from predicate-argument structures (PAS) aligned across monolingual comparable texts. We showcased the potential of this approach based on manual alignments. The examples shown in the previous section provide empirical evidence for three main ideas of this thesis:

- Implicit arguments can be identified by aligning predicate-argument structures in pairs of texts that describe the same events, states and entities.
- Aligned PAS from comparable texts, including implicit arguments, provide discourse contexts of implicit and explicit arguments alike.
- Discourse antecedents of implicit arguments can effectively be determined by looking for realizations that "match" the aligned explicit arguments.

As a basis for our heuristic induction technique, we first require monolingual comparable texts to be extracted from a large news corpus. We describe this extraction procedure and the resulting data set in Chapter 5. Secondly, predicate-argument structures need to be automatically aligned across comparable texts. We discuss a suitable model for this step in Chapter 6. Finally, we exploit complementary information realized across aligned PAS to identify implicit arguments and discourse antecedents. We present a computational approach for this step in Chapter 7. To demonstrate the utility of automatically induced instances of implicit arguments, we present extrinsic evaluations in implicit argument linking and coherence modeling in Chapter 8.

5. Creating a Corpus of Monolingual Comparable Texts

q_i: "How can we automatically identify pairs of comparable texts?"

As discussed in Chapter 4, our approach to automatically inducing implicit arguments relies on pairs of predicate-argument structures (PAS) that are aligned in monolingual comparable texts. In this chapter, we present a suitable data set of comparable texts, part of which we annotate with manual PAS alignments (cf. Chapter 6). While the annotated part serves primarily as a development and evaluation data set for automatic alignment approaches, we will use the full corpus to extract a large data set of corresponding PAS pairs in their respective discourse contexts. This data set forms the basis of our automatic approach to inducing instances of implicit arguments and discourse antecedents (cf. Chapter 7).

We previously described our corpus of monolingual comparable texts in the *SEM 2012 paper "Aligning Predicate Argument Structures in Monolingual Comparable Texts" (Roth and Frank, 2012a). This chapter is based on parts of this paper and divided into three sections: in Section 5.1, we describe Gigaword, a large corpus of news articles from different newswire sources; in Section 5.2 and Section 5.3, we present a method that extracts pairs of comparable texts and the resulting data set, respectively; finally, we summarize our results and briefly describe benefits of the created corpus for other research directions in Section 5.4.

5.1. Gigaword Corpus

The goal of Step 1 of our implicit argument induction technique is to extract a data set of comparable texts. To compile a sufficient amount of data, we make use of the Gigaword corpus, which to the best of our knowledge is the largest corpus of English newswire articles currently available. The current version of Gigaword, the "Fifth Edition", has been released in 2011 (Parker et al., 2011) and contains over 9.8 million newswire articles. The Gigaword corpus is particularly well suited for extracting pairs of comparable texts as it contains articles from seven distinct newswire sources, all of which report on real-world incidents. Each source is one of the following English newswire services by an international agency:

- Agence France-Presse, English Service (AFP)
- Associated Press Worldstream, English Service (APW)
- Central News Agency of Taiwan, English Service (CNA)

- Los Angeles Times/Washington Post Newswire Service (LTW)
- Washington Post/Bloomberg Newswire Service (WPB)
- New York Times Newswire Service (NYT)
- Xinhua News Agency, English Service (XIN)

All of these services release their own news reports on real-world events. The number of articles from each source span from 26,143 (WPB) to 3,107,777 (APW). To construct a data set of pairs of comparable texts, we make use of all combinations of agency pairs in Gigaword. All examples presented in the remainder of this chapter are taken from the agency pair AFP–APW.

5.2. Extraction Method

To identify pairs of articles describing the same news event, we compute pairwise similarities based on article headlines, using a method proposed by Wubben et al. (2009). To compute this similarity measure, the headlines of two documents are represented as sparse vectors $\vec{doc}_1, \vec{doc}_2$, in which each dimension corresponds to one word type. The actual similarity is then computed as the cosine of the angle between the two vectors, following Equation (5.1).

$$\cos(doc_1, doc_2) = \frac{\vec{doc}_1 \cdot \vec{doc}_2}{\|\vec{doc}_1\| * \|\vec{doc}_2\|} \tag{5.1}$$

When treating all word types (or dimensions) equally, the result of this measure is the same as a normalized count of the number of overlapping words in both headlines. To restrain the influence of words that commonly appear in news documents (e.g., function words) – and to strengthen the impact of words specific to an examined headline (e.g., proper names) – the effect of each occurring word type is computed as its TF-IDF score. As defined in Equation (5.2), this score is calculated for each vector space dimension as the *term frequency* of a word type w within a headline hl (TF) multiplied by the logarithm of its *inverse document frequency* (IDF). We compute the IDF value as the inverse ratio of headlines hl' that contain the word w in the subcorpus $paircorpus$ that contains all articles from the affected agency pair.

$$\text{TF-IDF}_{doc_i}(w) = |\{w \in hl_i\}| * \log \frac{|\{hl' \in paircorpus\}|}{|\{hl' \in paircorpus | w \in hl'\}|} \tag{5.2}$$

Figure 5.1 illustrates the similarity computation, including conversion of headlines to vector representations and application of TF-IDF scoring. As our corpus spans over almost two decades of news articles, we impose an additional date constraint to identify comparable texts more accurately – for example, we want to avoid pairing of news articles on 2003 elections and 2010 elections in Iraq. We apply this constraint by requiring a pair of articles to be published within the same two-day time frame in order to be considered as pairs of comparable news items.

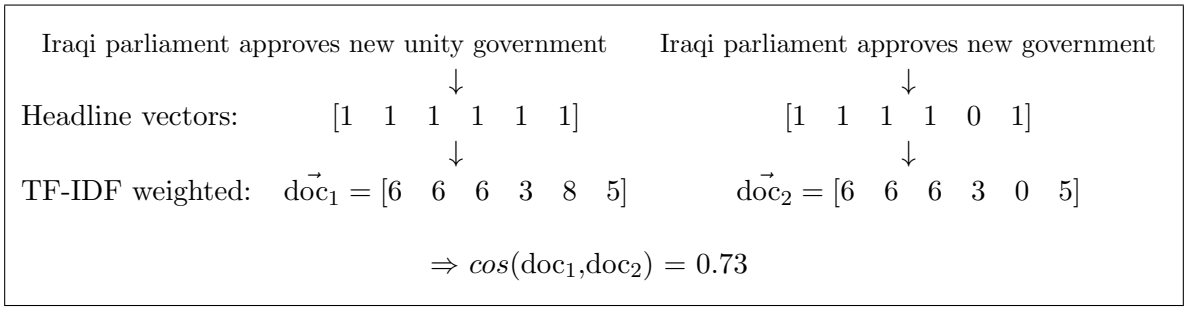

Figure 5.1.: Conversion of headlines to vector representations and similarity computation

5.3. Resulting Data Set

We applied the outlined procedure of identifying comparable text pairs to all documents from each pair of newswire sources. As a result, we extracted a total of 167,728 document pairs, an overall collection of 50 million word tokens. The distribution over pairs of newswire agencies is shown in Table 5.1. For the manual alignment of predicate-argument structures, we randomly selected 70 document pairs from the AFP–APW portion of the corpus. Our two annotators indicated that 69 of the 70 document pairs describe the same events, corresponding to a precision of 98.6%. This is in line with the results of Wubben et al. who reported a precision of 93% without explicitly imposing a date constraint. Overall, we found that most text pairs share a high degree of similarity and vary only in length (up to 7,564 words with a mean and median of 301 and 213 words, respectively) and detail. We examined a subset of 10 document pairs to identify discourse contexts, in which arguments have been non-locally realized, and found instances of this phenomenon in all pairs. The 10 document pairs are part of a manually annotated development and evaluation set for predicate alignment that we describe in Chapter 6.

Agency pair			Texts				
AFP	–	APW	52,300	XIN	–	APW	37,656
AFP	–	LTW	2,787	XIN	–	AFP	42,992
AFP	–	NYT	5,420	XIN	–	LTW	1,733
AFP	–	WPB	289	XIN	–	WPB	151
APW	–	LTW	4,054	NYT	–	XIN	3,649
APW	–	NYT	11,488	NYT	–	LTW	4,678
APW	–	WPB	335	NYT	–	WPB	196
				total			167,728

Table 5.1.: Distribution of comparable texts over pairs of newswire agencies.

5.4. Summary

In this chapter, we presented a new data set of monolingual comparable texts, extracted from the English Gigaword corpus. By combining an extraction method from the literature with an additional date constraint, we are able to retrieve pairs of comparable texts from Gigaword with high precision. In total, more than 160,000 document pairs were extracted using this approach.

In the next step, we make use of this data set to identify corresponding predicate-argument structures (PAS) across pairs of text. In Chapter 6, we describe an evaluation data set, containing manual PAS alignments, and a novel clustering approach to automatically align pairs of PAS. The resulting set of aligned PAS will be the basis for Step 3 of our induction approach, in which we aim to automatically induce instances of implicit arguments and their discourse antecedents (cf. Chapter 7). To ensure accurate results in the following steps, we will address them with high precision methods. On the downside, this means that we cannot anticipate high recall. The more than 160,000 document pairs in our new corpus will hence be a necessary prerequisite to induce a sufficient amount of argument instances to make use of in actual applications (cf. Chapter 8).

Our new corpus of comparable texts could also be useful for a range of other tasks. For example, it would be a suitable resource for extracting paraphrases (Wang and Callison-Burch, 2011) or sentences that are semantically similar (Agirre et al., 2012); more recently, "non-contradictory texts" have also been used for bootstrapping semantic analyzers (Titov and Kozhevnikov, 2010). We discuss some of these applications in more detail in Chapter 9.

6. Alignment Model

***q**ii: "How can we detect information that is shared across two texts?"*

Aligning words in texts is a well-studied task in natural language processing, with most approaches being word-based and assuming parallel data to be available. Our data set, in contrast, consists of comparable texts. This means that texts roughly contain the same information but there are variations in presentation: for example, information can be presented in a different order, it can be more or less detailed, and it can be presented from different points of view. In this chapter, we address the process of finding corresponding information in such pairs of texts as a new task, which we refer to as "aligning predicate-argument structures across monolingual comparable texts". We discuss annotation guidelines and an alignment model specifically designed for this task. To account for the fact that two comparable texts can realize complementary information, the alignment model makes use of a flexible clustering algorithm that does not align all predicate-argument structures. The clustering algorithm itself is based on pairwise similarities (or weights) between predicate-argument structures. We calculate each weight using a combination of various similarity measures that cover predicate-specific, argument-specific and discourse-specific information. We empirically validate the merits of our model in sentence-level and discourse-level evaluations.

We described the task of aligning predicate-argument structures in comparable texts in the *SEM 2012 paper "Aligning predicate argument structures in monolingual comparable texts: a new corpus for a new task" (Roth and Frank, 2012a). Parts of the alignment model have also been described in more detail in our EMNLP 2012 paper "Aligning predicates across monolingual comparable texts using graph-based clustering" (Roth and Frank, 2012b). This chapter represents an extension to both papers. In Section 6.1, we present the task of aligning predicate-argument structures, discuss annotation guidelines and introduce an annotated development and test data set. In Section 6.2, we describe a range of similarity measures that we use to identify pairs of structures that should be aligned. We make use of these similarity measures in a graph-based model, which we present together with a short introduction to graph clustering in Section 6.3. In Section 6.4, we present experiments and results that confirm that our new approach outperforms other models on the task of aligning predicate-argument structures across comparable texts. We further describe a tuning step that can be used in our graph-based approach to extract high precision alignments. These precise alignments between predicate-argument structures provide the foundation for the automatic induction of implicit arguments, which we discuss in Chapter 7. We summarize our results and outline other applications of predicate-argument structure alignments in Section 6.5.

6.1. Aligning Predicate-Argument Structures

As outlined in Chapter 4, the second step of our induction approach is to align predicate-argument structures (PAS) across comparable pairs of text. To perform this step reliably, we construct a development and evaluation set, on which we can test automatic alignment models. Both data sets are selected from the corpus of comparable texts described in Chapter 5. We make use of a state-of-the-art PropBank/NomBank-style semantic parser to first identify PAS in each text (Bohnet, 2010; Björkelund et al., 2010). We show a structured illustration of the output of this parser in Figure 6.1.

work	--A0--	The Russian military
	--A1--	to save a small submarine
	--MNR--	desparately
	--TMP--	on Friday
save	--A0--	The Russian military
	--A1--	a small submarine

Figure 6.1.: Parser output for the sentence "The Russian military worked desperately on Friday to save a small submarine." Here, predicates are displayed on the left, argument labels in the center and argument realizations on the right.

Based on the acquired predicate-argument structures, we create a manually annotated development and test set. This data set is crucial to develop and evaluate methods to automatically aligning PAS. We introduce such methods in Sections 6.3 and 6.4 of this chapter. In the remainder of this section, we describe the manual alignment process (Section 6.1.1) and resulting data set (Section 6.1.2).

6.1.1. Manual Annotation

We selected 70 document pairs from our data set of comparable texts (cf. Chapter 5) and asked two annotators to manually align predicate-argument structures obtained from pre-processing. Both annotators were students in Computational Linguistics, one undergraduate and one postgraduate. The texts were selected with the constraint that each text consists of 100 to 300 words. We chose this constraint as longer text pairs seemed to contain a higher number of unrelated predicates, making the alignment tasks difficult to manage for the annotators.

Both annotators received detailed guidelines that describe alignment requirements and the overall procedure (cf. Appendix A). We summarize essential details in the following.

Sure and possible links. Following standard practice in word alignment tasks (cf. Cohn et al., 2008), the annotators were instructed to distinguish between *sure* (S) and *possible* (P) alignments, depending on how certainly, in their opinion, two predicates (including their arguments) describe the same event, state or entity. The following examples show cases of predicate pairings marked as sure (18) and as possible alignments (19):

(18) The regulator ruled on September 27 that Nasdaq too was qualified to bid for OMX [...]

The authority [...] had already approved a similar application by Nasdaq.

(19) Myanmar's military government said earlier this year it has released some 220 political prisoners [...]

The government has been regularly releasing members of Suu Kyi's National League for Democracy party [...]

Replaceability. As a rule of thumb for deciding whether to align two structures, annotators were told to check how well the affected predicate-argument structures could be replaced by one another in their given context.

Missing context. In case one text does not provide enough context to decide whether two predicates in the paired documents refer to the same event, an alignment should not be marked as sure.

Similar predicates. Annotators were told explicitly that sure links can be used even if two predicates are semantically different but have the same meaning in context. Example (20) illustrates such a case:

(20) The volcano roared back to life two weeks ago.

It began erupting last month.

1-to-1 vs. n-to-m. We asked the annotators to find as many 1-to-1 correspondences as possible and to prefer 1-to-1 matches over n-to-m alignments. In case of multiple mentions of the same event, we further asked the annotators to provide only one sure link per predicate and mark remaining cases as possible links. Two comparable texts that involve multiple mentions of the same event are shown in Examples (21) and (22).

(21) a. Susan Boyle said she will sing in front of Britain's Prince Charles (...)

b. "It's going to be a privilege to be performing before His Royal Highness," the singer said (...)

c. British copyright laws will allow her to perform the hit in front of the prince and his wife.

(22) a. British singing sensation Susan Boyle is going to perform for Prince Charles.

b. (...) The show star will perform her version of Perfect Day for Charles and his wife Camilla.

Given such cases, annotators were asked to only align the PAS pairs with the highest information overlap with sure links. Following this rule, the predicate PERFORM in (23c) should be aligned with PERFORM in (24b) as both cases contain information on the song and the audience. If there is no difference in information overlap, the predicate pair that occurs first in both texts should be marked as a sure alignment. For example, this rule applies to the predicate pair SING and PERFORM in Example (23a) and (24a), respectively. The intuition behind this guideline is that the first mention introduces the actual event while later mentions just (co-)refer or add further information.

6.1.2. Development and Evaluation Data

In total, the annotators (A/B) aligned 487/451 sure and 221/180 possible alignments. Following Brockett (2007), we computed agreement on labeled annotations, including unaligned predicate pairs as an additional *null* category. Following Fleiss et al. (1981), the resulting Kappa score is 0.62, with per category scores of 0.74 and 0.19 for *sure* and *possible* alignments, respectively. The numbers show that both annotators substantially agree on which pairs of predicate-argument structures "surely" express the same proposition. Identifying further references to the same event or state, in contrast, can only be achieved with fairly low agreement. For the construction of a gold standard, we hence only view alignments as *sure* that were marked as such by both annotators and treat all *possible* alignments as optional. We further resolved cases that involved a sure alignment on which the annotators disagreed in a group discussion and added them to our gold standard accordingly. We split the final corpus into a development set of 10 document pairs and an evaluation set of 60 document pairs.

	Development	**Evaluation**
number of text pairs	10	60
number of pre-processed predicates		
all predicates (average)	395 (39.5)	3,453 (57.5)
nouns only (average)	168 (16.8)	1,531 (25.5)
verbs only (average)	227 (22.7)	1,922 (32.0)
number of alignments		
all alignments (average)	87 (8.7)	807 (13.4)
sure only (average)	35 (3.5)	446 (7.4)
possible only (average)	43 (4.3)	361 (6.0)
properties of aligned PAS		
same POS (nouns/verbs)	88.5% (24/42)	82.4% (242/423)
same lemma (total)	53.8% (42)	47.5% (383)
unequal number of arguments (total)	30.8% (24)	39.7% (320)

Table 6.1.: Statistics on predicates and alignments in the annotated data sets

Table 6.1 summarizes information about the resulting annotations in the development

and evaluation set. As can be seen from the numbers, the documents in the development set contain a fewer number of predicates (39.5 vs. 57.6) and alignments (8.7 vs. 13.4) on average. The fraction of aligned predicates is, however, about the same (22.0% vs. 23.3%). Across both data sets, the average number of observed predicates is approximately 55, of which 31 are verbs and 24 are nouns. In the development and evaluation sets, the average number of sure alignments are 3.5 and 7.4. From all aligned predicate pairs in both data sets, 82.6% are of the same part of speech (30.0% both nouns, 52.6% both verbs). In total, 48.0% of all alignments are between predicates of identical lemmata. As a rough indicator for diverging argument structures captured in the annotated alignments, we analyzed the number of aligned predicates that involve a different number of realized arguments. In both data sets together, this criterion applied in 344 cases (38.9% of all alignments).

In the next section, we discuss various similarity measures that can be employed to automatically identify corresponding pairs of predicate-argument structures.

6.2. Similarity Measures

The goal of Step 2 of our induction approach (cf. Chapter 4) is to automatically align predicates and their associated argument structures. We make use of the manually annotated data set, presented in Section 6.1, for developing and evaluating computational models for this task. To align structures with one another in our framework, we compare predicate-argument structures and compute pairwise similarities. We employ a total of seven different measures that can broadly be categorized into three classes: *predicate-specific*, *argument-specific* and *discourse-specific* measures. All seven similarity measures make use of complementary information that is typed-based or token-based (cf. Table 6.2).

	type based	token based
Predicate-specific measures		
Similarity in WordNet	X	-
Similarity in VerbNet	X	(X)
Similarity in a Semantic Space	X	-
Argument-specific measures		
Bag-of-Words Similarity	-	X
Head of Arguments Similarity	-	X
Discourse-specific measures		
Relative Discourse Position	-	X
Context Similarity	-	X

Table 6.2.: List of similarity measures applied in our alignment model.

Only the VerbNet similarity measure makes use of both type-based (mapping from predicates to VerbNet) and token-based information (word sense of a predicate in context). General information regarding instances of predicates (such as parts-of-speech and word senses), arguments (such as argument labels and syntactic structures), and joint occurrences (for example, frequency counts) are based on our pre-processed corpus introduced in Chapter 5. In this section, we discuss similarity measures in more detail. We present a graph-based model that combines these measures in Section 6.3.

6.2.1. Similarity in WordNet

WordNet is a large lexical database of English words (Fellbaum, 1998). It defines various semantic relations (e.g. hyponymy and meronymy) between so-called *synsets* – groups of nouns, verbs and adjectives that are "cognitive synonyms". A wide range of measures have previously been proposed in the literature to compute similarity based on this resource. These measures can be classified according to the kind of information they use: distance (or *path length*) between concepts in the (hyperonomy) taxonomy (for example, Leacock-Chodorow distance; Leacock and Chodorow, 1998), *information content* between two concepts and their *least common subsumer* – the closest common hyperonym of two synsets – (for example, Resnik's measure; Resnik, 1999), and textual overlap in WordNet glosses (for example, using the Lesk algorithm; Lesk, 1986).

In our approach, we rely on a measure based on information content, which we also refer to as "Lin's measure" (Lin, 1998). In preliminary experiments, we found this choice to be most reliable as distance-based measures suffer from the fact that the WordNet hierarchy is more detailed (or deep) for some concepts than others; similarly, overlap-based measures suffer from the fact that WordNet glosses can be of different length and are not available for all synsets. Given all *synsets* that contain the two predicates p_1, p_2, we compute the maximal similarity using the information theoretic measure described by Lin (1998). Our implementation exploits the WordNet hierarchy (Fellbaum, 1998) to find the synset of the least common subsumer (lcs) and uses pre-computed Information Content (IC) files from Pedersen et al. (2004) to compute similarity according to Equation (6.1). Following Pedersen et al., the information content of a concept represents its "specificity" and is defined as the relative frequency of instances that can be found in a sense-tagged corpus.

$$\text{sim}_{\text{WN}}(p_1, p_2) = \max_{s_1 \in \text{synsets}(p_1),\ s_2 \in \text{synsets}(p_2)} \frac{\text{IC}(\text{lcs}(s_1, s_2))}{\text{IC}(s_1) * \text{IC}(s_2)} \tag{6.1}$$

In order to compute similarities between verbal and nominal predicates, we further use derivational information from NomBank (Meyers et al., 2008): if a noun is a nominalization of a verbal predicate, we resort to the corresponding verb synset. In some cases, no relation can be found between two predicates, for example, because WordNet does not contain one of the predicates. Instead of leaving similarity in such cases at zero, we use a default value that we empirically set by computing the average similarity over one million randomly sampled predicate pairs that occur in our corpus and in WordNet.

