

Deep Learning With Sentiment Inference For Discourse-Oriented Opinion Analysis



Ana Marasović

Department of Computational Linguistics
Heidelberg University

This dissertation is submitted for the degree of
Doctor of Philosophy

Supervisor: Prof. Dr. Anette Frank
Second supervisor: Prof. Dr. Michael Strube
External examiner: Prof. Dr. Bonnie Webber

Submission date: 30.11.2018.

Acknowledgements

I would first like to thank my supervisor Anette Frank for taking me as her student although I had almost no experience in natural language processing. Since then, many doors have opened for me and I will always be grateful to Anette for it. Her valuable and regular guidance throughout my stay at Heidelberg University has shaped my thoughts on language, research, and life in general. Thank you, Anette, for also having patience and understanding for my numerous special requests.

I would like to thank my second supervisor Michael Strube and external examiner Bonnie Webber for proofreading this Ph.D. thesis. Special thanks to Jeff Kuntz for finding grammatical, punctuation, and spelling mistakes. I also thank Maria Becker, Bhushan Kotnis, and Todor Mihaylov for providing their feedback on a draft of this thesis.

I would like to thank everyone I have had the pleasure to meet over the past three years at the Department of Computational Linguistics at Heidelberg University, HITS, and TU Darmstadt as well as those whom I have encountered at conferences, summer schools, and during my internship in Amazon. I would like to give special thanks to Todor Mihaylov for helping me to deal with TensorFlow, encouraging me to apply to internships and future jobs, sharing his thoughts on many topics, and last but not least, teaching me how to juggle. Sincere thanks go to student researchers Leo Born, Juri Opitz, and Mengfei Zhou whose help in processing data enabled me to progress faster in my research. I also thank the researchers who gave excellent tutorials in Heidelberg and Darmstadt over the past three years. A special source of inspiration was the tutorial by Guillaume Bouchard and invited talks by Claire Cardie, Eduard Hovy, Sebastian Riedel, Dan Roth, and Bonnie Webber.

My three years stay in Heidelberg has been a wonderful journey. Sincere thanks go to all my flatmates for making my time in Heidelberg so enjoyable.

Finally, I could write as many words as there are in this thesis to thank my parents Davor and Silvana, my sister Andrea, and to my life companions Maša and Petar. Thank you for staying by my side.

Abstract

Opinions are omnipresent in written and spoken text ranging from editorials, reviews, blogs, guides, and informal conversations to written and broadcast news. However, past research in NLP has mainly addressed explicit opinion expressions, ignoring implicit opinions. As a result, research in opinion analysis has plateaued at a somewhat superficial level, providing methods that only recognize what is explicitly said and do not understand what is implied.

In this dissertation, we develop machine learning models for two tasks that presumably support propagation of sentiment in discourse, beyond one sentence. The first task we address is opinion role labeling, i.e. the task of detecting who expressed a given attitude toward what or who. The second task is abstract anaphora resolution, i.e. the task of finding a (typically) non-nominal antecedent of pronouns and noun phrases that refer to abstract objects like facts, events, actions, or situations in the preceding discourse.

We propose a neural model for labeling of opinion holders and targets and circumvent the problems that arise from the limited labeled data. In particular, we extend the baseline model with different multi-task learning frameworks. We obtain clear performance improvements using semantic role labeling as the auxiliary task. We conduct a thorough analysis to demonstrate how multi-task learning helps, what has been solved for the task, and what is next. We show that future developments should improve the ability of the models to capture long-range dependencies and consider other auxiliary tasks such as dependency parsing or recognizing textual entailment. We emphasize that future improvements can be measured more reliably if opinion expressions with missing roles are curated and if the evaluation considers all mentions in opinion role coreference chains as well as discontinuous roles.

To the best of our knowledge, we propose the first abstract anaphora resolution model that handles the unrestricted phenomenon in a realistic setting.

We cast abstract anaphora resolution as the task of learning attributes of the relation that holds between the sentence with the abstract anaphor and its antecedent. We propose a Mention-Ranking siamese-LSTM model (MR-LSTM) for learning what characterizes the mentioned relation in a data-driven fashion. The current resources for abstract anaphora resolution are quite limited. However, we can train our models without conventional data for abstract anaphora resolution. In particular, we can train our models on many instances of

antecedent-anaphoric sentence pairs. Such pairs can be automatically extracted from parsed corpora by searching for a common construction which consists of a verb with an embedded sentence (complement or adverbial), applying a simple transformation that replaces the embedded sentence with an abstract anaphor, and using the cut-off embedded sentence as the antecedent. We refer to the extracted data as silver data.