6.2.2. Similarity in VerbNet

We make use of an additional resource called VerbNet (Kipper et al., 2008), which has been specifically designed as a "hierarchical domain-independent, broad-coverage verb lexicon". VerbNet is more suitable to compute similarities between verbs as the verb hierarchy in WordNet exhibits systematic problems such as isolated synsets and cyclic hyperonymy relations (for details, see Richens, 2008). Verbs in VerbNet are categorized into classes according to their meaning as well as their syntactic behavior. A verb class C can recursively embed sub-classes $C_s \in \text{sub}(C)$ that represent finer semantic and syntactic distinctions. In Equation (6.2), we define a simple similarity function that defines fixed similarity scores between 0 and 1 for pairs of predicates p_1, p_2 depending on their relatedness within the VerbNet class hierarchy. Note that we can identify a unique class for each PropBank predicate using the mapping from PropBank to VerbNet defined in SemLink (Palmer, 2009).[1]

$$\text{sim}_{\text{VN}}(p_1, p_2) = \begin{cases} 1.0 & \text{if } \exists C : p_1, p_2 \in C \\ 0.8 & \text{if } \exists C, C_s : C_s \in \text{sub}(C) \\ & \quad \wedge\ ((p_1 \in\ C, p_2 \in\ C_s) \vee (p_1 \in\ C_s, p_2 \in\ C)) \\ \textit{default} & \text{else} \end{cases} \tag{6.2}$$

Again, we define the $default$ value empirically by computing the average similarity over one million predicate pairs that occur in our corpus. Note that while computing this average value, we assign predicate pairs a score of 0.0 if a predicate is not present in the VerbNet hierarchy.

6.2.3. Similarity in a Semantic Space

As predicates can be absent from WordNet and VerbNet, or distributed over separate hierarchies due to different parts-of-speech (verbal vs. nominal predicates), we additionally calculate similarity based on distributional meaning in a semantic space (Landauer and Dumais, 1997). This measure is based on the similarity of contexts of two given predicates over all their instances in a corpus. To compute this measure, we first calculate the Pointwise Mutual Information (PMI) for each predicate $p \in \{p_1, p_2\}$ and the n most frequent context words $c \in C$ following Equation (6.3).

$$\text{pmi}(p, c) = \frac{\text{freq}(p, c)}{\text{freq}(p) * \text{freq}(c)} \tag{6.3}$$

The joint frequency of two words $\text{freq}(p, c)$ is calculated as the number of times c appears in a context window of an occurrence of p. As we are dealing with predicates of different parts-of-speech, we define context windows in terms of neighboring words instead of relying on syntactic dependencies as proposed in more recent approaches to distributional semantics (Padó and Lapata, 2007; Erk and Padó, 2008; Baroni and Lenci,

[1] cf. `http://verbs.colorado.edu/semlink/`

2010). More precisely, we use context windows of five words to the left and to the right, and compute the PMI of each predicate and the 2,000 most frequent context words. The same setting has been successfully applied in related tasks, including word sense disambiguation (Guo and Diab, 2011) and measuring phrase similarity (Mitchell and Lapata, 2010). Considering context words $c_1 \dots c_{2,000} \in C$ as dimensions of a vector space, we represent predicates as vectors following Equation (6.4).

$$\vec{p} = (\text{pmi}(p, c_1), \text{pmi}(p, c_2), \dots, \text{pmi}(p, c_{2,000})) \tag{6.4}$$

Given the vector representations of two predicates, we calculate their similarity using the cosine function of the angle between the two vectors as defined in Equation (6.5).

$$\text{sim}_{Dist}(p_1, p_2) = \frac{\vec{p_1} \cdot \vec{p_2}}{\|\vec{p_1}\| * \|\vec{p_2}\|} \tag{6.5}$$

6.2.4. Bag-of-Words Similarity

The three similarity measures presented in the previous subsections only compare predicate pairs with each other. In addition to these measures, we define two token-based similarity measures that take into account the similarity between arguments realized in two predicate-argument structures. The first of these two measures is a simple measure based on bag-of-words, which we compute by measuring the overlap of word tokens over all (concatenated) arguments of each predicate-argument structure. Formally, the measure considers all arguments $a_1 \in A_1$ and $a_2 \in A_2$ associated with the predicates p_1 and p_2, respectively, and calculates overlap as defined in Equation (6.6).

$$\text{sim}_{\text{ABoW}}(p_1, p_2) = \frac{\sum_{w \in A_1 \cap A_2} \text{idf}(w)}{\sum_{w \in A_1} \text{idf}(w) + \sum_{w \in A_2} \text{idf}(w)} \tag{6.6}$$

In order to control the impact of frequently occurring words, we weight each word by its Inverse Document Frequency (IDF), which we calculate over all documents d in our corpus D:

$$\text{idf}(w) = \log \frac{|D|}{|\{d \in D | w \in \text{words}(d)\}|} \tag{6.7}$$

6.2.5. Head of Arguments Similarity

In addition to the bag-of-words similarity, which computes one score over all pairs of arguments in two predicate-argument structures (PAS), we define an argument-specific measure that only compares the semantic heads of arguments that have the same argument label. That is, given two PAS that consist of predicates p_1, p_2 and arguments labeled A0 and A1, we compute the similarity of the two arguments labeled A0 (also denoted as label(a) = 'A0') and the similarity of the two arguments labeled A1 (label(a) = 'A1'). Each similarity between argument heads is computed using Lin's measure as described in Section 6.2.1. We extract the semantic head of each argument

by considering the dependency tree of the parser output, in which we look for the noun or verb on the highest level within the argument span. Finally, we collapse all pairwise argument similarities into one measure by taking the average following Equation (6.8).

$$\text{sim}_{\text{Aheads}}(p_1, p_2) = \frac{\sum_{\{a_1,a_2|\text{label}(a_1)==\text{label}(a_2)\}} \text{sim}_{\text{WN}}(\text{head}(a_1), \text{head}(a_2))}{|\{a_1, a_2|\text{label}(a_1) == \text{label}(a_2)\}|} \tag{6.8}$$

Note that this measure is not applicable to arguments whose semantic head is a pronoun or a proper noun absent from WordNet. Thus, we expect that the measure is helpful in terms of precision but has a negative effect on recall.

6.2.6. Relative Discourse Position

The similarity measures introduced in the above sections consider type-based information about predicates (cf. Sections 6.2.1, 6.2.2 and 6.2.3) as well as token-based information regarding associated arguments (cf. Sections 6.2.4 and 6.2.5). In this and the following section, we introduce two additional measures that take into account discourse context.

One aspect of discourse is the relative position, in which a predicate-argument structure occurs. For example, we expect important information to be located towards the beginning of a newswire article, while background information and minor details should be conveyed at a later point. Given two predicates, we measure their similarity with respect to this assumption as one minus the absolute difference between their relative positions in discourse (cf. Equation (6.9)). The relative position in discourse is computed as the *sentence_index* in which the predicate p_1 or p_2 occurs, divided by the total number of sentences in the affected document (d_1 or d_2, respectively). To make sure that the relative position ranges from 0.0 (first sentence) to 1.0 (last sentence), we enumerate all sentences starting with zero and define the *length* of a document as the index of its last sentence.

$$\text{sim}_{DPos}(p_1, p_2) = 1 - \left(\left| \frac{\text{sentence\_index}(p_1)}{\text{length}(d_1)} - \frac{\text{sentence\_index}(p_2)}{\text{length}(d_2)} \right| \right) \tag{6.9}$$

That is, if one predicate is at the beginning of a text and the other is at the very end, the distance between their relative positions will be 1.0, leading to a similarity of 0.0. In contrast, if both predicates are in identical relative positions, the distance between them will be 0.0, meaning that the resulting similarity is 1.0.

6.2.7. Context Similarity

To compute the discourse similarity between two predicates, we further consider occurrences of other (shared) predicates in the immediate discourse context. Given two predicates, this similarity is computed as the relative number of overlapping predicate types within the preceding and succeeding n neighboring predicates as defined in Equation (6.10).

$$\begin{aligned} \text{sim}_{DCon}(p_1, p_2) =& \frac{\text{context}(p_1) \cap \text{context}(p_2)}{\text{context}(p_1) \cup \text{context}(p_2)}, \text{with} \\ \text{context}(p) =& \{p' | index(p') \in [index(p) - n : index(p) + n]\} \end{aligned} \tag{6.10}$$

We compute the *index* of a predicate as the number of preceding predicates within the same text. That is, the index of the first predicate in a text is zero, the index of the second predicate is one, etc. We experimented with different values for n on our development corpus (cf. Section 6.1.2) and empirically set this number to five.

6.3. Graph Representation and Clustering

Based on the similarity measures described in Section 6.2, we build a graph representation of each document pair from our corpus of comparable texts (cf. Chapter 5). In each graph, predicate-argument structures (PAS) are represented as nodes and pairs of PAS from two different texts are connected with undirected, weighted edges. The weight of each edge is computed using the described similarity measures. The purpose of the constructed graph representations is to provide input for a graph-based clustering approach, which we use to divide the set of all predicate-argument structures into subsets that should be aligned.

As a result of the graph construction, each document pair in our data set is represented by a weighted, undirected and bipartite graph. Given the sets of predicate-argument structures P_1 and P_2 of two comparable texts T_1 and T_2, respectively, we formally define the graph G_{P_1,P_2} following Equation (6.11).

$$G_{P_1,P_2} = \langle V, E \rangle \quad \text{where} \quad \begin{array}{l} V = P_1 \cup P_2 \\ E = P_1 \times P_2 \end{array} \tag{6.11}$$

Edge weights. We specify the edge weight between two nodes that represent the predicate-argument structures $p_1 \in P_1$ and $p_2 \in P_2$ as a weighted linear combination of the similarity measures described in Section 6.2.

$$w_{p_1 p_2} = \Sigma_i \quad \lambda_i * \text{sim}_\text{i}(p_1, p_2) \tag{6.12}$$

Initially we set all weighting parameters λ_i to have a uniform weight. In Section 6.4, we define an optimized weighting setting for the individual similarity measures. To make the range of each similarity measure comparable, we additionally normalize the output of each measure to have a mean value of 0.5 and a range of [0.0:1.0]. We compute the normalization parameters for this step using one million predicate pairs that we randomly sampled from our corpus of comparable texts.

Clustering. The goal of this step of our induction approach is to identify pairs of predicate-argument structures that describe the same event, state or entity. We represent pairs of comparable texts as graphs to leverage graph-clustering techniques, which are well-known for their strong performance on NLP tasks (Su and Markert, 2009; Chen and Ji, 2010; Cai and Strube, 2010, inter alia). In the next sections, we first give an overview of graph clustering and different algorithms (Section 6.3.1); we then describe our own adaptation of a specific clustering technique (Flake et al., 2004) to the task of aligning predicate-argument structures (Section 6.3.2).

6.3.1. Alignment via Graph Clustering

In general, *clustering* refers to the process of dividing a set of data points into groups that contain similar points. We use the term *graph clustering* to refer to clustering processes that are applied to graph structures, that is, sets of nodes and edges that connect these nodes. In context of our alignment approach, nodes represent predicate-argument structures (PAS) while edges represent undirected and weighted similarities between pairs of PAS. Each weight is calculated following Equation (6.12). Following the definition by Schaeffer (2007), the goal of graph clustering is to group the nodes in such a way that the sum of weights within a cluster should be as high as possibly while keeping the sum of weights between clusters as low as possible.

Given a set of nodes and edges, there are various different methods that can be applied to cluster groups containing similar elements. Most existing clustering methods can be classified into one of the following two categories: divisive and agglomerative.

- **Divisive algorithms** partition graphs in a top-down fashion. That is, they start from a state where all nodes of a graph are assigned to a single cluster and then recursively partition each cluster into smaller clusters until a pre-specified stopping criterion is reached. Examples of divisive algorithms are cut-based and spectral clustering techniques (Flake et al., 2004; Shi and Malik, 2000, inter alia).
- In contrast, **agglomerative algorithms** proceed from a state where each node is assigned to its own *singleton* cluster and clusters are iteratively merged or modified until a specific optimization criterion is reached. An example for this category is the Chinese Whispers algorithm proposed by Biemann (2006).

In our setting, each text pair has different properties: texts are of different length, text pairs can share a different amount of information and texts within each pair can be more or less detailed. As all of these factors influence properties of our graph representations and expected clustering outcome, it is difficult to make any general assumptions about the properties of desired clusters. As a consequence, clustering techniques that aim to maximize one and the same graph-theoretic measure for every input might not be well-suited for this task. We hence omit discussion on various optimization criteria that have been proposed in the literature. A comprehensive overview of optimization methods and criteria can be found, however, in survey articles by Newman (2004), Schaeffer (2007) and Chen et al. (2010).

There exists one restriction though that we know about the desired clustering outcome: namely the size that each cluster should have. More specifically, we want pairs of predicate-argument structures that should be aligned across two texts to form clusters of two nodes; in contrast, if a PAS only occurs in one of the two texts, the structure should be represented by one node in a singleton cluster. This precondition fits well with clustering approaches that can iteratively be applied until all clusters comply with the pre-specified cluster size. In the following sections, we describe two such techniques:

- In Section 6.3.2, we introduce a novel divisive clustering approach that makes use of minimum cuts (Flake et al., 2004) to recursively cluster a bipartite graph into smaller subgraphs.
- In Section 6.4.1, we describe an agglomerative approach – as a baseline model – that simply clusters the most similar pairs of predicate-argument structures.

6.3.2. Minimum Cuts

The clustering method used in our model relies on so-called minimum cuts (henceforth also called *mincuts*) in order to partition a bipartite graph, representing pairs of texts, into clusters of aligned predicate-argument structures. A mincut operation divides a given graph into two disjoint subgraphs. Each minimum cut is performed as a cut between some source node s and some target node t, such that (1) each of the two nodes will be in a different subgraph and (2) the sum of weights of all removed edges will be as small as possible. We implement basic graph operations using the freely available Java library JGraph[2] and determine each mincut using the method by Goldberg and Tarjan (1986).

In our initial graph representation, all nodes that represent predicates of one text are connected to each node that represents a predicate from a comparable text. As our goal is to induce clusters that correspond to pairs of corresponding structures, we set a maximum number of two nodes per cluster as stopping criterion. Given an input graph G, our algorithm recursively applies mincuts in three steps as described in Algorithm 1. Step 1 identifies the edge e with lowest weight in the given graph G. Step 2 performs the actual mincut operation on G. Finally, the stopping criterion and recursion are applied in Step 3. An example of a clustered graph is illustrated in Figure 6.2.

The advantage of our method compared to off-the-shelf clustering techniques is twofold: on the one hand, the clustering algorithm is free of any parameters, such as the number of clusters or a clustering threshold, that require fine-tuning; on the other hand, the approach makes use of a termination criterion that very well represents the nature of the goal of our task, namely to align pairs of predicate-argument structures across comparable texts. In Section 6.4, we further provide empirical evidence for the advantage of this approach.

[2] cf. `http://jgrapht.org/`

```
function CLUSTER(G)
    clusters ← ∅
    E ← GETEDGES(G)                                    ▷ Step 1
    e ← GETEDGEWITHLOWESTWEIGHT(E)
    s ← GETSOURCENODE(e)
    t ← GETTARGETNODE(e)
    G′ ← MINCUT(G, s, t)                               ▷ Step 2
    C ← GETCONNECTEDCOMPONENTS(G′)
    for all G_s ∈ C do                                 ▷ Step 3
        if SIZE(G_s) <= 2 then
            clusters ← clusters ∪ G_s
        else
            clusters ← clusters ∪ CLUSTER(G_s)
        end if
    end for
    return clusters;
end function
```

Algorithm 1: Pseudo code of our clustering algorithm

6.4. Experiments

This section describes the evaluation of our graph-based clustering model on the task of aligning predicate-argument structures across comparable texts. For comparison to related tasks and methods, we describe different evaluation settings, various baselines, and their results. In order to benchmark our model against traditional methods for word alignment (cf. Chapter 3), we first apply our graph-based alignment model on three sentence-based paraphrase corpora. For this task, we do not have any discourse-level information and hence only use a subset of the similarity measures defined in Section 6.2. More specifically, we use the following measures: Similarity in WordNet, Similarity in VerbNet, Similarity in a Semantic Space and Bag-of-Words Similarity. As this subset corresponds to the measures applied in our EMNLP paper (Roth and Frank, 2012b), we also refer to this simplified model as **EMNLP'12**.

For evaluation on the novel task of aligning predicate-argument structures across monolingual comparable texts, we use all similarity measures defined in Section 6.2. We henceforth refer to this model as **Full**. Both models, **Full** and **EMNLP'12**, make use of the clustering algorithm introduced in Section 6.3. In the setting with comparable texts, henceforth also called *discourse-level evaluation*, we evaluate our model against various baselines and against a model from the literature that has recently been proposed for this task (Wolfe et al., 2013). Similar to our model, Wolfe et al. use various resources to calculate the similarity of two predicate-argument structures. Differences to our approach lie in the utilized resources, the use of additional data to learn feature weights, and the fact that each alignment decision is made using a binary classifier.

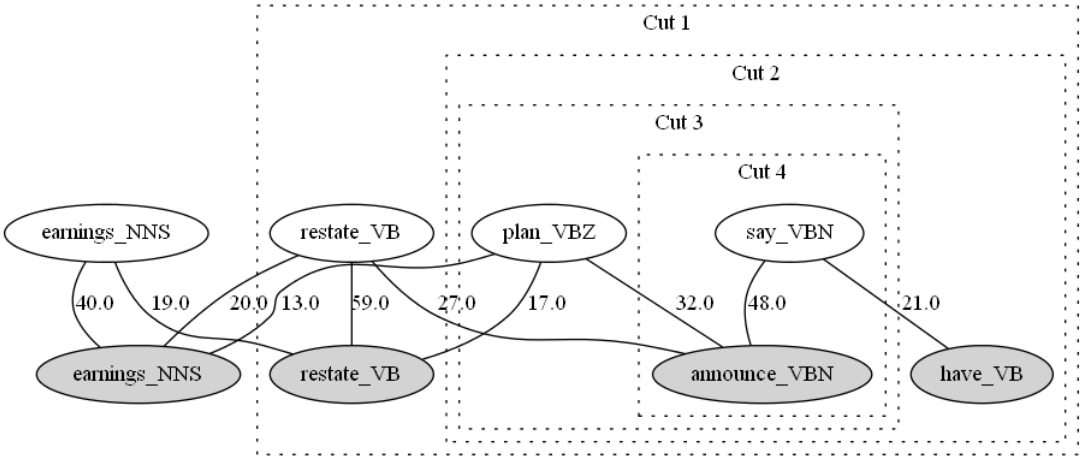

Figure 6.2.: The predicates of two sentences (white: "The company has said it plans to restate its earnings for 2000 through 2002."; grey: "The company had announced in January that it would have to restate earnings (...)") from the Microsoft Research Paragraph Corpus are aligned by computing clusters with minimum cuts.

Discourse-level evaluation is performed based on the manually annotated evaluation data set described in Section 6.1.

To gain more insight into the performance of our different similarity measures included in the **Full** model, we evaluate simplified versions that omit individual similarity measures (**Full**–[**measure name**]). The relative differences in performance against various baselines help us quantify difficulties between a traditional sentence-based word alignment setting and our novel alignment task that operates on full texts.

6.4.1. Baselines

A simple baseline for predicate alignment is to simply cluster all predicates that have identical lemmata (henceforth called **LemmaId**). To assess the benefits of the clustering step, we propose a second baseline that uses the same similarity measures as our **Full** model but does not use the mincut clustering described in Section 6.3.2. Instead, it greedily merges as many 1-to-1 alignments as possible, starting with the highest similarity (**Greedy**).

As a more sophisticated baseline, we make use of alignment tools commonly used in statistical machine translation (SMT, cf. Chapter 3). For the sentence-level evaluation, Cohn et al. (2008) readily provide GIZA++ alignments (Och and Ney, 2003) as part of their word-aligned paraphrase corpus. For the experiments in the discourse-level setting, we train our own word alignment model using the state-of-the-art word alignment tool Berkeley Aligner (Liang et al., 2006). As word alignment tools require pairs of sentences as input, we first extract paraphrases using a re-implementation of a previously proposed paraphrase detection system (Wan et al., 2006), which performs closely to the state-of-

the-art (Socher et al., 2011). In the following section, we abbreviate both baselines using SMT alignment tools as **WordAlign**.

6.4.2. Sentence-level Evaluation

For sentence-level evaluation, we make use of the following three corpora that are word-aligned subsets of the paraphrase collections described in Cohn et al. (2008): **MTC** consists of 100 sentence pairs from the Multiple-Translation Chinese Corpus (Huang et al., 2002), **Leagues** contains 100 sentential paraphrases from two translations of Jules Verne's "Twenty Thousand Leagues Under the Sea", and **MSR** is a subset of the Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005), consisting of 130 sentence pairs. All three paraphrase collections are in English. We use all sentence pairs as contained in the collections, meaning that we do not perform any paraphrase detection ourselves in this setting. To determine alignment candidates within sentence pairs, we apply the same pre-processing steps as used for our corpus of comparable texts (cf. Section 6.1). The semantic parser identified an average number of 3.8, 5.1 and 4.7 predicates per text (i.e., per paraphrase sentence) in **MTC**, **Leagues** and **MSR**, respectively. All models are evaluated against the subset of gold standard alignments between pairs of words marked as predicates (for details, cf. Cohn et al., 2008).

Results The results for **MTC**, **Leagues** and **MSR** are presented in Table 6.3. The numbers indicate that **WordAlign** consistently outperforms all other models on the three data sets in terms of precision and F_1-score. Not all differences in terms of recall are significant though due to high variance of results compared to data set sizes.

	MTC			**Leagues**			**MSR**		
	P	R	F_1	P	R	F_1	P	R	F_1
LemmaId	95**	75	84	97*	67**	79**	98**	91	94
Greedy	75**	**88****	81	75**	**86****	80	81**	**97****	88
WordAlign	**99****	87**	**93****	**99****	79	**87****	**100****	96*	**98****
EMNLP'12	92	72	81	93	69	79	95	88	91

Table 6.3.: Results for sentence-based alignment in the three benchmark settings **MTC**, **Leagues** and **MSR** (all numbers in %); results that significantly differ from **EMNLP'12** are marked with asterisks (* $p<0.05$; ** $p<0.01$).

The overall performance of **WordAlign** does not come much as a surprise, seeing that all three data sets consist of highly parallel sentence pairs. In fact, the results for **LemmaId** show that by aligning all predicates with identical lemmas, most of the sure alignments in the three settings are covered with high precision. Remaining errors of **LemmaId** result from the fact that the same lemma can occur multiple times in the same paraphrase, a phenomenon that is better handled by **WordAlign**. Interestingly, the **Greedy** model achieves the highest recall in all settings, demonstrating that our

combined similarity measures capture more information than any other method. The addition of our clustering method helps to achieve significantly higher precision ($p<0.01$). The reason why precision still lies below other models in this setting is related to the fact that the only stopping criterion for the clustering algorithm is that clusters should contain a maximum of two nodes. As a consequence, the clustering sometimes stops too early, leaving two predicate-argument structures in a cluster that are not related to one another. One way to solve this problem would be to require a small alignment threshold. In the discourse-level setting, we discuss the general need for a threshold during graph construction, hence this extra parameter will not be necessary in the clustering step.

6.4.3. Discourse-level Evaluation

For discourse-level evaluation, we compare the performance of all models against the annotated gold standard alignments between predicate-argument structures, as described in Section 6.1. Since all text pairs in our corpus comprise multiple sentences each, the average number of predicates per text to consider is much higher than in the sentence-level setting (approximately 28 vs. 5). As the full graph representation becomes rather inefficient to handle (by default, edges are inserted between all predicate pairs), we use the development set of 10 text pairs to estimate a similarity threshold for adding edges. We tested all thresholds from 0.0 to 1.0 with a step-size of 0.05 and found 0.75 to perform best. This threshold is applied in the evaluation of all models that make use of the similarity measures described in Section 6.2.

Results. The results for the discourse-level setting is presented in Table 6.4. From all approaches, **Greedy** and **WordAlign** yield lowest performance. For **WordAlign**, we observe two main reasons. On the one hand, sentence paraphrase detection does not perform perfectly. Hence, the extracted sentence pairs do not always contain gold alignments. On the other hand, even sentence pairs that contain gold alignments are less parallel than in the previous setting, which makes them harder to align in general. The increased difficulty can also be seen in the results for the **Greedy** model, which only achieves an F_1-score of 17.2% in this setting. In contrast, we observe that the majority of all sure alignments can be retrieved by applying the **LemmaId** model (60.3% recall).

The **Full** model achieves a recall of 48.9% but significantly outperforms all baselines ($p<0.01$) in terms of precision (71.8%). This is an important factor for us as we plan to use the alignments in subsequent tasks. With 58.2%, **Full** also achieves the best overall F_1-score. By comparing the results with those of the **EMNLP'12** model, we can see that the discourse-level similarity measures provide a significant improvement in terms of precision without a considerable loss in recall. This advantage in precision can also be seen in comparison to **Wolfe et al.**. In contrast, their system outperforms our model with respect to recall. There are two main reasons for this: on the one hand, their model makes use of much larger resources to compute alignments, including a paraphrasing database that contains over 7 million rewriting rules; on the other hand, their model is supervised and makes use of additional data to learn weights for each of their features. In contrast, **Full** and **EMNLP'12** only make use of a small development

	P	R	F_1
	Baselines		
LemmaId	40.3**	60.3*	48.3**
Greedy	12.5**	27.6**	17.2**
WordAlign	19.7**	15.2**	17.2**
	Previous work		
Wolfe et al.	52.4**	**64.0****	57.6
EMNLP'12	58.7**	46.6	52.0
	This thesis		
Full	**71.8**	48.9	**58.2**

	P	R	F_1
$-\mathbf{sim_{WN}}$	78.8	44.6**	57.0
$-\mathbf{sim_{VN}}$	78.7	44.6**	57.0
$-\mathbf{sim_{Dist}}$	77.3	45.5**	57.3
$-\mathbf{sim_{ABoW}}$	67.6*	49.3	57.0
$-\mathbf{sim_{Aheads}}$	68.9	**52.9****	**59.8**
$-\mathbf{sim_{DPos}}$	64.2	43.0**	51.6**
$-\mathbf{sim_{DCon}}$	69.3*	52.2**	59.6
Full	71.8	48.9	58.2
+HighPrec	**86.2**	29.1**	43.5**

Table 6.4.: Results for discourse-level alignment in terms of precision (P), recall (R) and F_1-score (all numbers in %); left: comparison of the **Full** model to baselines and previous work; right: impact of removing individual measures and using a tuned weighting scheme; results that significantly differ from **Full** are marked with asterisks (* $p<0.05$; ** $p<0.01$).

data set to determine a threshold for graph construction. Though the difference is not significant, it is worth noting that our model outperforms that by Wolfe et al. by 0.6 percentage points in F_1-score, despite not making use of any additional data.

Ablating similarity measures. All aforementioned results were conducted in experiments with a uniform weighting scheme of similarity measures as introduced in Section 6.2. Table 6.4 shows the performance impact of individual similarity measures by removing them completely (i.e., setting their weight to 0.0). The numbers indicate that not all measures contribute positively to the overall performance when using equal weights. Except for the argument head similarity (cf. Section 6.2.5), all ablation tests revealed significant drops in performance, either with respect to precision or recall. This result highlights the importance of incorporating predicate-specific, argument-specific and discourse-specific information regarding individual predications in this task.

Tuning Weights for high precision. Subsequently, we tested various combinations of weights on our development set in order to estimate a better weighting scheme. This tuning procedure is implemented as a brute-force technique, in which random weights between 0.0 and 1.0 are assigned to each measure. For graph construction, all weights are normalized again to sum to 1.0. We additionally try different thresholds for adding edges in the graph representation. To achieve high precision, we weight precision three times higher than recall while evaluating different parameters on our development set. We repeat this brute-force assignment of parameters for 2,000 iterations. Following this process, we found the best result to be achieved with a threshold of 0.85 and the following

weights:

- 0.11, 0.14 and 0.21 for sim_{WN}, sim_{VN} and sim_{Dist}, respectively, (i.e., 46% of the total weight for predicate-specific measures)
- 0.21 and 0.05 for sim_{ABoW} and $\text{sim}_{\text{Aheads}}$, respectively, and (i.e., 26% of the total weight for argument-specific measures)
- 0.21 and 0.07 for sim_{DPos} and sim_{DCon}, respectively. (i.e., 28% of the total weight for discourse-specific measures)

The weighting scheme shows that information from all categories contribute to achieving the best precision. When applying the tuned model on our evaluation data set, we note that results in recall drop to 29.1% (−19.8 percentage points). Precision, on the other hand, increases up to 86.2% (+14.4 percentage points).

6.5. Summary

In this chapter, we introduced the task of aligning predicate-argument structures across monolingual comparable texts. We designed annotation guidelines and created a data set of gold standard alignments. Based on this data set, we developed and evaluated a novel clustering-based alignment model that uses a combination of various similarity measures and a graph-based clustering algorithm that we specifically designed for this task. In our intrinsic evaluation, we showed that our novel model outperforms a range of baselines as well as previous approaches to this particular task. At the same time, our model achieves competitive performance in a traditional word alignment setting that involves texts that are parallel on the sentence level. As an additional contribution, we defined a tuning routine that can be utilized to train a high precision model for the discourse-level alignment task. Our results show that, by using this tuning step, corresponding structures in our evaluation set can be identified with a precision of 86.2%. This intermediate result is essential for the success of our overall framework. In the next chapter, we present Step 3 of our implicit argument induction technique, in which we examine pairs of automatically aligned predicate-argument structures as a means to identify and link implicit arguments.

Aligned pairs of structures could also be of interest for other lines of research. In particular, we expect that differing predicates or arguments in aligned structures could serve as a basis for inducing annotations for other linguistic phenomena. For example, two arguments in a pair of aligned structures that have the same label, but do not refer to the same entity, could be interesting instances of bridging relations (Hou et al., 2013). Wolfe et al. (2013) argue that alignments between predicate-argument structures could also be used to disambiguate events in text for tasks such as textual entailment and question answering. We discuss applications in more detail in Chapter 9.

7. Inducing Implicit Arguments

> $\boldsymbol{q}_{iii}$: *"How can we find antecedents for implicit arguments within their discourse context?"*

Pairs of aligned predicate-argument structures (PAS) present an ideal basis for the detection of implicit arguments. That is, arguments in each structure can be compared in order to find instances that are present, or *explicit*, in one PAS but absent, or *implicit*, from the aligned PAS. In this chapter, we present an approach to performing this comparison automatically based on semantic role annotations. More precisely, we extract annotations using a semantic role labeling (SRL) system and align its output using our high-precision alignment model introduced in Chapter 6. To determine antecedents of entity references in discourse, we further rely on automatic coreference resolution (CR) techniques. We define a few heuristic rules that can be applied on SRL and CR annotations to automatically determine instances of implicit arguments and their discourse antecedents. We hypothesize that, given the size of our corpus, we can induce a precise and reliable data set that can be helpful in various NLP applications. In particular, we aim to use the induced data set as additional training material for linking implicit arguments and for modeling realization decisions of arguments in discourse.

We previously described the automatic induction approach in the *SEM 2013 paper "Automatically Identifying Implicit Arguments to Improve Argument Linking and Coherence Modeling" (Roth and Frank, 2013). This chapter is based on parts of this paper, which have been revised and extended. In Section 7.1, we first discuss the preparation of a suitable data set, from which we induce instances of implicit arguments. In Section 7.2, we present our implementation of the induction approach outlined in Chapter 4 and based on the alignment model described in Chapter 6. Finally, in Section 7.3, we discuss the data set of implicit arguments and discourse antecedents that we obtain by applying this approach on the full corpus described in Chapter 5.

7.1. Data Preparation

As discussed in context of our induction framework, implicit arguments and their discourse antecedents can be effectively induced from pairs of comparable texts. In Section 7.2, we describe an implementation of the final stage, Step 3, of this approach. As a basis for the actual induction, we rely on several preparatory steps that identify information two documents have in common (cf. Figure 7.1). These steps are described in this section. In particular, we align corresponding predicate-argument structures using graph-based clustering as discussed in Chapter 6. We then determine co-referring entities

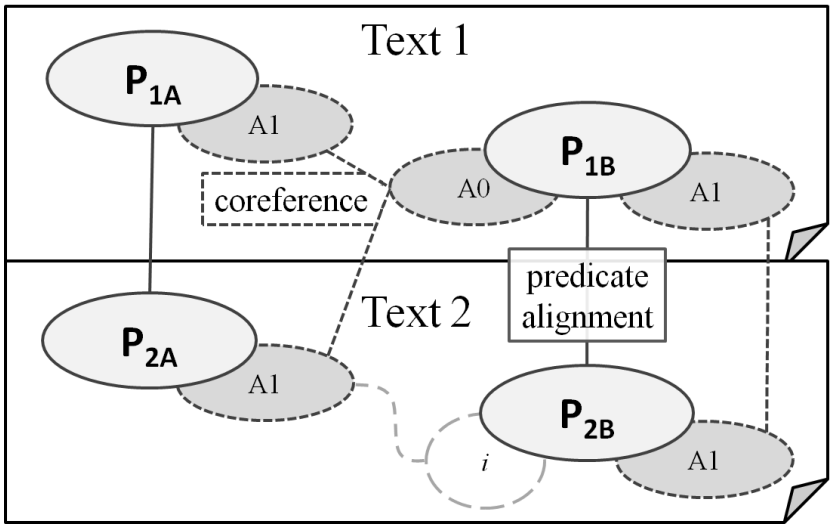

Figure 7.1.: Illustration of the induction approach: texts consist of PAS (represented by overlapping circles); we exploit alignments between corresponding predicates across texts (marked by solid lines) and co-referring entities (marked by dotted lines) to infer implicit arguments (marked by '*i*') and link antecedents (curly dashed line)

across pairs of texts using coreference resolution techniques on concatenated document pairs (Lee et al., 2012).

Single document pre-processing. We apply several preprocessing steps to all 167,728 document pairs in our data set of comparable texts, which we introduced in Chapter 5. We use the Stanford CoreNLP package[1] for tokenization and sentence splitting. We then apply MATE tools (Bohnet, 2010; Björkelund et al., 2010), including the integrated PropBank/NomBank-based semantic parser, to determine local predicate-argument structures (PAS). Finally, we resolve pronouns that occur in a PAS using the coreference resolution system by Martschat et al. (2012), which placed second for English in the CoNLL-2012 Shared Task (Pradhan et al., 2012).

High precision alignments. Once all single documents are pre-processed, we align predicate-argument structures across pairs of comparable texts. We want to induce reliable instances of implicit arguments based on aligned PAS pairs and hence apply our graph-based clustering technique using the high-precision tuning step described in Section 6.4.3. We run the high-precision model on all pairs of texts in our corpus. As a result, we extract a total number of 283,588 aligned pairs of predicate-argument structures. An overview of properties of this data set is given in Table 7.1.

[1] `http://nlp.stanford.edu/software/`

	high-precision alignments	
number of alignments	283,588	
same POS	278,970	(98.4%)
noun–noun	89,696	(31.6%)
verb–verb	189,274	(66.8%)
mixed POS	4,618	(1.6%)
same lemma	273,924	(96.6%)
different lemma	9,664	(3.4%)
same number of arguments	239,563	(84.5%)
unequal number of arguments	44,025	(15.5%)

Table 7.1.: Properties of the high precision alignment data set

Cross-document coreference. For each argument that is explicit in one PAS but implicit in an aligned PAS, we want to determine a suitable antecedent within the context of the implicit argument. We view the reference in the explicit argument as a cue and identify co-referring mentions in both texts by applying coreference resolution techniques across pairs of documents. In practice, we follow the methodology by Lee et al. (2012), who propose to apply standard coreference methods on pairs of texts by simply concatenating two documents and providing them as a single input document. We use the Stanford Coreference system (Lee et al., 2013), which applies a sequence of coreference "sieves" to the input, ordered according to their precision. To obtain a highly accurate and reliable output, we consider only the most precise resolution sieves:

- "String Match",
- "Relaxed String Match",
- "Precise Constructs",
- "Strict Head Match A", "Strict Head Match B", "Strict Head Match C", and
- "Proper Head Noun Match".

Note that none of these sieves involve pronoun resolution. Instead, we decided to use the resolved pronouns from the single-document coreference step. This decision is based on the fact that the system by Martschat et al. (2012) outperforms the Stanford system with all sieves on the CoNLL'11 test set by an average F_1-score of 3.0 absolute points. The high-precision sieves are, however, better suited for the cross-document task as we plan to rely on the resulting coreference chains for identifying potential antecedents of implicit arguments. That is, we prefer fewer but more reliable chains in order to minimize the impact of possible pre-processing errors.

7.2. Automatic Identification and Linking

Given a pair of aligned predicates from two comparable texts, we examine the parser output to identify arguments in each predicate-argument structure (PAS). We compare the set of labels assigned to the arguments in each structure to determine whether one PAS contains an argument (explicit) that has not been realized in the other PAS (implicit). For each implicit argument, we identify appropriate antecedents by considering the cross-document coreference chain of its explicit counterpart. That is, we specifically look for explicit arguments that are part of a coreference chain and that are implicit in an aligned structure. As our goal is to link arguments within discourse, we require candidate antecedents to be mentions that occur in the same document as the implicit argument. We impose a number of restrictions on the resulting pairs of implicit arguments and antecedents to reduce the impact of different types of pre-processing errors:

Mislabeled arguments. In some cases, the parser annotated the same argument in two texts using different labels. To ensure that mislabeled arguments are not recognized as implicit, we require that pairs of aligned PAS contain a different number of arguments.

Incorrectly resolved pronouns. The coreference resolution applied to single documents sometimes resolved pronouns incorrectly. To make sure that such errors do not affect implicit argument linking, we do not consider resolved pronouns as discourse antecedents.

Missed arguments. Depending on sentence structure, the semantic parser is sometimes unable to determine all local arguments. This often leads to the identification of erroneous implicit arguments. To intercept some of these cases, we require that all antecedents from the cross-document coreference chain must be outside of the sentence that contains the affected predicate-argument structure.

We further experimented with some additional restrictions such as reducing the search space to PAS with identical predicate POS and lemma. However, these additional heuristics only limited the amount of total alignments induced, without yielding improvements in terms of a lower error rate.

7.3. Resulting Data Set

We apply the outlined identification and linking approach to all text pairs in our corpus of comparable texts. As a result, we induce a total of 698 implicit argument and antecedent pairs. A summary of properties of the obtained pairs can be found in Table 7.3. The full data set involves 535 different predicates. Each pair was found in a separate document. Examples are displayed in Table 7.2. Note that 698 implicit arguments from 283,588 pairs of PAS seem to represent a fairly low recall. Most PAS pairs in the high precision data set do, however, consist of identically labeled argument sets (84.5%). In the remaining cases, in which an implicit argument can be identified (15.5%), an antecedent in discourse cannot always be found using the high precision coreference sieves. This does not mean that implicit arguments are rare in general. As discussed in Chapter 5, 38.9% of all manually aligned PAS pairs involve a different number of arguments.

[T-Online$_{\mathbf{i}}$], the leading Internet services provider in Europe and a unit of Deutsche Telekom, said Thursday its net loss more than doubled last year owing to its foreign activities and goodwill writedowns. (...) The [$\varnothing_{\mathbf{i}}$]$_{\text{A0}}$ [operating]$_{\text{A3}}$ **loss**, as measured by earnings before interest, tax, depreciation and amortisation, widened to 189 million euros last year from 121.6 million euros a year earlier.	[T-Online's$_{\mathbf{i}}$]$_{\text{A0}}$ [operating]$_{\text{A3}}$ **loss** – earnings before financial items such as interest, taxes, depreciation and amortization – also widened, to 189 million euros (dlrs 167 million) in 2001 from 122 million (dlrs 108 million).
[Mozambique$_{\mathbf{i}}$] police have arrested four foreigners in connection with an alleged plot to sabotage the African country's largest hydroelectric dam, officials said Wednesday. (...) It was handed over to Mozambican control last year, 33 years after [$\varnothing_{\mathbf{i}}$]$_{\text{A1}}$ **independence** in [in 1975]$_{\text{TMP}}$.	Its power lines and other infrastructure sustained severe damage during the 16-year civil war that followed [Mozambique's$_{\mathbf{i}}$]$_{\text{A1}}$ **independence** [in 1975]$_{\text{TMP}}$.
The accident occurred just after midnight on Sunday in Shanxi province but [local officials]$_{\text{A0}}$ failed to immediately **report** [the accident]$_{\text{A1}}$ [$\varnothing_{\mathbf{i}}$]$_{\text{A2}}$, the State Administration for Work Safety said on its website. [The government$_{\mathbf{i}}$] says 4,750 people died in coal mine accidents last year, an average of 13 a day. It is common for mine owners to delay reporting accidents or to not report them at all.	The explosion happened in a mine in the suburbs of Jincheng City on Sunday in Shanxi province, but [the coal mine owner]$_{\text{A0}}$ did [not]$_{\text{NEG}}$ [immediately$_{\text{TMP}}$] **report** [it]$_{\text{A1}}$ [to the government$_{\mathbf{i}}$]$_{\text{A2}}$, Xinhua News Agency said.

Table 7.2.: Three positive examples of automatically induced implicit arguments ($\varnothing$) and the cross-document coreference chains that include discourse antecedents (**i**); the right-hand side shows the aligned predicate-argument structures that were used to identify a suitable antecedent.

	implicit arguments and discourse antecedents
number of induced pairs	698
predicate counts	
number of nominal predicates	285 (40.8%)
number of verbal predicates	413 (59.2%)
total number of predicate types	535
label of induced argument	
proto-agent (A0)	423 (60.6%)
proto-patient (A1)	107 (15.3%)
other (A2–A5)	168 (24.1%)

Table 7.3.: Properties of automatically induced implicit arguments and antecedents

We asked one of our undergraduate students to examine a subset of 90 induced implicit arguments and report possible errors. In total, 80 of the predicted discourse antecedents were found to be correct (89%). Some incorrectly linked instances still result from preprocessing errors. In particular, combinations of errors can lead to incorrectly identified instances as showcased in Example (23):

(23) "The Guatemalan Congress on Thursday ratified 126-12 [a Central America-US]$_{\text{A0}}$ [free trade]$_{\text{A1}}$ agreement, lawmakers said."

Induced missing argument and discourse antecedent: [goods]$_{\text{A2/co-agent}}$

Instead of recognizing "Central America" and "US" as two separate arguments (agent and co-agent) of the predicate AGREE, the semantic parser labels both entities as one argument (A0, agent); our system hence tries to determine a discourse antecedent for an argument that is predicted to be missing despite being actually realized (A2, co-agent). In the aligned predicate-argument structure, the co-agent is realized as a prepositional phrase: "[with the United States]$_{\text{A2}}$". The cross-document coreference tool incorrectly predicts "the United States" to be coreferent with "U.S. goods and services"; our system hence detects "goods" as the antecedent for the erroneously predicted implicit argument. Further error sources are incorrectly extracted document pairs and alignments between predicate-argument structures that do not correspond to each other. Text pairs (24) and (25) show excerpts of one respective example each.