We evaluate our MR-LSTM models in a realistic task setup in which models need to rank embedded sentences and verb phrases from the sentence with the anaphor as well as a few preceding sentences. We report the first benchmark results on an abstract anaphora subset of the ARRAU corpus (Uryupina et al., 2016) which presents a greater challenge due to a mixture of nominal and pronominal anaphors as well as a greater range of confounders. We also use two additional evaluation datasets: a subset of the CoNLL-12 shared task dataset (Pradhan et al., 2012) and a subset of the ASN corpus (Kolhatkar et al., 2013a). We show that our MR-LSTM models outperform the baselines in all evaluation datasets, except for events in the CoNLL-12 dataset. We conclude that training on the small-scale gold data works well if we encounter the same type of anaphors at the evaluation time. However, the gold training data contains only six shell nouns and events and thus resolution of anaphors in the ARRAU corpus that covers a variety of anaphor types benefits from the silver data. Our MR-LSTM models for resolution of abstract anaphors outperform the prior work for shell noun resolution (Kolhatkar et al., 2013b) in their restricted task setup. Finally, we try to get the best out of the gold and silver training data by mixing them. Moreover, we speculate that we could improve the training on a mixture if we: (i) handle artifacts in the silver data with adversarial training and (ii) use multi-task learning to enable our models to make ranking decisions dependent on the type of anaphor. These proposals give us mixed results and hence a robust mixed training strategy remains a challenge.

Table of Contents

1	Introduction	1
1.1	Analyzing Opinions in Discourse	1
1.2	Fine-Grained Opinion Analysis	2
1.3	Abstract Anaphora Resolution	4
1.4	Research Questions	7
1.5	Contributions	7
1.6	Thesis Overview	8
1.7	Published Work	9
2	Background	11
2.1	Sentiment Inference	11
2.2	Fine-Grained Opinion Analysis	25
2.3	Abstract Anaphora Resolution	34
2.4	Neural Multi-Task Learning and Adversarial Training	47
3	Sentence-Level Neural Fine-Grained Opinion Analysis	57
3.1	Similarities and Divergences in the Opinion and Semantic Role Labeling Annotation Schema	58
3.2	Neural MTL for SRL and ORL	60
3.3	Experimental Setup	61
3.4	Experiments	64
3.5	Analysis of What Works and What is Next	66
3.6	Summary	73
4	Resolving Abstract Anaphors in a Relational Neural Model	75
4.1	Challenges and Working Toward the Models	76
4.2	MR-LSTM: Mention-Ranking LSTM-Siamese Neural Network	78
4.3	Training Data Extraction	81

4.4	Experimental Setup	84
4.5	Performance of the MR-LSTM Model	91
4.6	MR-LSTM with Adversarial Training	115
4.7	MR-LSTM with Multi-Task Learning	120
4.8	Summary	128
5	Outlook and Conclusions	131
5.1	Contributions and Takeaways	131
5.2	Discussion	134
5.3	Direction for Future Work	135
A	MPQA Pre-processing	139
B	Comparison between MR-LSTM and the Prior Work in Shell Noun Resolution	143
C	Quality Evaluation of the VC-SS-Extract Method	145
D	Data Management	181
	List of Figures	183
	List of Tables	185
	List of Abbreviations	193
	References	

Chapter 1

Introduction

1.1 Analyzing Opinions in Discourse

Modern NLP techniques perform well at tasks such as dependency parsing (Dozat and Manning, 2017), named entity recognition (Akbik et al., 2018), and semantic role labeling (He et al., 2017); but tasks like making conversation (Mrkšić et al., 2017), fact checking (Thorne and Vlachos, 2018), summarization (Chen and Bansal, 2018), or understanding opinionated text (Katiyar and Cardie, 2016) remain a challenge. These tasks require abstraction, cognition, reasoning, and knowledge about our world. In other words, it is not possible to solve these problems as long as we do not build systems that recognize both *what is said* and *what is not said but is implied*. Consider the following text snippet.¹

- (1) Registrar General Tobaiwa Mudede announced on state television that Mugabe was re-elected with 1,685,212 votes against 1,258,758 votes for Tsvangirai, leader of the Movement for Democratic Change (MDC). [...] "The election was massively rigged", he told a packed press conference. "We therefore as MDC do not accept this result".