(24) "[Production]$_{\text{A1}}$ rose [3.9 percent from October ...]$_{\text{A2}}$"

"[... manufacturing output]$_{\text{A1}}$ rose [2.6 percent]$_{\text{A2}}$ [in November]$_{\text{TMP}}$ ..."

(25) "[the president's]$_{\text{A0}}$ trip [to Indonesia]$_{\text{A1}}$"

"a [weeklong]$_{\text{TMP}}$ trip [to both countries ...]$_{\text{A1}}$"

7.4. Summary

In this chapter, we introduced a computational implementation of Step 3 of our induction method, which heuristically identifies implicit arguments and their discourse antecedents. Our approach depends on automatic annotations from semantic role labeling, predicate-argument structure alignment and coreference resolution. We implement two particular types of measures to minimize the impact of pre-processing errors: (1) we avoid imprecise input by applying high-precision tools instead of methods that are tuned for balanced precision and recall; (2) we circumvent some common pre-processing errors by formulating three constraints on resulting instances of implicit arguments and discourse antecedents. While our intrinsic analysis revealed that we cannot eliminate all error sources this way, we found the induced data set to be of high precision. This analysis, however, does not provide us with exact measurements on the actual utility of the induced data. To assess this value quantitatively, we perform several extrinsic evaluations, which we present in detail in the next chapter.

As discussed in Chapter 1 of this thesis, we assume that a proper treatment of implicit arguments can improve the performance of semantic parsers and text generation systems. In Chapter 8, we show how our data set of implicit arguments can be utilized in both of these areas. In particular, we demonstrate that the training process for models of implicit argument linking can be enhanced by including our data set of implicit arguments as additional training data. Furthermore, we show that implicit arguments can affect the perceived coherence of a text and that the instances of aligned (explicit and implicit) arguments in our data set can be employed to successfully predict this impact. We discuss potential benefits for other NLP tasks in Chapter 9.

Part III.

Applications and further directions

8. Applications

***Q2**: "How can we employ induced implicit arguments to improve existing SRL models?"*

***Q3**: "How can we predict coherent realizations of arguments in discourse context?"*

In the previous chapter, we introduced a novel data set that contains automatically induced instances of implicit arguments, discourse antecedents within the same document, and aligned explicit counterparts in comparable texts. In this chapter, we present several applications of this data set: firstly, we employ the automatically induced implicit arguments and discourse antecedents as additional training data for implicit argument linking; and secondly, we develop a model of local coherence that can be trained on instances of explicit and implicit arguments in discourse context. For the first task, we make use of a pre-existing model and evaluation scenario, in which we apply our novel data set as additional training material. For the latter task, we develop a model that emulates the decision process underlying (non-)realizations of arguments. We set up two evaluation scenarios to demonstrate the utility of this model.

Both applications are described in our *SEM 2013 paper "Automatically Identifying Implicit Arguments to Improve Argument Linking and Coherence Modeling" (Roth and Frank, 2013). This chapter is based on parts of this paper, which have been revised and extended. The first application in this chapter concerns the identification and linking of implicit arguments in discourse and is described in Section 8.1. In Section 8.2, we present a coherence model that can be applied to predict argument realization at a given point in discourse. In Section 8.3, we discuss an application of this coherence model on automatic summaries. Finally, we summarize our results in Section 8.4.

8.1. Linking Implicit Arguments in Discourse

Our first experiment assesses the utility of automatically induced pairs of implicit arguments and antecedents for the task of implicit argument linking. For evaluation, we use the data sets from the SemEval 2010 task on "Linking Events and their Participants in Discourse" (Ruppenhofer et al., 2010b, henceforth just *SemEval task*). For direct comparison with previous results and heuristic acquisition techniques, we apply the implicit argument identification and linking model by Silberer and Frank (2012, henceforth S&F) for training and testing. We briefly describe the SemEval task data and the model by Silberer and Frank in the next sections. More details can be found in Chapter 2.

8.1.1. Task Summary

Both the training and test sets of the SemEval task are text corpora extracted from Sherlock Holmes novels, with manual frame semantic annotations including implicit arguments (cf. Section 2.1.3). In the actual linking task ("NI-only"), gold labels are provided for local arguments and participating systems have to perform the following three sub-tasks: (1) identify implicit arguments (IA), (2) classify IAs as definite (DNI) or indefinite null instantiations (INI) and, if possible, (3) find an appropriate antecedent.

The task organizers provide two versions of their data sets: one based on FrameNet annotations and one based on PropBank/NomBank annotations. We found, however, that the latter only contains a subset of the implicit argument annotations from the FrameNet-based version. As all previous results in this task have been reported on the FrameNet data set, we adopt the same setting. Note that our automatically induced data set, which we want to apply as additional training data, is automatically labeled with a PropBank/NomBank-style parser. That is, we need to map our annotations to FrameNet in order to make use of them in this task. The organizers of the SemEval task provide a manual mapping dictionary for predicates in the annotated data set. We make use of this manual mapping and additionally use SemLink 1.1[1] for mapping predicates and arguments not covered by the dictionary.

8.1.2. Model Details

We make use of the system by S&F to train a new model for the NI-only task. As mentioned in the previous subsection, this task consists of three steps: In step (1), implicit arguments are identified as unfilled and non-redundant FrameNet core roles; in step (2), an SVM classifier is used to predict whether implicit arguments are definite based on a small number of features – semantic type of the affected Frame Element, the relative frequency of its realization type in the SemEval training corpus, and a boolean feature that indicates whether the affected sentence is in passive voice and does not contain a (deep) subject. In step (3), we apply the same features and classifier as S&F to find appropriate antecedents for (predicted) definite arguments (for details, cf. Chapter 2). S&F report that their best results were obtained when considering all entities as candidate antecedents that are syntactic constituents from the present and the past two sentences, or entities that occurred at least five times in the previous discourse ("Chains+Win" setting). In their evaluation, the latter of these two restrictions crucially depended on gold coreference chains. As the automatic coreference chains in our data are rather sparse (and noisy), we only consider syntactic constituents from the present and the past two sentences as antecedents ("SentWin" setting).

Before training and testing a new model with our own data, we perform feature selection using 10-fold cross validation. To find the best set of features, we run the feature selection on a combination of the SemEval training data and our additional data set. The only features that were selected in this process concern the "prominence" of the candidate antecedent, its semantic agreement with the selectional preferences of the

[1] `http://verbs.colorado.edu/semlink/`

	Precision	Recall	F_1-score
Chen et al. (2010)[2]	**0.25**	0.01	0.02
Tonelli and Delmonte (2011)	0.13	0.06	0.08
Laparra and Rigau (2012)	0.15	**0.25**	**0.19**
Laparra and Rigau (2013b)	0.14	0.18	0.16
Gorinski et al. (2013)[3]	0.14	0.12	0.13
S&F (no additional data)	0.06	0.09	0.07
S&F (best additional data)	0.09	0.11	0.10
This thesis	0.21	0.08	0.12

Table 8.1.: Results for identifying and linking implicit arguments in the SemEval test set.

predicate, the part-of-speech-tags used in each reference to the candidate entity and the semantic types of all roles that the entity fills according to local role annotations. These features are a subset of the best features found by Silberer and Frank (cf. Section 2.1.4).

8.1.3. Results

For direct comparison in the full task, both with S&F's model and other previously published results, we adopt the precision, recall and F_1 measures as defined in Ruppenhofer et al. (2010b).

We compare our results with those previously reported on the SemEval task (see Table 8.1 for a summary): the best performing system in the actual task in 2010 was developed by Chen et al. (2010) and is an adaptation of the semantic role labeling system SEMAFOR (Das et al., 2010). In 2011, Tonelli and Delmonte presented a revised version of their SemEval system (Tonelli and Delmonte, 2010), which outperforms SEMAFOR in terms of recall (6%) and F_1-score (8%). The best results in terms of recall and F_1-score up to date have been reported by Laparra and Rigau (2012), with 25% and 19%, respectively. Our model outperforms their state-of-the-art system in terms of precision (21%) but achieves a lower recall (8%). Two influencing factors for their high recall are probably (1) their improved method for identifying (resolvable) implicit arguments, and (2) their addition of lexicalized and ontological features.

Comparison to the original results reported by S&F, whose system we use, shows that our additional data improves precision (from 6% to 21%) and F_1-score (from 7% to 12%). The loss of 1 percentage point in recall is marginal given the size of the test set (only 259 implicit arguments have an annotated antecedent). Our result in precision is the second highest score reported on this task. Interestingly, the improvements are higher than those achieved in the original study by Silberer and Frank (2012), even though their best additional training set is three times bigger than ours and contains

[2]Results as reported by Tonelli and Delmonte (2011)

[3]Results computed as an average over the scores given for both test files; rounded towards the number given for the test file that contained more instances.

manual semantic annotations. We conjecture that their low gain in precision could be a side effect triggered by two factors: as discussed in the motivation of our work, the heuristically created training instances by Silberer and Frank might not represent implicit argument instances adequately (cf. Section 2.1.5); on the other hand, their model relies on coreference chains, which are automatically generated for the test set and hence are rather noisy. In contrast, our heuristically created data does not contain manual annotations on semantic roles and coreference chains, hence we do not make use of coreference information during training and testing. Despite this limitation, the results show that our new model outperforms previous models trained using the same system, indicating the utility and high reliability of our automatically induced data.

8.2. Modeling Local Coherence

In our second experiment, we examine the effect of implicit arguments on local coherence. That is, how does local argument (non-)realization affect the perceived coherence of a discourse segment? We approach this question as follows: first, we assemble a data set of document pairs that differ only with respect to a single realization decision (Section 8.2.1); given each pair in this data set, we ask human annotators to indicate their preference for the implicit or explicit argument instance in the pre-specified context (Section 8.2.2); finally, we attempt to emulate the decision process computationally using a discriminative model based on discourse and entity-specific features (Section 8.2.3). To assess the performance of the new model, we train it on automatically induced training data (Section 8.2.4), evaluate it against human annotations (Section 8.2.5) and compare its results to those of previous models of local coherence (Section 8.2.6).

8.2.1. Data Compilation

We use the data set of automatically induced implicit arguments (henceforth *source data*), described in Chapter 7, as a starting point for composing a set of document pairs that involve implicit and explicit arguments. To make sure that each document pair in this data set only differs with respect to a single realization decision, we first create two copies of each document from the source data: one copy remains in its original form, and the other copy will be modified with respect to a single argument realization. Example (26) illustrates an original and modified (marked by a question mark) sentence:

(26) [The Dalai Lama's]$_{\text{A0}}$ **visit** [to France]$_{\text{A1}}$ ends on Tuesday.

? [The Dalai Lama's]$_{\text{A0}}$ **visit** ends on Tuesday.

Note that adding and removing arguments at random can lead to structures that are semantically implausible. Hence, we restrict this procedure to predicate-argument structures (PAS) that actually occur and are aligned across two texts. Given a pair of PAS that differ with respect to an argument realization, we create modifications by replacing the specific implicit or explicit argument in one text with the corresponding

argument in the comparable text. Examples (26) and (27) show two such comparable sentences. The original PAS in Example (26) contains an explicit argument that is implicit in the aligned PAS and hence removed in the modified version. Vice versa, the original text in (27) involves an implicit argument, which is made explicit in the modified version.

(27) [The Dalai Lama's]$_{\text{A0}}$ **visit** coincides with the Beijing Olympics.

? [The Dalai Lama's]$_{\text{A0}}$ **visit** [to France]$_{\text{A1}}$ coincides with the Beijing Olympics.

We ensure that the modified structure fits into the given context grammatically by only considering pairs of PAS with identical predicate form and constituent order. We found that this restriction constraints affected arguments to be modifiers, prepositional phrases and direct objects. We argue that this is actually a desirable property because more complicated alternations could affect coherence by themselves. In other words, resulting interplays would make it difficult to distinguish between the isolated effect of argument realization itself and other effects, triggered for example by sentence order (Gordon et al., 1993).

8.2.2. Annotation

We set up a web experiment using the evaluation toolkit by Belz and Kow (2011) to collect ratings of local coherence for implicit and explicit arguments. For this experiment, we compiled a data set of 150 document pairs. As described in Section 8.2.1, each text pair consists of mostly the same text, with the only difference being one argument realization.

We presented all 150 pairs to two annotators[4] and asked them to indicate their preference for one alternative over the other using a continuous slider scale. The annotators got to see the full texts, with the alternatives presented next to each other. To make texts easier to read and differences easier to spot, we collapsed all identical sentences into one column and highlighted the aligned predicate (in both texts) and the affected argument (in the explicit case). An example is shown in Figure 8.1. To avoid any bias in the annotation process, we shuffled the sequence of text pairs and randomly assigned the side of display (left/right) of each realization type (explicit/implicit). Instead of providing a definition of local coherence ourselves, we asked annotators to rate how "natural" a realization reads given the discourse context. This procedure is in line with previous work by Pitler and Nenkova (2008) who "view text readability and text coherence as equivalent properties".[5]

We found that annotators made use of the full rating scale, which spans from -50 to $+50$, with the extremes indicating either a strong preference for the text on the left hand side or the right hand side, respectively. However, most ratings were concentrated more towards the center of the scale (i.e., around zero). This seems to imply that the

[4] Both annotators are undergraduate students in Computational Linguistics.

[5] Pitler and Nenkova explicitly restrict this statement to "competent language users". While our annotators are non-native speakers of English, both of them are highly proficient in the English language.

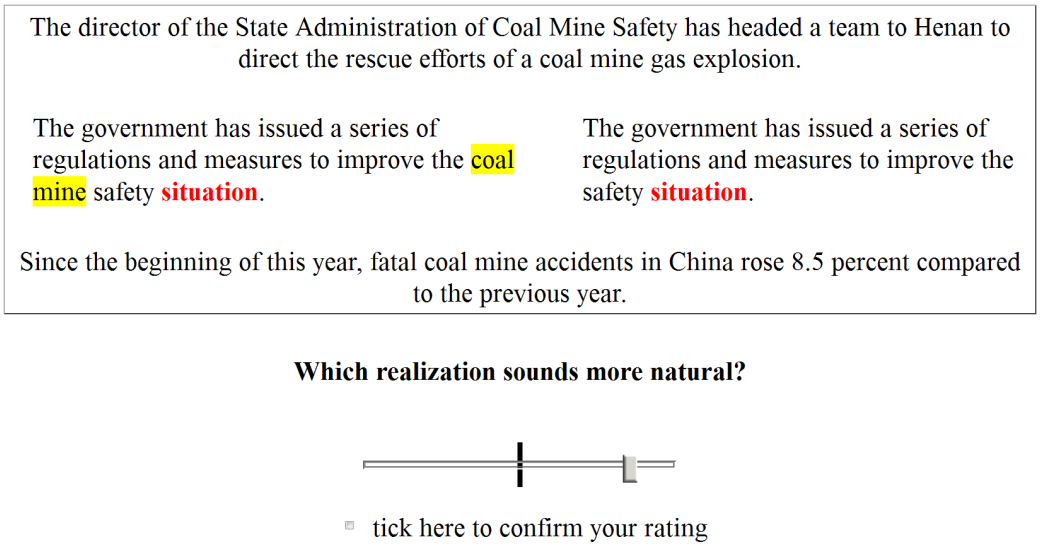

Figure 8.1.: Texts as displayed to the annotators.

use of implicit or explicit arguments did not make a considerable difference most of the time. We confirmed this assumption and resolved disagreements between annotators in several group discussions. The annotators also affirmed that some cases do not read naturally when a specific argument is omitted or redundantly realized at a given position in discourse. For example, the text fragment in Example (28) shows two sentences, in which an argument has been realized twice, leading to a perceived redundancy in the second sentence (A4, `destination`); vice versa, Example (29) showcases an excerpt, in which a non-redundant argument (A2, `co-signer`) has been omitted.

(28) ? The remaining contraband was picked up at Le Havre. The containers had arrived [in Le Havre] from China.

(29) ? Lt.-Gen. Mohamed Lamari (...) denied his country wanted South African weapons to fight Muslim rebels fighting the government. "We are not going to fight a flea with a hammer," Lamari told reporters after signing the agreement of intent [∅].

Following discussions with the annotators, we discarded all items from the final data set for which no clear preference could be established (72%) or the annotators had different preferences (9%). We mapped all remaining items into two classes according to whether the affected argument had to be implicit (9 texts) or explicit (20 texts). All 29 uniquely classified texts are used as a small gold standard test set for evaluation.

8.2.3. Coherence Model

We model the decision process that underlies the (non-)realization of arguments using a SVM classifier (LIBSVM, Chang and Lin, 2011) and a range of discourse features. We

define all features based on the following three factors: the affected predicate-argument structure (**Parg**), the (automatic) coreference chain of the affected entity (**Coref**), and the discourse context (**Disc**).

Parg. The first group of features is concerned with the complexity of the affected predicate-argument structure: this includes the absolute and relative number of explicitly realized arguments in the structure, the number of modifiers in it, and the total length of the structure as well as of the complete sentence (in words).

Coref. The coreference-specific features include transition patterns as inspired by the entity grid model (cf. Section 2.2.2), the absolute number of previous and follow-up mentions, their POS tags, and the distance between the current PAS to the closest previous and follow-up mention (in number of words and sentences). In contrast to previous work on the entity grid model, we do not type transition features with respect to the grammatical function of explicit realizations. The reason for skipping this information lies in the insignificant amount of relevant samples in our training data (cf. Section 8.2.4).

Disc. On the discourse level, we define a small set of additional features that include the total number of coreference chains in the text, the occurrence of pronouns in the current sentence, lexical repetitions in the previous and follow-up sentence, the current position in discourse (begin, middle, end), and a feature indicating whether the affected argument occurred in the first sentence.

Most of these features overlap with those successfully applied in previous work. For example, the transition patterns are inspired by the entity grid model. In addition to entity-grid like features, Pitler and Nenkova (2008) also use text length, word overlap and pronoun occurrences as features for predicting readability. Our own contribution lies in the definition of PAS-specific features and the adaptation of all features to the task of predicting (non-)realization of arguments in a predicate-argument structure. In the evaluation (cf. Section 8.2.6), we report results for two models: a **simplified model**, which only makes use of entity-grid like features, and the **full model**, which uses all features described above. To learn feature weights, we make use of the training data described in the following section.

8.2.4. Training data

Our model does not make use of any manually annotated data for training. Instead, we rely solely on the automatically induced source data, described in Section 8.2.1, for learning. This source data consists of texts, in which implicit and explicit arguments were automatically detected in aligned predicate-argument structures.

For training, we prepare this data set as follows: firstly, we remove all data points that also occur in the test set; secondly, we split all pairs of texts into two groups – texts that contain a predicate-argument structure, in which an implicit argument has been identified (IA), and their comparable counterparts, which contain the aligned PAS with an explicit argument (EA). All texts are labeled according to their group. For all texts in

group EA, we remove the explicit argument from the aligned PAS. This way, the feature extractor always gets to see the text and automatic annotations as if the realization decision had not been performed and can thus extract unbiased feature values for the affected entity and argument position. Given each feature representation, we train a classifier using the default parameters of the LIBSVM package.[6]

8.2.5. Evaluation Setting

The goal of this task is to correctly predict the realization type (implicit or explicit) of an argument that maximizes the perceived coherence of the document. As a proxy for coherence, we use the naturalness ratings given by our annotators. We evaluate classification performance on the 29 data points in our annotated test set, for which clear preferences have been established. We report results in terms of precision, recall and F_1-score. We compute precision as the fraction of correct classifier decisions divided by the total number of classifications; and recall as the fraction of correct classifier decisions divided by the total number of test items. Note that precision and recall are identical when the model provides a class label for every test item. We compute F_1 as the harmonic mean between precision and recall. To get a better insight into strengths and deficits of different models, we further compute the ratio of correct classifications separately for arguments that are annotated as implicit ($P_{implicit}$) and explicit ($P_{explicit}$).

For comparison, we apply a couple of coherence models proposed in previous work: the original entity grid model by Barzilay and Lapata (2005), a modified version that uses topic models (Elsner and Charniak, 2011a) and an extended version that includes entity-specific features (Elsner and Charniak, 2011b); we further apply the discourse-new model by Elsner and Charniak (2008) and the pronoun-based model by Charniak and Elsner (2009). A more detailed description of these models can be found in Chapter 2. For all of the aforementioned models, we use their respective implementation provided with the Brown Coherence Toolkit. [7] Note that the toolkit only returns one coherence score for each document. To use it for argument classification, we use two documents per data point – one that contains the affected argument explicitly and one that does not (implicit argument) – and treat the higher scoring variant as classification output. If both documents achieve the same score, we neither count the test item as correctly nor as incorrectly classified. We further apply our own model on each data point in the small annotated test set, where we always treat the affected argument, regardless of its actual annotation, as implicit to extract unbiased feature values for classification. Based on the features described in Section 8.2.3, our model predicts the realization type of each argument in the given context. We note that our model has an advantage here because it is specifically designed for this task. Yet, all models compute local coherence ratings based on entity occurrences and should thus be able to predict which realization type coheres best with the given discourse context. That is, since the input document pairs are identical except for the affected argument position, the coherence scores assigned by each model to pairs of text only differ with respect to the affected entity realization.

[6] The default settings in LIBSVM are: equal costs of both classes and use of a sigmoid kernel.
[7] `http://www.ling.ohio-state.edu/%7Emelsner/`

	$P_{implicit}$	$P_{explicit}$	$P_{overall}$	$R_{overall}$	F_1-score
Entity grid models	–	–	–	–	–
Baseline entity grid	0.50	0.05	0.15**	0.14**	0.15**
Extended entity grid	0.56	0.00	0.19**	0.17**	0.18**
Topical entity grid	0.86	0.20	0.34**	0.34**	0.34**
Other models	–	–	–	–	–
Pronouns	0.60	0.37	0.43**	0.34**	0.38**
Discourse-newness	**1.00**	0.25	0.48**	0.48**	0.48**
This thesis	–	–	–	–	–
Our (full) model	0.78	0.95	**0.90**	**0.90**	**0.90**
Simplified model	0.56	**1.00**	0.83	0.83	0.83
Majority class	0.00	1.00	0.69*	0.69*	0.69*

Table 8.2.: Results in Precision, Recall and F_1-score for correctly predicting argument realization; Significant differences in overall scores from our (full) model are marked with asterisks (* $p<0.1$; ** $p<0.01$)

8.2.6. Results

The results are summarized in Table 8.2. As all models provided class labels for almost all test instances, we focus our discussion on F_1-scores. The majority class in our test set is the explicit realization type, making up 20 of the 29 test items (69%).

The original entity grid model produced differing scores for the two realization types only in 26 cases. The model exhibits a strong preference for the implicit realization type: it predicts this class in 22 cases, resulting in only 5% of all explicit arguments being correctly classified. Overall, the entity grid achieves an F_1-score of 15%. Taking a closer look at the features of the model reveals that this is an expected outcome: in its original setting, the entity grid learns realization patterns in the form of sentence-to-sentence transitions. Most entities are, however, only mentioned a few times in a text, which means that non-realizations constitute the 'most frequent' class – independently of whether they are relevant in a given context or not. The models by Charniak and Elsner (2009) and Elsner and Charniak (2011a), which are not based on an entity grid, suffer less from such an effect and achieve better results, with F_1-scores of 38% and 48%, respectively. The topical refinement to the entity grid model also alleviates the bias towards non-realizations, resulting in improved F_1-scores of 34%. To counterbalance this issue altogether, we train a simplified version of our own model that only uses features that involve entity transition patterns. The main difference between this simplified model and the original entity grid model lies in the different use of training data: while entity grid models treat all non-realized items equally, our model gets to "see" actual examples of entities that are implicit. In other words, our simplified model takes into account implicit mentions of entities, not only explicit ones. The results confirm that this extra information has a significant impact ($p<0.01$, using a randomization test; Yeh, 2000) on test set performance, and raises the ratio of correctly classified explicit arguments to

100%. Yet, the simplified model only yields a precision of 56% on implicit arguments, leading to an overall F_1-score of 83%. As demonstrated by the performance of our full model, a combination of all features is needed to achieve the best overall results of 90% precision and recall. Applied to the two classes separately, our model achieves a precision of 78% on arguments that are annotated as implicit and 95% on explicit arguments.

Weight	Group	Feature description
+55.38	Coref	The entity is mentioned within two sentences
+25.37	Coref	The entity has previously been mentioned as a proper noun
+19.14	Coref	The entity has previously been mentioned as a pronoun
+14.75	Parg	The PAS consists of at least 2 words
+12.82	Parg	The sentence contains at least 20 words
+12.65	Parg	The sentence contains at least 40 words
+12.12	Parg	The PAS consists of at least 3 words
+11.32	Coref	The entity is mentioned in the next but not in the previous sentence
+11.23	Coref	The entity is mentioned within the previous or next 10 tokens
+10.79	Coref	The entity is mentioned within the previous two sentences
−4.72	Parg	The absolute number of arguments and modifiers in the PAS
−5.80	Coref	The entity is mentioned two sentences ago but not in the previous
−6.38	Parg	The previous entity mention was a definite noun phrase
−6.94	Disc	The PAS occurs in the first sentence of the discourse
−7.11	Parg	The absolute number of arguments in the affected PAS
−7.22	Coref	The entity is mentioned in the next sentence but not in the previous
−8.79	Coref	The entity is mentioned within the previous or next three sentences
−9.42	Coref	The entity is mentioned within the previous three sentences
−10.38	Coref	The entity is mentioned in the previous sentence
−32.70	Coref	The next mention is a pronoun

Table 8.3.: Weights assigned to each feature in our model; list includes the top 10 features for implicit (positive weight) and explicit arguments (negative).

To determine the impact of the three different feature groups, we derive the weight of each feature from the model learned by LIBSVM. Table 8.3 gives an overview over the ten highest weights for implicit and explicit realization classification. We use the following terminology in the feature description: "the entity" refers to the entity that is referred to by the to-be-classified argument, "next/previous mention" denotes a co-referring mention to the same entity, "the PAS" refers to the predicate-argument structure, which contains the affected argument (implicitly), and "the sentence" refers to the sentence, in which the PAS is realized. All "distances" refer to the number of tokens that appear between the predicate that heads the PAS and the previous or next mention of the entity. As shown in Table 8.3, the strongest feature for classifying an argument as implicit is whether the entity is also realized in the preceding or following two sentences. The strongest feature for classifying an argument as explicit is whether the next mention is a

pronoun. With one exception, all of the listed features are Boolean. That is, they only take values of −1 (false) and +1 (true). In contrast, the number of realized arguments and modifiers in a PAS varies from zero to six in our automatically annotated training data. The SVM normalizes these raw numbers to a scale from −1 (0 constituents) to +1 (6 constituents), so that all feature values are within the same range. To classify a particular argument instance, all features are taken into account. For example, if an entity has been mentioned once before as a proper noun but never as a pronoun, the classifier will take the feature "mentioned before as a proper noun" into account with a value of +1; the fact that the entity has *not* been mentioned as a pronoun will be taken into account with a value of −1.

In the next section, we demonstrate how our novel coherence model can be applied in the task of post-processing automatically generated summaries. We discuss other application scenarios and potential improvements in Chapter 9.

8.3. Multi-Document Summarization

In our third experiment, we apply the coherence model that we introduced in the previous section to the task of multi-document summarization (MDS). The task of MDS is particularly appealing for applying our model because it involves the combination of information from comparable, hence overlapping, textual sources. We expect this to frequently result in automatic summaries that contain multiple references to the same entity, when one reference might be sufficient. We aim to reduce this redundancy on the level of shallow semantic analysis by predicting whether a semantic argument should be explicit or implicit at a specific point in discourse. Given an automatically generated summary, we hypothesize that our model can be used this way to correctly predict entity references that can be omitted and references that have to remain explicit to establish local coherence. By removing arguments that are predicted to be implicit, we expect summaries to become more coherent and shorter in text. Vice versa, we expect that summaries remain informative and coherent by keeping arguments in place that are classified as explicit.

As a starting point for this experiment, we use the output of the two best-performing systems on a multi-document summarization data set. We describe this data set and the two systems in more detail in Section 8.3.1. In Section 8.3.2, we outline the modification procedure, which uses predictions from our coherence model to revise automatic summaries. In Section 8.3.3, we describe a comparative evaluation of original and modified automatic summaries. Finally, we discuss the results of this evaluation in Section 8.3.4.

8.3.1. Summarization data

From 2002 to 2007, the Document Understanding Conference (DUC) organized an annual task on generating automatic summaries (Hahn and Harman, 2002). From 2008 to 2011, this annual challenge was part of the Text Analysis Conference (TAC). In this experiment, we focus on the data set from the summarization task that was part of TAC 2008 (Dang and Owczarzak, 2008). We deliberately choose this data set because it

The A380 superjumbo, which will be presented to the world in a lavish ceremony in southern France on Tuesday, will be profitable from 2008, its maker Airbus told the French financial newspaper La Tribune.
One problem that Airbus is encountering with its new A380 is that the craft pushes the envelope on the maximum size of a commercial airplane.
French President Jacques Chirac immediately hailed the total success of the first test flight of the Airbus A380.
The superjumbo Airbus A380, the world's largest commercial airliner, took off into cloudy skies over southwestern France for its second test flight.

Figure 8.2.: Output from the summarization system by Berg-Kirkpatrick et al. (2011).

addresses summarization from multiple documents and it is one of the most recent data sets, on which new systems are being developed. While the original TAC task consists of two sub-tasks – producing an initial summary and an "update summary" – we here only consider data from the first task. Given a set of comparable texts as input, the goal of this task is to generate a "well-organized, fluent summary" of at most 100 words. The task data consists of 48 document sets, each comprising 10 newswire articles from a subset of the Gigaword corpus. In total, 33 teams participated in this task, resulting in 71 submissions (up to 3 per team). In addition to system summaries, the organizers asked 8 human peers to write reference summaries for each document set.

During the evaluation phase of both human and system summaries, the organizers found a high correlation between automatic scoring metrics and human judgments. According to the ROUGE metric (Lin, 2004), which measures the quality of a summary with respect to its overlap in content with a reference summary, the best performing systems on the TAC data set nowadays are those by Berg-Kirkpatrick et al. (2011) and Woodsend and Lapata (2012). We describe both systems in the following paragraphs. Example summaries of their systems are shown in Figures 8.2 and 8.3, respectively. Both examples exhibit a problem that we frequently find in automatic summaries: news articles typically focus on a specific topic or entity, which frequently leads to multiple (redundant) references to the same topic or entity. When reading the automatic summaries in Figures 8.2 and 8.3, it is not clear how the sentences within a summary relate to each another. One reason for this is that it is hard to distinguish between already mentioned and actual new information in each sentence. We argue that removing redundant mentions could shorten each sentence to contain only essential information and thus make the summaries as a whole read more coherently.

Berg-Kirkpatrick et al. The system by Berg-Kirkpatrick et al. (2011) generates automatic summaries by extracting and compressing sentences that occur in the documents that are to be summarized. To select important text fragments, sentences are represented as bi-grams, which are weighted by features that include the frequency of (stemmed) words, overlap with stop words, and bi-gram position in the document. Sentence compression is modeled by learning feature weights for deleting specific subtrees

The United States said Tuesday it was "deeply troubled" by an apparent setback to democracy in Nepal, where King Gyanendra has dismissed the government and imposed a state of emergency.
It was the second time in three years that the king has taken control of the tiny South Asian constitutional monarchy.
Japan, the top donor to aid-dependent Nepal, on Thursday voiced concern at King Gyanendra's power grab and urged him to restore democracy and release any detained political leaders.
King Gyanendra on Tuesday sacked prime minister Sher Bahadur Deuba for the second time in three years and assumed control.

Figure 8.3.: Output from the summarization system by Woodsend and Lapata (2012).

of a sentence, including adjuncts, prepositional phrases and relative clauses. Finally, summaries are generated by extracting and compressing sentences from newswire documents, such that a joint objective for all features is maximized. The system output is a simple concatenation of the selected and compressed sentences. No post-processing steps are applied to establish local or global coherence. An example is shown in Figure 8.2.

Woodsend and Lapata. Like Berg-Kirkpatrick et al., Woodsend and Lapata (2012) model generation as an extraction and compression task that is solved by maximizing a joint objective function. In contrast to the simple compression model used by Berg-Kirkpatrick et al., however, Woodsend and Lapata learn a Quasi-synchronous Tree Substitution Grammar (Eisner, 2003) to combine and rewrite extracted sentence fragments. For content selection, Woodsend and Lapata rate each phrase in the original documents using part-of-speech features (e.g., "phrase contains a noun") and the bi-gram features by Berg-Kirkpatrick et al. The extraction step in the system also takes into account lexical and stylistic features to ensure that certain words and phrases, which are learned to be inappropriate in summaries, are unlikely to be extracted. This includes, for example, personal pronouns, questions and quotations, and lexemes such as "say", "go", "last" and "tell". The system output consists of extracted sentence fragments that are rewritten to form grammatical sentences via the tree substitution grammar. No post-processing steps are applied to establish local or global coherence. An example summary is given in Figure 8.3.

8.3.2. Applying the Coherence Model

As exemplified in the previous section, extractive summarization systems do not always generate coherent output. One obvious reason for this lack of coherence is that sentences and sentence fragments are extracted independent of their original discourse context. When rewriting extracted fragments, the generated sentences are simply concatenated without considering the change in context. As a result, adjacent sentences might seem unrelated, or they might contain redundant repetitions of the same information. In this

experiment, we address apparent incoherence caused by the latter issue by removing redundant argument realizations. We achieve this goal by predicting the realization type of each argument using the coherence model that we introduced in Section 8.2.3. This application involves three processing steps, which we outline below.

First, we apply the same pre-processing methods that we used on the training data of our coherence model (cf. Section 8.2.1). That is, we run the pre-processing pipeline that comes with MATE tools (Björkelund et al., 2010; Bohnet, 2010) to identify predicate-argument structures (PAS) and we use the coreference toolkit by Martschat et al. (2012) to identify co-referring arguments in the identified PAS. Since the output of both summarization systems consist of text fragments extracted from newswire articles, no domain adaptation is required for processing them.

Given the predicate-argument structures predicted by the semantic parser, we process each text with our coherence model to identify arguments that should remain explicit and arguments that should be implicit. To predict the realization type of each argument, the coherence model takes into account properties of the affected predicate-argument structures, entity mentions that co-refer with the affected argument and discourse-specific features such as the position of the affected sentence in discourse. For a complete list of features, see Section 8.2.3.

Finally, we put the predictions by our coherence model into practice: arguments classified as 'explicit' remain realized in text, and arguments classified as 'implicit' will be removed. As in the previous experiment, we restrict the modification procedure to semantic arguments that can be removed without affecting constituent order. We implement this restriction by only removing semantic arguments that the syntactic-semantic parser recognized as optional dependents (modifiers, adverbials, appositions). If the removed argument formed the beginning of a noun phrase, we replace it by the definite determiner 'the' to retain grammaticality. An example is illustrated in sentence (30):

(30) ~~Washington's~~ secretary of state, the ~~state's~~ top election official, must certify the vote next week.

→ The secretary of state, the top election official, must certify the vote next week.

8.3.3. Evaluation Setting

We applied our model on all summaries produced by the systems from Woodsend and Lapata (henceforth *WL*), and Berg-Kirkpatrick et al. (henceforth *BK*). That is, for each syntactically optional argument, we used our model to predict whether the argument should be explicit or implicit. For evaluation, the model automatically creates modified versions of summaries, in which all arguments are removed that our model predicted to be implicit. Applied to the automatically generated summaries, our model predicted all arguments to be explicit in 83 out of 96 cases. In the remaining 13 summaries, our model finds at least one explicit argument that it predicts to be implicit. To evaluate the quality of summaries before and after modification, we set up two web experiments, Study 1 and Study 2, described below. Since the modifications only affect arguments classified as implicit by our model, we set up an additional experiment, Study 3, to

examine the performance of our model with respect to arguments that are classified as explicit.

Study 1. In the first experiment, we present original and modified versions of each summary next to each other, and ask annotators to indicate whether they prefer one alternative with respect to grammaticality, informativeness and coherence. Our hypotheses for Study 1 is that grammaticality and informativeness are not affected when removing arguments that are actually redundant, and that annotators would find the modified summaries to be more coherent. At the same time, we expect that informativeness and coherence are negatively affected when non-redundant references are being removed. For grammaticality and informativeness, the annotators can indicate whether they prefer one alternative (and if so, which one) or not. For coherence, the annotators also have the possibility to indicate the strength of their preference on a slider scale.

Study 2. Since annotators get to see the original and modified summary in Study 1, resulting judgments only represent relative preferences. In the second experiment, we only present one of the two versions to each annotator and ask for a single coherence rating. Since we are interested in ratings of local coherence, we always mark a specific sentence. Annotators do not get to see any alternative for the affected sentence and hence have to give absolute coherence scores. In this experiment, we use a fixed 5-point scale and provide the following definitions to the annotators to help them choose an appropriate rating:

5 – The sentence fits excellent into the text.
4 – The marked sentence reads naturally in the given context.
3 – Only parts of the sentence fit into the context, the rest sounds odd.
2 – The marked sentence sounds strange in the given context.
1 – The sentence in question is not related to the context at all.

As in the first experiment, our hypothesis is that modified summaries are more coherent. In contrast to Study 1, however, annotators are not biased by the fact that they get to see both options. By making use of the absolute scores from this experiment, we can further test whether original and modified summaries significantly differ in how coherent they are according to human judgments.

Study 3. In our third experiment, we test whether our model can reliably predict whether an argument realization contributes to local coherence. In contrast to the Study 1 and Study 2, we modify summaries for this experiment with respect to an argument that is predicted to be *explicit* according to our model. As in Study 1, we present pairs of summaries to our annotators and ask them to indicate their preference regarding grammaticality, informativeness and coherence. Each pair consists of an original summary from the two state-of-the-art systems introduced in Section 8.3.1 and a modified version, in which an argument is removed that should be explicit according to our model. We

create the same number of summary pairs as in Study 1 by randomly sampling and removing arguments that are syntactically optional and classified as explicit by our model. Our hypothesis for this experiment is that the removed arguments contribute to local coherence. Hence, we expect higher coherence ratings for the variant of each summary that contains the affected argument realization.

Setup All three studies are designed as web-based experiments, in which annotators rate original and modified summaries. For Study 1 and 3, we use the NLTG package (Belz and Kow, 2011) to set up online evaluation forms. In total, we collect 40 ratings in Study 1 and 42 ratings in Study 3. For Study 2, we set up an annotation task on CrowdFlower[8]. This way, we collect a higher number of ratings (842 in total) for each summary in isolation. To ensure a high quality of annotations, we required annotators to correctly spot "random" summaries, in which parts of the summary were replaced by sentences from an unrelated text, and to have a high level of proficiency in English. In Study 1 and Study 3, we enforced the second criterion by only admitting participants that use English as their primary medium of communication at work. In Study 2, we explicitly required annotators to be citizens of the UK or the US.

8.3.4. Results

Table 8.4 presents the mean ratings collected for each summary. For Study 1 and Study 2, each number represents the average preference for the summary proposed by our model. All numbers are scaled to range from -1 (highest possible disagreement) to $+1$ (highest possible agreement). For Study 2, the listed ratings are absolute differences between the average score (over all annotators) for the original and modified summary. For better comparability, we also re-scaled the differences in results of this experiment to range from -1 to $+1$. Since the differences found in Study 2 are generally low – and none of them have been found to be significant –, we concentrate our discussion on the results from Study 1 and Study 3.

In general, we observe a positive agreement between human ratings and the predictions made by our our model. According to the ratings, there is only one case, in which our model has a negative impact on grammaticality, and two cases, where informativeness is negatively affected. Regarding coherence, the ratings reveal a positive correlation ($\geq$0.0) with our model in 19 out of 26 cases (73%). As found already in the results of our first coherence experiment (cf. Section 8.2.6), we observe that our model performs better at predicting explicit arguments (Study 3) than predicting implicit arguments (Study 1). We discuss the results and potential errors by our model in more detail in the following paragraphs.

Grammaticality. In Study 1, the ratings show that, in the case of one summary, grammaticality was negatively affected by a modification proposed by our model. We note

[8] `http://www.crowdflower.com`

ID	system	grammaticality (relative)	informativeness (relative)	coherence (relative)	coherence (absolute)
		Study 1			Study 2
D0803	BK	–	–	+1.00	+0.04
D0806	BK	–	–	+0.20	−0.07
D0808	BK	–	–	−0.20	−0.05
D0816	BK	−0.7	–	−0.53	±0.0
D0844	BK	–	–	+0.60	−0.07
D0801	WL	–	–	−0.33	±0.0
D0803	WL	–	–	+0.20	−0.11
D0805	WL	–	–	+0.08	−0.04
D0808	WL	–	−0.25	+0.40	±0.0
D0812	WL	–	–	−0.10	+0.05
D0813	WL	–	–	−0.40	−0.04
D0822	WL	–	–	±0.0	+0.03
D0833	WL	–	−0.5	±0.0	−0.05
		Study 3			
D0801	BK	–	–	±0.0	
D0818	BK	–	–	+0.53	
D0820	BK	+0.33	–	+0.27	
D0842	BK	–	–	+0.33	
D0846	BK	+0.33	–	+0.47	
D0804	WL	+0.67	+0.33	+1.0	
D0807	WL	–	–	+0.46	
D0811	WL	–	–	−0.26	
D0817	WL	–	–	+0.26	
D0820	WL	–	+0.67	−0.20	
D0822	WL	–	+0.33	+0.93	
D0836	WL	–	–	±0.0	
D0842	WL	–	+1.0	+0.73	
agreement with our model		25/26	24/26	19/26	6/13

Table 8.4.: Rating differences regarding grammatically, informativeness and coherence after removing explicit arguments. Positive values indicate a positive correlation between our model prediction and annotations.

that the reason for this lies in an incorrect syntactic analysis produced by the preprocessing pipeline, which led to the removal of a semantic argument in a subject position. The affected original and modified sentence (marked by a question mark) from summary D0816 (BK) is shown in Example (31):

(31) Jan 10 Pyongyang withdraws from the nuclear Non-Proliferation Treaty NPT.

? Jan 10 withdraws from the nuclear Non-Proliferation Treaty NPT.

Informativeness. As a result of Study 1, we observe that annotators agree that most modifications, performed according to our model, do not affect the informativeness of a summary. Only in two cases, some but not all annotators marked summaries, modified according to model predictions, as containing less information than the original version. One reason as to why not all annotators marked the specific cases as less informative might be that the affected entities are also realized elsewhere within the same sentence. We list the sentences, which are taken from summary D0808 (WL) and summary D0833 (WL), with the removed arguments striked out, in Example (32) and (33), respectively.

(32) More than 1,500 members of an Iraqi Christian group have gone ~~to Northern Iraq~~ to try to protect Christians following attacks on churches in Baghdad and Mosul.

(33) The Indian Space Research Organization (ISRO) has short-listed experiments from five nations including the United States, Britain and Germany, for a slot on ~~India's~~ unmanned moon mission Chandrayaan-1 to be undertaken by 2006-2007, the Press Trust of India (PTI) reported Monday.