Applying state-of-the-art opinion analysis systems (Yang and Cardie, 2013, 2014a) on the final sentence: *We therefore as MDC do not accept this result*, would inform us that *MDC* explicitly expressed a negative attitude toward the *result* by not accepting it. A human reader can easily infer that the *result* refers to the fact that *Mugabe was re-elected with 1,685,212 votes against 1,258,758 votes for Tsvangirai* and understand what is implied: if *MDC* is negative toward the *re-election* then it is likely that *MDC* is negative toward the person who was re-elected, *Mugabe*, and positive toward the person who was not elected,

¹Examples (1–2) are taken from the Multi-Perspective Question Answering Corpus (MPQA) dataset (Wiebe et al., 2005).

Tsvangirai. This inference step—enabled once we realized what the *result* is—gives us a deeper understanding of opinionated text since we also observe implicit attitudes which are a distinctive aspect of subjective language.²

Deng et al. (2013a, 2014) and Deng and Wiebe (2014, 2015a) proposed the first computational approaches to sentiment inference. They focus on sentiment implicatures that arise in the presence of explicit sentiments and events that positively (GOODFOR) or negatively (BADFOR) affect entities. In Example (1), *re-elected* is a GOODFOR event as it benefits the theme—the person who is re-elected (*Mugabe*). Their systems are designed to utilize the fact that *MDC* is explicitly negative toward this event and propagate the negative attitude to *Mugabe*—the theme who benefits from the event *MDC* is negative about.

However, computationally modeling this inference step is difficult as it involves an interplay of some challenging subtasks:

- (i) Detecting who *explicitly* expressed what kind of attitude toward what or whom (e.g., that *MDC* is negative about the *result*).
- (ii) Finding a (typically) non-nominal antecedent of pronouns and noun phrases that refer to abstract objects like facts, events, actions, or situations in the preceding discourse (e.g., detecting that *this result* refers to the fact that *Mugabe was re-elected with 1,685,212 votes against 1,258,758 votes for Tsvangirai*).
- (iii) Resolving noun phrases referring to concrete objects or entities in the real world (e.g., realizing that *we* refers to *MDC*).
- (iv) Propagating sentiment in discourse, beyond one sentence (e.g., inferring that *MDC* is negative toward *Mugabe* and positive toward *Tsvangirai*).

In the research presented in this thesis we focus on the first and second subtasks. We aim to develop machine learning models for them and hence potentially support propagation of sentiment in discourse beyond one sentence. In the remainder of this chapter, we further motivate the research conducted in the thesis (Sections 1.2–1.3), formulate main research questions (Section 1.4), summarize our contributions (Section 1.5), present the outline of the thesis (Section 1.6), and describe which parts of the thesis were published (Section 1.7).

1.2 Fine-Grained Opinion Analysis

Sentiment inference systems of Deng and Wiebe (2014, 2015a) and Deng et al. (2014) require information about who explicitly expressed what kind of sentiment toward what or whom.

²We use attitude, sentiment, and opinion interchangeably.

Significant research efforts have been invested on document- and sentence-level subjectivity detection and polarity classification (Pang and Lee, 2008; Liu, 2015). In contrast, there has been less work on other tasks in opinion analysis such as detecting entities that express an attitude (*holders*) and objects that the attitude is directed towards (*targets*). These tasks are vital for various applications such as opinion summarization and sentiment inference. Furthermore, it is often even impossible to determine the document- and sentence-level polarity³ since a given text may contain opposing views. For example, in (2) the *team* is positive toward the *presidential election*, but the *West* expressed negative attitude toward the claim that *the presidential election was substantially free and fair*.

- (2) The team met President Mugabe at Zimbabwe House where it briefed him of its opinion that the presidential election was substantially free and fair, despite a view to the contrary by the West.

In contrary to determining overall polarity of text, Fine-Grained Opinion Analysis (FGOA) aims to: (i) detect opinion expressions that convey attitudes such as sentiments, agreements, beliefs, or intentions (like *do not accept* in Example (1) on page 1), (ii) measure their intensity (e.g., strong), (iii) identify their holders i.e. entities that express an attitude (e.g., *we* or *MDC*), (iv) identify their targets i.e. entities or propositions at which the attitude is directed (e.g., *this result*), and (v) classify their target-dependent attitude (e.g., negative sentiment). Hovy (2011) argues that all subtasks (i–v) need to be addressed to properly define and understand opinions. In his own words,

An opinion is a decision made by someone (the holder) about a topic⁴. This decision assigns the topic to one of a small number of classes (the valences)⁵ that affect the role that the topic will play in the holder’s future goals and planning decisions.