For comparison, a negative impact on informativeness was observed in four cases in Study 3, where modifications are against the classification by our model. In one particular case, all annotators agreed that the original version – as predicted by our model – contained more information. We show the affected sentence from summary D0842 (WL) in Example (34):

(34) Vatican-pope-Poland WARSAW – As the world prepared for the funeral of John Paul II tens of thousands of ~~the pope's~~ Polish compatriots head to Rome to bid him a dieu.

Coherence. In Study 1 and Study 3, we find that annotators found the predictions of our model to be preferable in 15 out 26 cases. When additionally taking into account cases, in which no preference was found (± 0.0), the agreement between ratings and model predictions increases to 19/26 cases (73%). In Study 1, annotators rated the coherence of modified summaries higher in six cases. At the same time, the original summary was rated more coherent in five other cases. We observe high gains in coherence ratings when removing arguments that are indeed redundant. Two examples are given in (35) and (36), which show the summaries D0803 and D0806 from system BK:

(35) a. The country's work safety authority will release the list of the first batch of coal mines to be closed down said Wang Xianzheng deputy director of the National Bureau of Production Safety Supervision and Administration. China is seeking solutions from the world to improve its coal mining safety system.

b. The death toll in China's disaster-plagued coal mine industry is rising according to the latest statistics released by the government. Fatal coal mine accidents ~~in China~~ rose 8.5 percent in the first eight months of this year with thousands dying despite stepped-up efforts to make the industry safer.

(36) a. President George W. Bush on Monday nominated White House counsel Harriet Miers to replace retiring Justice Sandra Day O'Connor on the Supreme Court reaching into his loyal inner circle for a pick that could reshape the nation's judiciary for years to come.

b. Doubts about Miers – thanks to her friends in the White House, Harriet Miers will face the Senate Judiciary Committee bearing a greater burden than John Roberts now the chief justice of the Supreme Court.

c. The chaos surrounding the Supreme Court nomination ~~of Harriet Miers~~ gets curiouser and curiouser.

Not all arguments removed by our model are redundant though. For example, the lowest rating for a modified summary correlates with Example (32), in which the subject was removed from a sentence. Syntactic optionality, however, does not guarantee redundancy either. For example, removing a modifier can also cause incoherence as illustrated by summary `D0808` (BK) in Example (37):

(37) a. Christians make up just 3 percent of Iraq's population of about 25 million. Two churches in the northern Iraqi city of Mosul came under simultaneous attack on Tuesday witnesses and clerics said but there were no immediate reports of casualties.

b. Most of Christians in Iraq are in Baghdad and northern cities.

c. ~~Iraq's~~ Christian parties complain they lack funding.

One problem in Example (37) might be that coherence is affected by an interplay with other factors. More specifically, it seems unclear how "the Christians in ... Baghdad and northern cities" are related to "(Iraq's) Christian parties" in the first place – is it a reference to the same set of entities? Or is one a subset of the other? By removing the modifier "Iraq's", the question becomes even more difficult to answer. In contrast, if the previous sentence had already talked about religious parties, the reference might have been clear and the modifier "Iraq's" could have been removed. In general, we find that other factors related to coherence have to be considered in order to improve the performance of our model on this task. This insight is different from what we found in our

previous experiment on newswire reports, where input texts were presumably coherent except for a single argument realization (cf. Section 8.2). As a consequence, we were able to assume in the previous experiment that complementary factors, such as smooth transitions of entity mentions, already contributed to establishing local coherence. This assumption seems inappropriate in a setting that involves automatically generated summaries. Another example to support this claim is given in snippet (38) from summary `D0805` (WL):

(38) Medicare is going to pay the insurance companies $700 a year for everyone who signs up for a drug plan.

(orig.) Millions of people are asking that question as they consider Medicare's new prescription drug plan, which rolls out Jan. 1.

(mod.) Millions of people are asking that question as they consider the new prescription drug plan, which rolls out Jan. 1.

In this example, the (original) second sentence refers to "that question" and "Medicare's new prescription drug plan". While the previous sentence indeed refers to "Medicare" and "a drug plan", no question or 'new' drug plan has been mentioned in the preceding context. While our model is not able to recognize this difficulty, the particular case illustrates a good example, in which a combination of different coherence models could be useful. The effect of mentions to discourse-new and discourse-old entities has, in fact, been addressed before. We discussed one such model (Elsner et al., 2007), together with other attempts to modeling local coherence, in Chapter 2. As shown in the previous experiment, however, none of the previously proposed models is able to correctly capture the impact of implicit vs. explicit arguments on perceived coherence (cf. Section 8.2.6). Combing models is a non-trivial task though as the different factors interfere with one another. One way of considering various factors at the same time would be to process text in sequential order. As decisions at the beginning of a text can have an impact, however, on what the content of the preceding text is expected to be, coherence-related realization decisions should ideally be addressed already during summarization. To the best of our knowledge, the first and only current model, which performs coherence modeling and sentence extraction for resulting summaries jointly, has recently been proposed by Christensen et al. (2013).

From both coherence experiments together, we conclude that our model provides a suitable prerequisite to predict argument realizations that maximize the perceived coherence of a text, but developing a combination with other models will be necessary to successfully apply these predictions in unrestricted NLG tasks.

8.4. Summary

In this chapter, we presented several applications of the automatically induced data set of implicit arguments that we introduced in Chapter 7. This data set has been induced

from pairs of comparable text and is a unique resource in that it contains automatic annotations of implicit arguments, aligned explicit arguments and discourse antecedents. In the first application, "Linking events and their participants in discourse", we evaluated the utility of our data using a pre-existing model and evaluation data set. For the second and third application, we developed a novel coherence model that simulates argument realization decisions. We trained this model on our automatically induced training data using features regarding predicate-argument structures affected by realization decisions, entity coreference and discourse context. Given an entity in discourse context, the trained model is able to predict whether the entity should explicitly be realized in a predicate-argument structure to establish coherence or whether it is redundant and could cause apparent incoherence. We applied this model on an annotated data set of news reports and on automatically generated summaries.

In the linking experiment, we observed improved results when using our data set of automatically induced implicit arguments as additional training data. While the model cannot compete with state-of-the-art systems, the addition of our data led to an enhanced performance compared to the same system with different and without additional training data. Compared to the model without additional training data, our induced data set increased results in terms of precision and F_1-score by 15 and 5 percentage points, respectively. In our experiments on perceived coherence, we found that the use of implicit vs. explicit arguments, while often being a subtle difference, can have a clear impact on readability ratings by human annotators. We showed that our novel coherence model, which is solely trained on automatically induced data, is able to predict this difference in newswire articles with a precision of up to 90%. Unfortunately, similar results could not be achieved on automatic summaries. We found that one reason for this lies in the interplay with other coherence-related factors that need to be explicitly considered in a setting that involves machine generated texts. Modeling this interplay is non-trivial but could also be beneficial for other tasks related to coherence modeling. We discuss this point in more detail, together with other potential applications of resources created in the context of our work, in Chapter 9.

9. Discussion

In this chapter, we discuss the benefits of our work for other lines of research in computational linguistics. We focus our discussion on applications that can directly benefit from the resources created in our induction framework. Particular resources are: our corpus of comparable texts (cf. Chapter 5), our graph-based model for aligning predicate-argument structures (cf. Chapter 6), our heuristic approach to inducing instances of implicit arguments (cf. Chapter 7), and our argument-based model of local coherence (cf. Chapter 8). We briefly review each of the four resources and outline a few potential applications. A summary of our contributions, together with directions for future work, is given in Chapter 10.

The contributions described in this chapter are divided into four parts: in Section 9.1, we give a brief summary of our corpus of comparable texts and discuss how it could be used in tasks that rely on pairs of texts or text fragments; in Section 9.2, we describe how our data set of aligned predicate-argument structures (PAS) could be used to acquire training data for other linguistic phenomena and how PAS alignments could be useful in natural language processing tasks; in Section 9.3, we reiterate properties of our induction approach and discuss how NLP applications could benefit from it; finally, in Section 9.4, we outline how our argument-based coherence model can be applied in tasks that have been addressed by previous models of local coherence. An overview of data sets created as part of our overall framework can be found in Table 9.1.

Type of data set	data points	description	technical details
Comparable text pairs	167,728	Chapter 5	Appendix B.1
Manually aligned predicate pairs	885	Chapter 6	Appendix B.2
Automatic high-precision alignments	283,588	Chapter 7	Appendix B.3
Induced implicit arguments	698	Chapter 7	Appendix B.4

Table 9.1.: Overview of our data sets

9.1. Comparable Texts

As a first step in our induction framework, we extracted pairs of comparable texts from a large corpus of news reports from different newswire agencies (cf. Chapter 5). We showed that by combining a simple extraction method from the literature (Wubben et al., 2009) with an additional date constraint, more than 160,000 pairs of documents can be extracted with high precision. In previous work, comparable texts have been used for a range of tasks in natural language processing (cf. Chapter 3). We briefly

discuss the benefits of our corpus for such tasks in the following paragraphs. A less task-oriented application of our data set would be in the area of digital humanities: by containing descriptions of real-word incidents from various international sources, the extracted pairs of texts could be used to detect different points of view and cultural differences that may reflect the fact that the news agencies in our corpus are based in different countries (Al Khatib et al., 2012).

Paraphrase detection. In previous work, comparable corpora have successfully been used as a resource for extracting paraphrases (Munteanu and Marcu, 2006; Dolan et al., 2004, cf. Section 3.1). Compared to previous corpora, our data set has two considerable advantages: firstly, it is bigger than corpora used in small-scale studies that focus on a limited domain (Regneri and Wang, 2012; Belz and Kow, 2010a); and secondly, our extraction method produces more precise results than previous attempts to perform this step automatically (Wang and Callison-Burch, 2011). We hence believe that our resulting corpus could be well suited to extract more paraphrases or to extract them with higher precision.

Semantic textual similarity (STS). Agirre et al. (2012) recently proposed STS as a unified framework for the extrinsic evaluation of semantic processing modules. The goal of the semantic textual similarity task is to automatically compute similarity scores for pairs of sentences. Computed scores are evaluated by comparison against ratings provided by human annotators. Given the high precision of our extraction method, we argue that our corpus would be a suitable resource to create data for the STS task. For example, we assume that the head lines in our comparable texts will have a high similarity, and that sentences of varying similarity could be extracted from the content of each document pair. As an indicator for similarity in our data set, we can count the relative number of aligned predicate-argument structures. Compared to previous data sets used in the STS task, our corpus further has the advantage of providing full discourse contexts. That is, since no contexts were given in previous instances of the task, only semantic phenomena on the sentence level were addressed. In contrast, our data set contains full discourse contexts and could hence be used to extend future tasks to also address discourse-level phenomena including, for example, implicit arguments.

9.2. Benefits of Aligning Predicate-Argument Structures

In the second step of our induction approach, we proposed to align predicate-argument structures that correspond to each other in comparable texts (cf. Chapter 6). In the context of our framework, we only took a closer look at pairs of aligned structures that involve a differing number of realized arguments. Examining pairs of aligned structures, in which the same arguments are realized, could be relevant for a range of other tasks though. In practice, we assume that aligned predicates refer to the same event or state, meaning that pairs of predicates can be used to disambiguate one another. Furthermore, we expect that identically labeled arguments in aligned structures are typically related

to one another, for example, in terms of coreference or bridging. These two observations give rise to several potential applications. For example, alignments can be utilized in tasks such as multi-document summarization to avoid the extraction and generation of multiple references to the same event. furthermore, alignments can also be explored as a way of assessing the similarity of two texts. As an instance of this application, Reiter (submitted) recently proposed to use our graph-based alignment approach, among other models, to find structural similarities in narrative texts. We describe two applications in more detail in the following paragraphs.

Event disambiguation. By viewing aligned predicate-argument structures as references to the same event, a further application would be to disambiguate between different events. This proposal has recently been made by Wolfe et al. (2013), who argue that such a disambiguation step is necessary to distinguish between events across documents. Potential applications for this process can be found in tasks related to question answering or information retrieval, where information on specific events and states needs to be found in documents that are not necessarily comparable. In context of our work, we briefly discussed the need for disambiguating between events in our induction framework (cf. Chapter 4). To illustrate this issue, we considered the pairs of texts shown in Example (39) and (40):

(39) It's a [private]$_{\text{MNR}}$ visit [by Bilal]$_{\text{A0}}$.

(40) a. [Bilal]$_{\text{A0}}$ is on a visit [to India]$_{\text{A1}}$. (...)

b. [Sonia Gandhi]$_{\text{A0}}$ has also been invited to visit [Pakistan]$_{\text{A1}}$

As a reader, it is easy to see that the two VISIT predicates in (39) and (40a) describe the same event and that this event differs from the visit mentioned in sentence (40b). In contrast, a machine first has to determine, for example, that "Bilal" and "Sonia Gandhi" are references to two different entities. In context of our work, we explicitly identified realizations of the same event by aligning predicates, together with their associated argument structures, in pairs of comparable texts. This effort could be extended to larger sets of documents by extending our clustering approach to sets of more than two documents.

Bridging anaphora resolution. For two predicate-argument structures to be aligned, their associated arguments should also refer to the same entities or properties. While this means that arguments in aligned structures are typically related by coreference, there are cases that involve more complex relationships. An example for this are bridging anaphora as exhibited in the aligned structures shown in Example (41) and (42a).

(41) [Japan's Foreign Ministry]$_{\text{A0}}$ issued [a travel alert]$_{\text{A1}}$ ...

(42) a. [Japan]$_{\text{A0}}$ has issued [a travel alert]$_{\text{A1}}$ (...)

b. [The Foreign Ministry's]$_{A0}$ announcement called on Japanese citizens to be cautious ...

Sentence (41) explicitly conveys the information that "Japan's Foreign Ministry" issued a travel alert. In contrast, sentence (42a) simply states that "Japan" issued an alert, omitting any more specific information on the issuer (A0). While the follow-up sentence (42b) mentions "The Foreign Ministry's announcement", the text does not make explicit that the announcement and the travel alert are identical, nor does it explicitly state that the "Foreign Ministry" is indeed "Japan's Foreign Ministry". The alignment between the PAS in (42a) with the one in (41) makes clear, however, that both entities should be related. By applying a similar technique as the induction method discussed in Chapter 7, such pairs of aligned predicate-argument structures could be exploited to automatically induce a data set that contains annotated instances of bridging anaphora. Such a data set would be particularly useful for ongoing research in the area of resolving bridging anaphora as there currently exists only very little training data that contains unrestricted bridging instances (Hou et al., 2013).

9.3. Implicit Arguments in Applications

In the third step of our framework, we introduced a heuristic approach to automatically inducing instances of implicit arguments and discourse antecedents from pairs of comparable texts (cf. Chapter 7). In Chapter 2, we motivated this framework by discussing two particular tasks, in which implicit arguments are important: namely semantic role labeling and entity-based coherence modeling. As demonstrated in Chapter 8, our automatically induced data can successfully be employed to enhance the performance of existing systems that identify and link implicit arguments in discourse. The precision of such systems, however, is currently still rather low. In contrast, the methods applied in our induction approach aim for high precision. As an alternative to identify and link implicit arguments using a SRL-based system, we could also apply our methods directly in actual tasks. Consequently, applications could benefit from insights regarding implicit arguments in two different ways: by linking implicit arguments in a pre-processing step (within text) or by aligning and merging pairs of predicate-argument structures (within or across texts). In the following, we discuss two specific applications: textual entailment and relation extraction.

Recognizing textual entailment (RTE). Since 2009, the annually organized RTE challenge involves a discourse-level task (Bentivogli et al., 2009, cf. Chapter 3). In this task, also called "search task" or "RTE within a corpus", entailing sentences for a hypothesis have to be found among a set of automatically extracted candidates from a corpus. Mirkin et al. (2010a,b) examined the data set of the search task in 2009 to study the impact of discourse referents on textual entailment. Their analysis revealed that almost half of the entailment pairs involved reference relations whose resolution was essential for correctly predicting entailment. In (43) and (44), we illustrate an example of such a

text-hypothesis pair. Resolution could be performed in this example by identifying an implicit argument and linking it to a suitable discourse antecedent.

(43) T: Bomb explosions tore through three subway trains and a red double-decker bus [in a coordinated terror attack$_{\mathbf{i}}$] (...) Authorities said 22 of the [700 people]$_{\text{A0}}$ injured [$\varnothing_{\mathbf{i}}$]$_{\text{A1}}$ remained in critical condition.

(44) H: [About 700 people]$_{\text{A0}}$ were injured [in the attack$_{\mathbf{i}}$]$_{\text{A1}}$.

As can be seen in (44), two arguments are realized in the hypothesis. In contrast, only one explicit argument can be found in the corresponding predicate-argument structure in text (43). We can make use of the preceding context, however, to look for an antecedent that matches the extra argument "the attack" from the hypothesis. In the given example, we indeed find a suitable antecedent, namely "coordinated terrorist attack", and can hence infer that there does exist an entailment relation between the text and hypothesis.

Relation extraction. Relation extraction can broadly be defined as the task of extracting relations between entities from natural language text. A specific instance of relation extraction is the TAC Knowledge Base Population task (Ji et al., 2010). In this task, entities and relations are given and the goal is to determine which relations hold between which entities. As an example, instances of parent–child relationships have to be found in a text given a pre-specified list of persons. One challenging aspect of this task is that pairs of related entities might not occur within the same sentence. According to a recent study by Ji and Grishman (2011), in fact, only 60.4% of all relation instances in the TAC data set can be extracted from within a single sentence. We illustrate one case, in which cross-sentential inference is necessary, in Example (45).

(45) a. Lahoud is married to an Armenian and [the couple]$_{\texttt{ego}}$ have three children$_{\text{KINSHIP}}$ [DNI]$_{\texttt{alter}}$.

b. Eldest son$_{\text{KINSHIP}}$ [Emile Emile Lahoud]$_{\texttt{alter}}$ [DNI]$_{\texttt{ego}}$ was a member of parliament.

In the TAC task, the goal would be to find a parent–child relation between "the couple" Lahoud and his Armenian wife and "Emile Emile Lahoud". As shown in Example (45), neither of the two sentences contains references to both entities. To correctly extract the relation nonetheless, we can apply frame-semantic role annotation and cast this problem as an implicit argument linking task. In fact, we find that the parent–child relation represents a specific instance of the KINSHIP frame in FrameNet (cf. Chapter 1), with the `ego` and `alter` roles corresponding to the parent and child, respectively. Using this knowledge, the relation can correctly be extracted by aligning and merging the predicate-argument structure in (45a) with the corresponding structure in (45b).

9.4. Employing our Model of Local Coherence

Based on the automatically induced instances of implicit arguments and their explicit counterparts in aligned predicate-argument structures, we developed a model that predicts whether an argument should be realized explicitly in a given discourse context. We introduced this coherence model together with two evaluation experiments in Chapter 8. In our experiments, we found that the performance of the model heavily depends on the provided input text. Our analysis revealed that one reason for this outcome might be a lack of coherence caused by other discourse-level factors. One way to address this problem would be to combine our model with approaches that cover other coherence-related factors (cf. Section 2.2) and to apply specific criteria already during summary generation, rather than to post-process the output of a generation system. A combination of models could also be advantageous in tasks, to which models of local coherence have previously been applied. In the following paragraph, we describe sentence ordering as an example of such a task.

Sentence ordering. The ordering of information is an essential step in text generation. In the past couple of years, a range of models were proposed to distinguish between randomly shuffled sentences and texts in their actual, i.e. "correct", order. Many of these models are entity-based, following the ideas put forward in the Centering framework (cf. Chapter 2). A recent empirical study on a newswire corpus has shown, however, that 37–51% of all adjacent sentence pairs do not share any entity references (Louis and Nenkova, 2010). Consequently, it will be difficult to correctly predict sentence order in these cases for entity-based models of coherence proposed in previous work. Example (46) shows two sentences from Louis and Nenkova's analysis and one sentence of follow-up context.

(46) a. Authorities in Hawaii said the wreckage of a missing commuter plane with 20 people aboard was spotted in a remote valley on the island of Molokai.

b. There wasn't any evidence of survivors.

c. The plane failed to reach Molokai's airport Saturday while on a flight from the neighboring island of Maui.

We applied a couple of previously proposed models of local coherence to score this text in its original sentence order and in a permuted order in which the first two sentences are switched (cf. Example 47). We find that models based on the entity grid give a higher score to the incorrect sentence order as it brings co-referring mentions closer together ("Molokai"–"Molokai's", "commuter plane"–"The plane"). Other models, including the pronoun-based model and the discourse-new model, return the same score for both permutations as the type and order of co-referring mentions remain the same. In contrast, our model can predict that the original sentence order is more coherent, given the explicitly realized arguments. In the reversed order, our model detects an explicit entity reference, which it predicts to be implicit, as illustrated in Example (47):

(47) a. There wasn't any evidence of survivors.

b. Authorities in Hawaii said the wreckage of a missing commuter plane with 20 people aboard was spotted in a remote valley on the island of Molokai.

c. The plane failed to reach (the) ~~Molokai's~~ airport Saturday while on a flight from the neighboring island of Maui.

Our model in isolation, however, is not very useful for this task as it only covers this one particular phenomenon. To cover many different types of coherence-related factors, it will be necessary to integrate various coherence models in a way that combines the strengths of each of them. We discuss this idea, together with other directions for future work, in the next chapter.

10. Conclusions

In this chapter, we summarize the contributions of this thesis and describe promising further directions. We further discuss insights gained from our experiments that will be beneficial for future work in natural language processing.

10.1. Contributions

In this thesis, we introduced a framework for inducing instances of implicit arguments and their discourse antecedents from pairs of comparable texts. As described in Chapter 4, this framework is designed as a pipeline architecture that consists of three steps: extracting pairs of comparable texts, aligning predicate-argument structures across pairs of texts, and identifying implicit arguments and antecedents. In the following paragraphs, we summarize each of these steps and highlight our contributions.

Large corpus of comparable texts. Comparable texts are a useful resource for the acquisition of paraphrases and form the basis for several NLP tasks, including multi-document summarization and (discourse-level) textual entailment recognition. In previous work, corpora of comparable texts were manually compiled or extracted with low precision from the web and other textual resources. In Chapter 5, we introduced a new corpus of comparable texts that we constructed in such a way that it does not suffer from the size and noise issues that we observed in previous corpora. To achieve this goal, we combined an established method from the literature with an additional date constraint and applied it to a large collection of newswire articles. As a result, we extracted a collection of more than 160,000 document pairs. In a sample of 70 document pairs from this corpus, our two annotators found 69 to be comparable texts, reflecting a sample precision of 98.6%.

Predicate-argument structure alignments. With the goal of inducing instances of implicit arguments, we proposed a novel task that aims at aligning pairs of predicate-argument structures (PAS) across pairs of comparable texts. In Chapter 6, we introduced a manually annotated data set for the development and evaluation of models for this particular task. We found that pairs of PAS can be aligned across documents with good inter-annotator agreement given appropriate annotation guidelines. Based on the development part of our corpus, we designed and fine-tuned a novel graph-based clustering model. To apply this model, we represent predicate-argument structures in pairs of documents as bipartite graphs and recursively divide this graph into subgraphs. All clustering decisions by the model are based on pairwise similarities between PAS, combining

information on predicates, associated arguments and their respective discourse contexts. We empirically evaluated our model against various baselines and a competitive model that has recently been proposed in the literature. The results of our evaluation show that our model outperforms all other models on the discourse-level PAS alignment task by a margin of at least 0.6 percentage points in F_1-score, despite only a single threshold parameter being adjusted on our development set. As an additional contribution, we defined a tuning procedure, in which we adjust our method for high precision. Following this tuning routine, our model is capable of aligning PAS pairs with a precision of 86.2%.

Heuristic induction method. Based on aligned pairs of predicate-argument structures, the last step in our induction framework is to identify instances of implicit arguments. In Chapter 7, we described a computational implementation of this step, in which aligned argument structures are automatically compared and discourse antecedents for implicit arguments are found by means of entity coreference chains across documents. To reduce the effect of pre-processing errors, our implementation makes use of precise pre-processing methods and a small set of restrictions that exclude instances whose automatic annotations are likely to be erroneous. We found that by combining information from different pre-processing modules, we can induce instances of implicit arguments and discourse antecedents with a sample precision of up to 89%.

Coherence modeling and implicit argument linking. To examine the utility and reliability of our data set, we additionally performed extrinsic evaluations in task-based settings. In Chapter 8, we described two particular applications of our data: linking implicit arguments to discourse antecedents and predicting coherent argument realizations. In the first application, we employed our data set as additional training data to enhance a pre-existing system for identifying and linking implicit arguments in discourse. Experimental results showed that the addition of our training data can improve model performance by 15 percentage points in precision and 5 percentage points in F_1-score.

For the second application, we developed a novel model of local coherence that predicts whether a specific argument should be explicitly realized in a given context or not. Our experiments revealed that this model, when trained on the unannotated data that we automatically induced, can predict human judgements on argument realization in coherent newswire text with a precision and recall of 90%. In comparison, we found that previous models of local coherence only achieve precision and recall scores below 50%, showing that they do not capture this phenomenon appropriately. In our final experiment, we applied our model on automatically generated summaries, with the goal of improving textual coherence and shortening summary length. We performed several studies, in which we presented original and modified summaries to human annotators and asked them to indicate their preferences. In all studies, we used our model to predict whether an argument should be implicit or explicit and automatically modified summaries accordingly. In our analysis of the collected coherence ratings, we found that human preferences overlap with the decisions of our model in 19 out of 26 cases (73%).

10.2. Potential Improvements

There are a number of ways in which methods and models developed in this thesis could be enhanced. As an example, we discussed in Chapter 9 that combining our coherence model with models from previous work will be useful to make more informed predictions of argument realization. Given our analysis of results on multi-document summarization (cf. Chapter 8), we expect this to be a necessary step to make our model more robust across different settings. In another application of our data set, namely the linking of implicit arguments in discourse, we saw that induced instances of implicit arguments can be applied as training material to enhance the precision of automatically learned models. The recall achieved by the enhanced model is, however, still low compared to current state-of-the-art models (8% vs. up to 25%). We believe that future work will have to closer examine the strengths of each model in order to come up with ways to achieve better results both in terms of precision and recall.

Revising precision and recall requirements. One potential way of achieving better results in applications is to induce more diverse training instances. Using the current induction method, which we described in Chapter 7, only around 700 instances of implicit arguments and antecedents were extracted from a corpus of over 160,000 document pairs. Given the high precision of pre-processing models and hard restrictions in our heuristic induction method, our approach is artificially limited to specific kinds of instances. One way to improve the resulting coverage would be to replace some of the existing steps with probabilistic methods that can assign confidence values to all automatic annotations (semantic role labels, alignments, coreference chains). Given such confidence values, we could then minimize the impact of unreliable annotations and other error sources by jointly optimizing confidence thresholds while training models in task-based settings.

Removing lexical restrictions. A related problem lies in the fact that our induction framework mainly builds on predicate-specific argument labels. By following the PropBank/NomBank-paradigm, we benefit from robust and precise shallow semantic parsers. It can be difficult, however, to identify different predicate types that convey information about the same event and to relate their arguments to one another. This problem does not only affect our method for inducing implicit arguments but also the coverage of our PAS alignment model (cf. Chapter 6). In the alignment model, we can solve this issue partially by mapping predicates and arguments to frames and frame elements in FrameNet via the VerbNet mappings provided by SemLink. Yet, new errors can emerge in this process, affecting both correctness and completeness. In particular, this is the case for ambiguous mappings from PropBank arguments to FrameNet. Furthermore, VerbNet and FrameNet are incomplete resources, meaning that they do not cover all predicates of the English language and their different senses. An alternative to employing hand-crafted resources would be to rely on unsupervised semantic role labeling techniques. In recent years, such techniques have received an increasing interest as they are not bound to a specific domain or manually designed role sets.

10.3. Directions for Future Work

There are a number of applications that, in future work, can benefit from the insights gained in this thesis. As a starting point, we discussed several example applications for each of our contributions in Chapter 9. In particular, we expect that the identification of implicit arguments can provide a boost to tasks related to information extraction and question answering. Furthermore, natural language generation applications can benefit from coherence models that cover argument realization as an additional aspect.

Information access. While we have seen that identifying implicit arguments and antecedents is a difficult task in general, current methods can already be applied reliably in more restricted settings. Empirical evidence for this claim can be found in previous work by Gerber and Chai (2012) and Moor et al. (2013), who report precision figures of 57.9% and 47.7%, respectively, when using models of implicit argument linking with predicate-specific features and gold labels of arguments that are resolvable (cf. Chapter 2). Another, yet unexplored way to apply models with high precision would be to focus on domain-specific factors: for example, we find that the first sentence of a news article typically conveys information on the same event as its head line, but the realized details can complement each other; furthermore, implicit arguments in citations often refer to entities that are mentioned in the immediate context

Coherence modeling. In our experiments on local coherence, we observed that argument realization can affect readability according to human judgments (cf. Chapter 8). We found that both implicit and explicit arguments can be preferable, depending on the given contexts. Based on this observation, we developed a model that can predict the preferred realization type of an argument. Applied to full news texts, our model achieves a high precision on this task despite being trained only on 698 automatically annotated data points. The results on automatically generated summaries, however, were not as high. As discussed in Chapter 9, potential improvements could be achieved by combining several models of coherence that cover different but interfering phenomena.

In summary, a considerable amount of work still needs to be done to enhance models for handling implicit arguments in discourse. In the long run, however, this research direction will be beneficial for any application that involves the understanding or generation of text beyond the sentence level. In this thesis, we provided several research contributions that form a reliable basis for future work. In particular, we developed a framework for automatically inducing instances of implicit arguments, and we designed a novel coherence model that predicts the effect of argument realizations on perceived textual coherence. From a theoretical perspective, we validated that both explicit and implicit arguments can affect coherence and that automatically induced training data can be utilized to model this phenomenon appropriately. We further showed that our induced data set, which contains instances of implicit arguments and discourse antecedents, can be applied to enhance current models for implicit argument linking. Future work will be able to build on these insights, further enhance existing models and apply them in order to improve current state-of-the-art NLP systems.

A. Guidelines for Aligning Predicate-Argument Structures

A.1. Introduction

Annotators are provided with pairs of newswire articles describing the same news from the perspective of two distinct sources. The descriptions can vary both with regard to content and linguistic realization. As an annotator, your task will be mark corresponding pairs of predicates across documents. Note that for this task, it is important to also take the context of the predicate into consideration! The predicates have been (automatically) pre-selected and marked in boldface to ease annotation. We further provide indices for all predicates, so that you can uniquely identify them in a given text. If you notice any unmarked predicate, which you do want to align, please note them separately. We show an example text pair and annotation in the following:

A "Peru's Luis Horna **clinched**$_1$ his second career ATP **title**$_2$ with a 7-5, 6-3 **win**$_3$ over local favourite Nicolas Massu on the **clay**$_4$ of Vina del Mar on Sunday."

A' "Luis Horna of Peru **defeated**$_1$ hometown favorite Nicolas Massu 7-5 , 6-3 for the first time in the Movistar Open final on Sunday and **claimed**$_2$ his second career ATP **title**$_3$."

Annotation (A–A'): 1–2 2–3 3–1

As you can see in the given example, the differences between corresponding predicates and their contexts can be very small: in some cases, a synonymous predicate (e.g., "win"–"defeat", "clinch"–"claim") was used, and in other cases, some extra information was introduced (e.g., "for the first time", "on Sunday"). However, not all cases are this simple. There might be a correspondency, which only becomes apparent when considering the actual meaning of the concerned predicate argument structures in context. Even if the correspondency seems rather loose, we aim to also take these cases into account. Here is an example illustrating two such correspondencies:

B "(...) **Spokespeople**$_1$ at Pfizer's China **operations**$_2$ were not immediately available to **comment**$_3$ on Monday."

B' "(...) Phone **calls**$_1$ to Pfizer's China **headquarters**$_2$ in Beijing were not **answered**$_3$."

Annotation (B–B'): 2–2 3–3–P

Your task as an annotator will be to mark alignments as in the above examples. We provide pairs of texts in a simple text format for this task. As in the example annotations, you should write down the index pairs of corresponding predicates. We will provide you with a separate file for these annotations. Here are some additional guidelines to follow:

1. You have the option to mark alignments as "possible" using the suffix '–P' for cases, in which you feel uncertain or in which the text does not make clear, which event, state or object is being referred to. However, your main focus should lie on "sure" alignments, i.e., alignments that you are certain about.

2. You should prefer to mark alignments on a 1-to-1 basis whenever it is possible. However, you can indicate n-to-m correspondences when necessary.

3. Spend as much as time as needed to think about the meaning of marked predicates to make sure that you do not miss complex correspondencies that seem unlikely on first sight.

The next section describes the overall annotation process in a bit more detail.

A.2. Details

Before starting to annotate a pair of text, please make sure to read both texts carefully from beginning to end. This allows you to get an overall picture of the content and details included in each text. Depending on the length of a text, you might also want to pre-structure the document and remove paragraphs that are only contained in one of the two texts. However, please keep the predicate indices as they are for your annotation!

Once you have a good feeling for what the content of each text is, you can start the actual task. We do not provide a strict definition as to when two predicates correspond and should be aligned. As a rule of thumb, you can think of correspondency as a measure for how well one predicate argument structure can be replaced in context with another. If it is possible to exchange both predicate argument structures without changing the meaning of a text, then you should probably align the two.

As mentioned before, you should try to mark alignments on a 1-to-1 basis. However, there are cases, where this rule is not possible due to syntactic constructions and the meaning of predicates. For example, "**rear** and **spew**" are two predicates that can have the same meaning as the single predicate "**erupt**", depending on the context. In these cases, you should align all affected predicates in one text with all affected predicates in the other (for example, "1–1", "2–1").

Apart from simply marking two predicates as corresponding, you have the option to mark alignments as "possible". In particular, you should make use of this option, if you think that a correspondency between two predicates depends on one particular interpretation of one of the predicates. We have seen one such example in the previous section:

> "spokespeople were not available to **comment**"

"phone calls were not **answered**"

A.3. Special Cases

There are a number of special cases that you should pay attention to in this task:

Exact correspondency. If two predicates are identical and their arguments overlap, they should almost certainly be annotated using a sure alignment. The only exception from this rule would be if the arguments occurred in reverse order and led to a contradictive meaning. In other words, you should not align cases such as C but you should always align an example such as D:

C "VW **bought** Porsche" – "Porsche **bought** VW" (incorrect)

D "VW **bought** Porsche" – "VW **bought** Porsche for USD 5.6bn" (correct)

Pronouns. When comparing the arguments and other contexts of two predicates, you should also check whether pronouns in one structure correspond to anything in the other. Here is another example:

E "He was **joined** by the Bassac River by his wife"

E' "Hun Sen's wife **stood** at her husband's side"

Spelling mistakes. Some newswire articles contain spelling mistakes and other errors. You can simply ignore them as long as the actual meaning of the text is still clear.

Approximate correspondencies. Two predicates can correspond, even if they are not synonymous. In particular, this can be the case even if one predicate describes a different perspective on an event, state or object (e.g., **buy** vs. **sell**). It can also be the case that one predicate only describes a part of the concept described by the other (cf. example E). If it is clear that the event, state or object is the same though, you should also annotate these pairs using sure alignments.

E "The soldier was **killed** during a patrol in the area south of Baghdad."

E' "The soldier **died** in an attack close to the capital Baghdad."

Repetitions. If one newswire article refers to the same event, state or object multiple times, but the other article only once, then only the first correspondency should be marked as a sure alignment[1]. Further correspondencies should also be annotated but

[1] The intuition behind this guideline is that the first mention introduces the actual concept while later mentions just (co-)refer or add further information, i.e., they serve a different function with respect to the discourse.

only as possible alignments ('–P'!). In general, if there are multiple references in both texts, each reference should be annotated using a sure alignment at most once. In these cases, you should mark the predicates with the highest information overlap as "sure". Here is an example:

F "Susan Boyle said she will **sing**$_1$ in front of Britain's Prince Charles (...) 'It's going to be a privilege to be **performing**$_2$ before His Royal Highness', the singer said (...) British copyright laws will allow her to **perform**$_3$ the hit in front of the prince and his wife."

F' "British singing sensation Susan Boyle is going to **perform**$_1$ for Prince Charles (...) The show star will **perform**$_2$ her version of Perfect Day for Charles and his wife Camilla."

Annotation (F–F'): 1–1 1–2–P 2–1–P 2–2–P 3–1–P 3–2

Note that the example annotation for F–F' only shows one possible way of aligning the occurring predicates. Depending on the interpretation of each predicate and its contexts, a different annotation might be equally good.

B. Description of Data Sets

B.1. Pairs of Comparable Texts

The data set of comparable text pairs can be downloaded using the following URL:

`http://projects.cl.uni-heidelberg.de/india/files/gigapairs-doc-ids.tar.gz`

The archive contains IDs of all 167,728 text pairs that we automatically extracted from the English Gigaword Fifth Edition (Parker et al., 2011). For each pair of newswire sources, there exists one file that contains the IDs of all document pairs that are predicted to be comparable. Each file lists these IDs in a tab-separated format, with each line corresponding to one document pair. In addition to document IDs, the files contain corresponding similarity scores, computed using the method described in Chapter 5 of a document pairing in the first column each.

B.2. Manual Predicate Alignments

The data set of manual predicate alignments can be downloaded using the following URL:

`http://projects.cl.uni-heidelberg.de/india/files/manual-alignments.tar.gz`

The archive contains gold standard predicate alignments for 70 comparable text pairs extracted from the English Gigaword Fifth Edition (Parker et al., 2011). We provide these alignments in two separate ways, described in the following paragraphs.

XML. The most simple way to view the annotations is to run the provided script (`run.sh`), which extracts news reports from the Gigaword corpus, creates XML documents and automatically inserts all alignments. To run the script, simply execute the command `sh run.sh [GIGAWORDDIRECTORY]`. The script will create a directory `XML/` in the current directory, which will contain two subdirectories `dev/` and `test/` for the development and testing documents, respectively. Note that alignments will be inserted into the XML files in form of `ALIGNED` tags, each of which contains two attributes: `type` and `set`. The type attribute refers to whether the alignment has been marked as 'sure' or 'possible'. The set attribute serves as a unique identifier for each alignment pair. Note that if a predicate has not been aligned, there will be an `ALIGNED` element with the alignment type 'none'; if a predicate has been aligned multiple times, it will be marked by multiple tags.

Stand-off annotation. If you do not own a license of Gigaword or if you just want to see the word forms of aligned predicates, you can view the file `stand_off_annotations.txt` (or `stand_off_annotations.no_unaligned.txt`) using any standard text editor. Note that each line refers either to a single document ID, a unique handle for pairs of documents, or an aligned predicate. Each alignment annotation has the following form:

```
[PARAGRAPH_NUMBER] [OCCURRENCE_NUMBER] [PREDICATE] [ALIGNMENTID] [ALIGNMENTTYPE]
```

You can manually investigate aligned predicates by comparing the `alignment_id` columns underneath the document IDs that occur within the same directory handle.

B.3. Automatic High-Precision Alignments

The data set of predicate-argument structures that have been aligned with our high precision model can be downloaded using the following URL:

```
http://projects.cl.uni-heidelberg.de/india/files/automatic-alignments.tar.gz
```

The archive contains automatic high-precision alignments for 283,588 predicate pairs that occur in our data set of comparable texts (cf. Appendix B.1) Each file in this archive provides a list of aligned predicates for a given pair of newswire sources in the English Gigaword Fifth Edition (Parker et al., 2011). For example, `afp-apw.out` contains all predicate alignments between documents from Agence France-Press (AFP) and from Associate Press Worldstream (APW). Each alignment is specified in the following format:

```
[DOCID],[SENTENCEID],[TOKENID],[WORD] \t [DOCID],[SENTENCEID],[TOKENID],[WORD]
```

The document IDs refer to the original IDs as contained in Gigaword. Sentence and token IDs refer to automatically-generated annotation using Stanford CoreNLP (Toutanova et al., 2003). We additionally provide the word form of each predicate for performing automatic sanity checks.

All alignments are automatically computed using the high-precision version of our clustering approach, described in Chapter 6: instead of tuning F_1-score on the development set, we tuned $F_{0.33}$, i.e., we weighted precision three times higher than recall. On the manually aligned test set (cf. Appendix B.2), the tuned method achieves a precision and recall of 86.2% and 29.1%, respectively.

B.4. Induced Implicit Arguments and Discourse Antecedents

The data set of automatically induced instances of implicit arguments and discourse antecedents can be downloaded using the following URL:

```
http://projects.cl.uni-heidelberg.de/india/files/implicit-arguments.tar.gz
```

The archive contains annotations for 698 instances of implicit arguments and discourse antecedents that were automatically extracted from our data set of comparable texts (cf. B.1). Each implicit argument is specified in the following format:

`[DOCID],[SENTID],[TOKENID],[WORD],[ARG_LABEL],[ARGSENTID],[ARGTOKENID],[ARGWORD]`

All IDs refer to documents in Gigaword, with sentence and token numbers refering to annotations produced by Stanford CoreNLP (cf. Appendix B.3). `SENTID` and `TOKENID` indicate the position of the predicate, for which an implicit argument with label `ARG_LABEL` was detected. `ARGSENTID` and `ARGTOKENID` describe the position of the discourse antecedent. Note that the identified antecedent can span multiple words. In this case, the position is given as the index of the first and the last token in the span, concatenated by two dots. For sanity checking, `WORD` and `ARGWORD` provide the word form of the predicate and the head word of the argument, respectively.

Bibliography

Agirre, Eneko, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluations*. Montreal, Canada. To appear.

Al Khatib, Khalid, Hinrich Schütze, and Cathleen Kantner. 2012. Automatic detection of point of view differences in Wikipedia. In *Proceedings of COLING 2012*, 33–50. Mumbai, India: The COLING 2012 Organizing Committee.

Allerton, D. J. 1982. *Valency and the English Verb*. Academic Press.

Álvez, Javier, Jordi Atserias, Jordi Carrera, Salvador Climent, Egoitz Laparra, Antoni Oliver, and German Rigau. 2008. Complete and consistent annotation of WordNet using the top concept ontology. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*.

Bär, Daniel, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. UKP: Computing semantic textual similarity by combining multiple content similarity measures. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 435–440. Montréal, Canada: Association for Computational Linguistics.

Baroni, Marco, and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36(4):673–721.

Barzilay, Regina, and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing,* Sapporo, Japan, 11–12 July 2003, 25–32.

Barzilay, Regina, and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Michigan, USA, 25–30 June 2005, 141–148.

———. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics* 34(1):1–34.

Barzilay, Regina, and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics,* Toulouse, 50–57.

Baum, Leonard E. 1972. An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities* 3:1–8.

Belz, Anja, and Eric Kow. 2010a. Extracting parallel fragments from comparable corpora for data-to-text generation. In *Proceedings of the 6th International Natural Language Generation Conference (INLG'10)*, 167–171. Trim, Ireland.

———. 2010b. The GREC challenges 2010: Overview and evaluation results. In *Proceedings of the Sixth International Natural Language Generation Conference*, 219–229.

———. 2011. Discrete vs. continuous rating scales for language evaluation in NLP. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 230–235. Portland, Oregon, USA: Association for Computational Linguistics.

Belz, Anja, Eric Kow, Jette Viethen, and Albert Gatt. 2008. The GREC challenge 2008: Overview and evaluation results. In *Proceedings of the Fifth International Natural Language Generation Conference*, 183–191.

———. 2009. The GREC main subject reference generation challenge 2009: overview and evaluation results. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, 79–87.