Although FGOA is crucial for understanding opinionated text, the state-of-the-art Conditional Random Field (CRF) model (Yang and Cardie, 2013) and the neural competitor (Katiyar and Cardie, 2016) achieve about 55% F1 score for predicting which targets relate to which opinions in the benchmark corpus MPQA (Wiebe et al., 2005).⁶ Thus, these models are not yet ready to answer the question this line of research is generally motivated with:

³At least in the standard categories: positive, negative, and neutral.

⁴That is, a target.

⁵Hovy (2011) defines two kinds of opinions: judgment and belief. The valance of judgment opinions can be positive, negative, mixed, neutral, or unstated. The valance of belief opinions can be believed, disbelieved, unsure, neutral, or unstated. In practice, a much broader valance set is used: positive, negative, neutral.

⁶To be precise, they address (i), (iii), and (iv), which is still generally considered as the fine-grained opinion analysis task.

"Who expressed what kind of sentiment toward what?", and consequently assist applications such as sentiment inference.

We recognize benefits of the neural FGOA model (Katiyar and Cardie, 2016), despite the fact that it still lags behind the feature-based CRF (Yang and Cardie, 2013) for labeling of opinion holders and targets, i.e. for Opinion Role Labeling (ORL). In particular because neural models are adaptive to heterogeneous sources since they do not depend on external resources such as dependency parsers, named entity recognizers, sentiment lexicons, and other resources that are usually available only for English and for certain domains.

In this thesis we aim to investigate the limitations of neural models in solving ORL and to gain a better understanding of what is solved and what is next. We speculate that scarcity of labeled data is a major obstacle for neural ORL models and address this problem using Multi-Task Learning (MTL) with appropriate auxiliary tasks. A promising auxiliary task candidate for ORL is Semantic Role Labeling (SRL), the task of predicting predicate-argument structure of a sentence, which answers the question: "Who did what to whom, where and when?". However, obstacles for properly exploiting SRL training data with MTL could be specificities, inconsistencies, and the incompleteness of the MPQA annotations (see the discussion on challenges of the MPQA corpus in more detail in Chapter 3). This observation leads to the first main research question: **can we improve neural opinion role labeling models by using MTL with a related task which has substantially more data, i.e., SRL, even though there are divergences in the annotation schemes of opinion and semantic role labeling in the benchmark corpora?** We address this research question by adopting one of the recent successful architectures for SRL (Zhou and Xu, 2015) and by applying different MTL schemes.

1.3 Abstract Anaphora Resolution

In addition to FGOA, we need to resolve different types of anaphors to be able to properly utilize proposed sentiment inference systems. For Example (1) on page 1, we need to understand what the *result* is to infer that *MDC* is negative toward *Mugabe* and positive toward *Tsvangirai*.

Current research in anaphora (or coreference) resolution is focused on noun phrases referring to concrete objects or entities in the real world (e.g., *we* and *MDC* which represent the same real-world organization "The Movement for Democratic Change"). Distinct from these are diverse types of abstract anaphora (Asher, 1993) where reference is made to propositions, events, properties, or facts such as that *Mugabe was re-elected with 1,685,212 votes against 1,258,758 votes for Tsvangirai*.

Abstract Anaphora Resolution (AAR) is a difficult task because there is a number of lexical, semantic, and syntactic properties associated with abstract anaphora. These properties are outlined in Chapter 2 (Section 2.3.2). The standard coreference resolution features such as agreement or saliency cannot be applied to AAR. For example, the state-of-the-art neural coreference resolver (Lee et al., 2017) is not designed to capture abstract anaphora and hence is not able to point to what the *result* in (1) refers to.⁷

Furthermore, even humans do not generally agree on the exact boundaries of the antecedent. The reported inter-annotator agreement on exact match for antecedent selection is around Krippendorff’s alpha of 0.55 in the ARRAU corpus (Poesio and Artstein, 2008). The difficulty of annotating data for AAR resulted in limited labeled data for this task.