Belz, Anja, and Sebastian Varges. 2007. Generation of repeated references to discourse entities. In *Proceedings of the 11th European Workshop on Natural Language Generation,* Schloss Dagstuhl, Germany, 17–20 June 2007, 9–16.

Bentivogli, Luisa, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of TAC*.

Berg-Kirkpatrick, Taylor, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 481–490. Portland, Oregon, USA: Association for Computational Linguistics.

Biemann, Chris. 2006. Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, 73–80. New York City, New York, USA: Association for Computational Linguistics.

Björkelund, Anders, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstration Volume*, 33–36. Beijing, China: Coling 2010 Organizing Committee.

Bohnet, Bernd. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 89–97. Beijing, China.

Bonial, Claire, Jena Hwang, Julia Bonn, Kathry Conger, Olga Babko-Malaya, and Martha Palmer. 2012. English PropBank Annotation Guidelines. Tech. Rep., University of Colorado at Boulder. Annotation Guidelines (Version 3.1).

Bresnan, Joan. 1982. *Mental Representation of Grammatical Relations.* Cambridge, Mass.: MIT Press.

Brockett, Chris. 2007. Aligning the RTE 2006 corpus. Tech. Rep., Microsoft Research.

Brown, Peter F., Vincent J. Della Pietra, Stephan A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19:263–311.

Burchardt, Aljoscha, Anette Frank, and Manfred Pinkal. 2005. Building Text Meaning Representations from Contextually Related Frames – A Case Study. In *Proceedings of the 6th International Workshop on Computational Semantics (IWCS-6)*, 66–77. Tilburg, The Netherlands.

Burstein, Jill, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 681–684. Los Angeles, California, USA: Association for Computational Linguistics.

Cahill, Aoife, and Arndt Riester. 2009. Incorporating information status into generation ranking. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 817–825. Suntec, Singapore.

Cai, Jie, and Michael Strube. 2010. End-to-end coreference resolution via hypergraph partitioning. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 143–151. Beijing, China.

Chang, Chih-Chung, and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(27):1–27.

Charniak, Eugene, and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 148–156. Athens, Greece.

Chen, Desai, Nathan Schneider, Dipanjan Das, and Noah A. Smith. 2010. SEMAFOR: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 264–267. Uppsala, Sweden.

Chen, Zheng, and Heng Ji. 2010. Graph-based clustering for computational linguistics: A survey. In *Proceedings of TextGraphs-5 - 2010 Workshop on Graph-based Methods for Natural Language Processing*, 1–9. Uppsala, Sweden: Association for Computational Linguistics.

Choi, Jinho D., and Martha Palmer. 2011. Transition-based semantic role labeling using predicate argument clustering. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, 37–45. Portland, Oregon, USA.

Christensen, Janara, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards coherent multi-document summarization. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1163–1173. Atlanta, Georgia: Association for Computational Linguistics.

Cohn, Trevor, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for development and evaluation of paraphrase systems. *Computational Linguistics* 34(4).

Cote, Sharon. 1998. Ranking forward-looking centers. In *Centering in discourse*, ed. M.A. Walker, A.K. Joshi, and E.F. Prince, 55–69. Oxford, U.K.: Oxford University Press.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges*, ed. J. Quiñonero-Candela, I. Dagan, and B. Magnini, 177–190. Heidelberg, Germany: Springer.

Dahl, Deborah A. 1986. Focusing and reference resolution in PUNDIT. In *Proceedings of the 5th National Conference on Artificial Intelligence,* Philadelphia, Pennsylvania, USA, 11–15 August 1986, 1083–1088.

Dale, Robert. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes.* Cambridge, Massachusettes, USA: MIT Press.

Dale, Robert, and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 18:233–263.

Dang, Hoa Trang, and Karolina Owczarzak. 2008. Overview of the TAC 2008 update summarization task. In *Proceedings of the First Text Analysis Conference.* National Institute of Standards and Technology.

Das, Dipanjan, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 948–956. Los Angeles, California, USA: Association for Computational Linguistics.

Das, Dipanjan, and Noah A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* Portland, Oregon, 19–24 June 2011.

Daumé III, Hal. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`.

Delmonte, Bristot Antonella Piccolino Boniforti Marco Aldo Tonelli Sara, Rodolfo. 2006. Another evaluation of anaphora resolution algorithms and a comparison with GETARUNS' knowledge rich approach. In *Proceedings of the 4th International workshop on RObust Methods in Analysis of Natural language Data (ROMAND 2006)*, 3–10.

Delmonte, Tonelli S. Piccolino Boniforti M. A. Bristot A. Pianta E., R. 2005. VENSES A linguistically-based system for semantic evaluation. In *Proceedings of the first PASCAL RTE Workshop.*

DeNero, John, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 314–323. Honolulu, Hawaii.

Di Eugenio, Barbara. 1990. Centering theory and the Italian pronominal system. In *Proceedings of the 13th International Conference on Computational Linguistics,* Helsinki, Finland, 20–25 August 1990, vol. 2, 270–275.

Dolan, Bill, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of Coling 2004*, 350–356. Geneva, Switzerland.

Dolan, William B., and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing.*

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67(1): 547–619.

Eisner, Jason. 2003. Learning non-isomorphic tree mappings for machine translation. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, 205–208. Sapporo, Japan: Association for Computational Linguistics.

Elsner, Micha, Joseph Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. 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, 436–443. Read this version: http://www.cs.brown.edu/ melsner/order.pdf.

Elsner, Micha, and Eugene Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings of ACL-08: HLT, Short Papers*, 41–44. Columbus, Ohio.

———. 2011a. Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1179–1189. Portland, Oregon, USA: Association for Computational Linguistics.

———. 2011b. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 125–129. Portland, Oregon, USA: Association for Computational Linguistics.

Erk, Katrin, and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25-27 October 2008.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Massachusetts, USA: MIT Press.

Feng, Lijun, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, 276–284. Beijing, China: Coling 2010 Organizing Committee.

Feng, Vanessa Wei, and Graeme Hirst. 2014. Patterns of local discourse coherence as a feature for authorship attribution. *Literary and Linguistic Computing* 29:191–198.

Filippova, Katja, and Michael Strube. 2007. Generating constituent order in German clauses. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June 2007, 320–327.

Fillmore, Charles J. 1969. *Types of lexical information*, 370–392. Semantics, Cambridge University Press.

———. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, vol. 280, 20–32.

———. 1986. Pragmatically controlled zero anaphora. In *Proceedings of the twelfth annual meeting of the Berkeley Linguistics Society*, 95–107.

Fillmore, Charles J., and Collin F. Baker. 2001. Frame semantics for text understanding. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*. Association for Computational Linguistics.

Flake, Gary William, Robert E. Tarjan, and Kostas Tsioutsiouliklis. 2004. Graph Clustering and Minimum Cut Trees. *Internet Mathematics* 1(4):385–408.

Fleiss, Joseph L, Bruce Levin, and Myunghee Cho Paik. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions* 2:212–236.

Fraser, Bruce, and John R. Ross. 1970. Idioms and unspecified NP deletion. *Linguistic Inquiry* 1:264–265.

Gerber, Matthew, and Joyce Chai. 2010. Beyond nombank: A study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the*

Association for Computational Linguistics, 1583–1592. Uppsala, Sweden: Association for Computational Linguistics.

———. 2012. Semantic Role Labeling of Implicit Arguments for Nominal Predicates. *Computational Linguistics* 38(4):755–798.

Gerber, Matthew, Joyce Chai, and Adam Meyers. 2009. The role of implicit argumentation in nominal SRL. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 146–154. Boulder, Colorado, USA: Association for Computational Linguistics.

Gildea, Daniel. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 80–87. Sapporo, Japan: Association for Computational Linguistics.

———. 2004. Dependencies vs. constituents for tree-based alignment. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* Barcelona, Spain, 25–26 July 2004, 214–221.

Goldberg, Andrew V., and Robert E. Tarjan. 1986. A new approach to the maximum flow problem. In *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, 136–146. New York, NY, USA.

Gordon, Peter C., and Davina Chan. 1995. Pronouns, passives and discourse coherence. *Journal of Memory and Language* 34:216–231.

Gordon, Peter C., Barbara J. Grosz, and Laura A. Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science* 17:311–347.

Gorinski, Philip, Josef Ruppenhofer, and Caroline Sporleder. 2013. Towards weakly supervised resolution of null instantiations. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, 119–130. Potsdam, Germany.

Grosz, Barbara J. 1977. The representation and use of focus in a system for understanding dialogs. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence,* Cambridge, Massachusetts, USA, 22–25 August 1977, vol. 1, 67–76.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2):203–225.

Guinaudeau, Camille, and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, 93–103. Sofia, Bulgaria: Association for Computational Linguistics.

Guo, Weiwei, and Mona Diab. 2011. Semantic topic models: Combining word distributional statistics and dictionary definitions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 552–561.

Hahn, Udo, and Donna Harman, eds. 2002. *Text Summarization – Proceedings of the Workshop*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Halliday, M. A. K., and Ruqaiya Hasan. 1976. *Cohesion in English*. London, U.K.: Longman.

Hou, Yufang, Katja Markert, and Michael Strube. 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*, 907–917. Atlanta, Georgia: Association for Computational Linguistics.

Huang, Shudong, David Graff, and George Doddington. 2002. *Multiple-Translation Chinese Corpus*. Linguistic Data Consortium, Philadelphia.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering* 11(3):311–325.

Ji, Heng, and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1148–1158. Portland, Oregon, USA: Association for Computational Linguistics.

Ji, Heng, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC 2010 knowledge base population track. In *Proceedings of the third Text Analysis Conference (TAC 2010)*.

Johansson, Richard, and Pierre Nugues. 2006. A framenet-based semantic role labeler for swedish. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 436–443. Sydney, Australia.

Joshi, Aravind K., and Scott Weinstein. 1981. Control of inference: Role of some aspects of discourse structure – centering. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence,* Vancouver, British Columbia, Canada, 24–28 August 1981, 385–387.

Kamp, Hans. 1981. A theory of truth and semantic representation. In *Formal methods in the study of language*, ed. J. Groenendijk, Th. Janssen, and M. Stokhof, 277–322. Amsterdam: Mathematisch Centrum Tracts.

Kay, Martin, and Martin Röscheisen. 1993. Text-translation alignment. *Computational Linguistics* 19(1):121–142.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation Journal* 42(1): 21–40.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, vol. 5, 79–86.

Kozhevnikov, Mikhail, and Ivan Titov. 2013. Crosslingual transfer of semantic role models. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria: Association for Computational Linguistics. To appear.

Krahmer, Emiel, Sebastiaan van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics* 29(1):53–72.

Landauer, T. K., and S. T. Dumais. 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.

Laparra, Egoitz, and German Rigau. 2012. Exploiting explicit annotations and semantic types for implicit argument resolution. In *Proceedings of the Sixth IEEE International Conference on Semantic Computing (ICSC 2010)*, 75–78. Palermo, Italy: IEEE Computer Society.

———. 2013a. ImpAr: A deterministic algorithm for implicit semantic role labelling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1180–1189. Sofia, Bulgaria: Association for Computational Linguistics.

———. 2013b. Sources of evidence for implicit argument resolution. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, 155–166. Potsdam, Germany.

Leacock, Claudia, and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In *WordNet. An Electronic Lexical Database*, ed. C. Fellbaum, chap. 11, 265–283. Cambridge, Massachusetts, USA: MIT Press.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics* 39(4). Accepted for publication.

Lee, Heeyoung, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 489–500. Jeju Island, Korea.

Lesk, Michael. 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, 24–26.

Liang, Percy, Benjamin Taskar, and Dan Klein. 2006. Alignment by agreement. In *North American Association for Computational Linguistics (NAACL)*, 104–111. Association for Computational Linguistics.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, ed. Stan Szpakowicz Marie-Francine Moens, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Lin, Dekang. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning,* Madison, Wisconsin, USA, 24–27 July 1998, 296–304.

Liu, Hugo, and Push Singh. 2004. ConceptNet – a practical commonsense reasoning tool-kit. *BT technology journal* 22(4):211–226.

Louis, Annie, and Ani Nenkova. 2010. Creating local coherence: An empirical assessment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 313–316. Los Angeles, California, USA: Association for Computational Linguistics.

MacCartney, Bill, Michael Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25-27 October 2008.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics* 19(2):313–330.

Martschat, Sebastian, Jie Cai, Samuel Broscheit, Éva Mújdricza-Maydt, and Michael Strube. 2012. A multigraph model for coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, 100–106. Jeju Island, Korea.

Matthews, Peter H. 1981. *Syntax.* Cambridge University Press.

McIntyre, Neil, and Mirella Lapata. 2009. Learning to tell tales: A data-driven approach to story generation. 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, 217–225.

———. 2010. Plot induction and evolutionary search for story generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1562–1572. Uppsala, Sweden: Association for Computational Linguistics.

McKeown, Kathleen R., and Dragomir Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval,* Seattle, Washington, USA,

9–13 July 1995, 74–82. Reprinted in *Advances in Automatic Text Summarization*, Mani, I. and Maybury, M.T. (Eds.), Cambridge, Mass.: MIT Press, 1999, pp.381-389.

Meyers, Adam, Ruth Reeves, and Catherine Macleod. 2008. *NomBank v1.0.* Linguistic Data Consortium, Philadelphia.

Mirkin, Shachar, Jonathan Berant, Ido Dagan, and Eyal Shnarch. 2010a. Recognising entailment within discourse. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 770–778. Beijing, China: Coling 2010 Organizing Committee.

Mirkin, Shachar, Ido Dagan, and Sebastian Padó. 2010b. Assessing the role of discourse references in entailment inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010.

Mitchell, Jeff, and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science* 34(8):1388–1429.

Mittwoch, Anna. 1971. Idioms and unspecified NP deletion. *Linguistic Inquiry* 2:255–259.

Moor, Tatjana, Michael Roth, and Anette Frank. 2013. Predicate-specific annotations for implicit role binding: Corpus annotation, data analysis and evaluation experiments. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, 369–375. Potsdam, Germany.

Munteanu, Dragos Stefan, and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 81–88. Sydney, Australia: Association for Computational Linguistics.

Newman, Mark E. J. 2004. Analysis of weighted networks. *Physical Review E* 70(5).

Och, Franz Josef, and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

Padó, Sebastian, and Mirella Lapata. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics* 33(2):161–199.

———. 2009. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research* 36:307–340.

Palmer, Martha. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference.*

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–105.

Palmer, Martha, Daniel Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling.* Synthesis Lectures on Human Language Technologies, Morgan & Claypool.

Palmer, Martha S., Deborah A. Dahl, Rebecca J. Schiffman, Lynette Hirschman, Marcia Linebarger, and John Dowding. 1986. Recovering implicit information. In *Proceedings of the 24th annual meeting of the association for computational linguistics,* new york, n.y., 10–13 june 1986, 10–19.

Parker, Robert, David Graff, Jumbo Kong, Ke Chen, and Kazuaki Maeda. 2011. *English Gigaword Fifth Edition.* Linguistic Data Consortium, Philadelphia.

Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. 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, 267–270.

Pitler, Emily, and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 186–195. Honolulu, Hawaii.

van der Plas, Lonneke, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 299–304. Portland, Oregon, USA: Association for Computational Linguistics.

Poesio, Massimo. 2004. The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue,* Cambridge, Massachusetts, USA, 30 April–1 May 2004, 154–162.

Postolache, Oana, Dan Cristea, and Constantin Orasan. 2006. Transferring coreference chains through word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 889–892. Genoa, Italy.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, 1–40. Jeju Island, Korea.

Quirk, Chris, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP 2004*, 142–149. Barcelona, Spain.

Regneri, Michaela, and Rui Wang. 2012. Using discourse information for paraphrase extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 916–927. Jeju Island, Korea.

Reiter, Ehud, and Robert Dale. 2000. *Building Natural Language Generation Systems.* Cambridge, U.K.: Cambridge University Press.

Reiter, Nils. submitted. Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms. Ph.D. thesis, Heidelberg University.

Resnik, Philip. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* 11:95–130.

Resnik, Philip, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a parallel corpus: Annotating the book of 2000 tongues. *Computers and the Humanities* 33(1–2):129–153.

Richens, Tom. 2008. Anomalies in the WordNet verb hierarchy. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 729–736. Manchester, UK: Coling 2008 Organizing Committee.

Roth, Michael, and Anette Frank. 2012a. Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, 218–227. Montreal, Canada.

———. 2012b. Aligning predicates across monolingual comparable texts using graph-based clustering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 171–182. Jeju Island, Korea.

———. 2013. Automatically identifying implicit arguments to improve argument linking and coherence modeling. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, 306–316. Atlanta, Georgia, USA: Association for Computational Linguistics.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010a. *FrameNet II: Extended Theory and Practice.* .

Ruppenhofer, Josef, Russell Lee-Goldman, Caroline Sporleder, and Roser Morante. 2012. Beyond sentence-level semantic role labeling: linking argument structures in discourse. *Language Resources and Evaluation* 1–27.

Ruppenhofer, Josef, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010b. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluations*, 45–50. Uppsala, Sweden.

Schaeffer, Satu Elisa. 2007. Graph Clustering. *Computer Science Review* (1):27–64.

Shen, Siwei, Dragomir R. Radev, Agam Patel, and Güneş Erkan. 2006. Adding syntax to dynamic programming for aligning comparable texts for the generation of paraphrases. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 747–754. Sydney, Australia.

Shi, Jianbo, and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8):888–905.

Shinyama, Yusuke, Satoshi Sekine, Kiyoshi Sudo, and Ralph Grishman. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of the human language technology conference,* san diego, cal., 24–27 march 2002, 40–46.

Sidner, Candace L. 1979. Towards a computational theory of definite anaphora comprehension in English. Tech. Rep. AI-Memo 537, Massachusetts Institute of Technology, AI Lab, Cambridge, Mass.

———. 1981. Focusing for interpretation of pronouns. *American Journal of Computational Linguistics* 7(4):217–231.

Silberer, Carina, and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, 1–10. Montréal, Canada.

Socher, Richard, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems (NIPS 2011)*, 801–809.

Spreyer, Kathrin, and Anette Frank. 2008. Projection-based acquisition of a temporal labeller. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, 489–496.

Strube, Michael, and Udo Hahn. 1996. Functional centering. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics,* Santa Cruz, California, USA, 24–27 June 1996, 270–277.

Su, Fangzhong, and Katja Markert. 2009. Subjectivity recognition on word senses via semi-supervised mincuts. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 1–9. Boulder, Colorado: Association for Computational Linguistics.

Surdeanu, Mihai, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, 159–177. Manchester, England: Coling 2008 Organizing Committee.

Tetreault, Joel. 2002. Implicit role reference. In *Proceedings of the International Symposium on Reference Resolution for Natural Language Processing*, 109–115.

Titov, Ivan, and Mikhail Kozhevnikov. 2010. Bootstrapping semantic analyzers from non-contradictory texts. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, 958–967.

Tonelli, Sara, and Rodolfo Delmonte. 2010. VENSES++: Adapting a deep semantic processing system to the identification of null instantiations. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 296–299. Uppsala, Sweden.

———. 2011. Desperately seeking implicit arguments in text. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, 54–62. Portland, Oregon, USA.

Toutanova, Kristina, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 173–180. Association for Computational Linguistics.

Turan, Ümit Deniz. 1995. Null vs. overt subjects in Turkish: A centering approach. Ph.D. thesis, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

Uryupina, Olga. 2003. High-precision identification of discourse new and unique noun phrases. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, 80–86. Sapporo, Japan: Association for Computational Linguistics.

Versley, Yannick. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC 2010)*, 83–92. Tartu, Estonia: Northern European Association for Language Technology.

Viethen, Jette, and Robert Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation,* Sidney, Australia, 15–16 July 2006, 63–70.

Walker, Marilyn A., Masayo Iida, and Sharon Cote. 1990. Centering in Japanese discourse. In *Proceedings of the 13th International Conference on Computational Linguistics,* Helsinki, Finland, 20–25 August 1990. Appendix, 6pp.

Wan, Stephen, Mark Dras, Robert Dale, and Cecile Paris. 2006. Using dependency-based features to take the "Para-farce" out of paraphrase. In *Proceedings of the Australasian Language Technology Workshop*, 131–138.

Wang, Rui, and Chris Callison-Burch. 2011. Paraphrase fragment extraction from monolingual comparable corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, 52–60. Portland, Oregon: Association for Computational Linguistics.

Whittemore, Greg, Melissa Macpherson, and Greg Carlson. 1991. Event-building through role-filling and anaphora resolution. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 17–24. Berkeley, California, USA: Association for Computational Linguistics.

Wolfe, Travis, Benjamin Van Durme, Mark Dredze, Nicholas Andrews, Charley Beller, Chris Callison-Burch, Jay DeYoung, Justin Snyder, Jonathan Weese, Tan Xu, and Xuchen Yao. 2013. PARMA: A predicate argument aligner. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 63–68. Sofia, Bulgaria: Association for Computational Linguistics.

Woodsend, Kristian, and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 233–243. Jeju Island, Korea: Association for Computational Linguistics.

Wubben, Sander, Antal van den Bosch, Emiel Krahmer, and Erwin Marsi. 2009. Clustering and matching headlines for automatic paraphrase acquisition. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, 122–125. Athens, Greece: Association for Computational Linguistics.

Yarowsky, David, and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the 2nd Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 200–207. Pittsburgh, Pennsylvania, USA: Association for Computational Linguistics.

Yeh, Alexander. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics*, 947–953. Saarbrücken, Germany.

Zarrieß, Sina, and Jonas Kuhn. 2013. Combining referring expression generation and surface realization: A corpus-based investigation of architectures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1547–1557. Sofia, Bulgaria: Association for Computational Linguistics.

Zhao, Hai, Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Multilingual dependency learning: Exploiting rich features for tagging syntactic and semantic dependencies. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, 61–66. Boulder, Colorado, USA.