Since designing a good feature space and annotating data is challenging, automatic resolution of abstract anaphors is still relatively unexplored, although abstract anaphors are frequent across languages (Vieira et al., 2005; Poesio and Artstein, 2008; Dipper and Zinsmeister, 2009). Moreover, it has been shown that their understanding is valuable for computational systems in machine translation (Le Nagard and Koehn, 2010; Hardmeier et al., 2015), summarization (Steinberger et al., 2005; Orăsan, 2007), question answering (Quarteroni, 2007; Vicedo and Ferrández, 2008), and, as we hypothesize in Chapter 2 (Section 2.1), most likely for sentiment inference too. This insight leads us to the second main research question of the thesis: **can we apply computational methods to resolve abstract anaphors automatically?**

We approach this challenging research question following the intuition that we can learn what is the correct antecedent for a given abstract anaphor by learning the relation that holds between the sentence with the anaphor and its antecedent. In Example (1) this relation is something *MDC* may not accept. Since antecedents differ in distance, syntactic type, and other properties, we opt for a neural model that learns relevant features from data and does not force us to make certain assumptions that might be limiting.

However, even if our modeling assumptions are plausible, we still have the problem of the scarcity of labeled training data. Fortunately, we can train our models to learn what characterizes mentioned relations, such as what *MDC* might not accept, if we train them on many instances of antecedent-anaphoric sentence pairs that we are able to automatically extract. We harvest such pairs from parsed corpora in the following way: (i) we search for constructions with a verb with an embedded sentential argument (3a), (ii) apply a simple transformation that replaces the embedded sentence with an abstract anaphor (such as *this* in (3b)), and (iii) use the cut-off embedded sentence as the antecedent (3c).⁸

⁷According to the available demo: <http://demo.allennlp.org/coreference-resolution/NDA2NjAw>.

⁸This is a real example used in experiments.

- (3) a. While few lawmakers [VP anticipated [S' that [S the humanitarian aid would be cut off next month]]], Mr. Ortega's threat practically guarantees that the humanitarian aid will be continued.
- b. **Anaphoric sentence:** While few lawmakers anticipated this, Mr. Ortega's threat practically guarantees that the humanitarian aid will be continued.
- c. **Antecedent:** The humanitarian aid would be cut off next month.

Some follow-up challenges emerge. First, extracted antecedents do not occur in a natural context once they are extracted from their original sentences. For this reason, we cannot use sequence labeling techniques that label each word of a given text as the part of the antecedent or outside of it. Therefore we use ranking models that rank each candidate for the antecedent where candidates are sub-constituents of few preceding sentences and the antecedent. Our models calculate a score for each candidate from a joint representation of the *anaphoric sentence* (the sentence in which the anaphor occurs) and the candidate. That is, we do not use the context of the candidate for the antecedent to calculate its score. Candidates are then ranked using the calculated score.

Moreover, an extracted antecedent does not have a distance from the anaphor. However, information about the distance between the anaphor and the antecedent plays an important role in anaphora resolution systems since it can narrow down the search scope of candidates for antecedents (Mitkov, 2014, p. 17). Therefore, we sample a distance for every extracted antecedent on the basis of a distribution calculated from natural abstract anaphors from a development set.

Next, we make pre-processing decisions to make our harvested examples more similar to natural cases. However, there still might be other more complex properties that cannot be captured with pre-processing decisions. For instance, in (4b) the anaphor does not occur in a natural position in the sentence. A more plausible anaphoric sentence is in (4c). To ensure that our models do not fit such artifacts we propose *adversarial training*. In theory, it should force our models to capture only features which occur in both harvested and natural data.

- (4) a. "The main feature of the new organization [VP is [S' that [S each local manager will have both the authority and accountability for profitable and technically sound operations]]], said Charles E. Spruell, president of the Mobil Unit.
- b. **Anaphoric sentence:** "The main feature of the new organization is this", said Charles E. Spruell, president of the Mobil Unit.
- c. **More natural anaphoric sentence:** "This is the main feature of the new organization", said Charles E. Spruell, president of the Mobil Unit.

- d. **Antecedent:** Each local manager will have both the authority and accountability for profitable and technically sound operations.

Finally, we aim to propose a single model for all types of abstract anaphors: pronouns, noun phrases, shell nouns, or eventualities. However, different features may be relevant for different anaphora types and resolving each type could be considered as a different related task. For this reason, we extend our models with *multi-task learning*.

1.4 Research Questions

In Sections 1.2 and 1.3, we raised two main research questions. Some follow-up questions and research opportunities naturally emerge from them. We organize them in two groups.

Opinion Role Labeling (ORL). In this thesis, we investigate whether exploiting data of a similar task (SRL) through Multi-Task Learning (MTL) helps to improve ORL models. In particular, we wonder whether MTL can overcome divergences in the annotation schemes of opinion and semantic role labeling. We finalize the ORL study by reflecting on what is solved and what is next for ORL.

Resolving abstract anaphors automatically. In this thesis, we raise a series of questions to determine whether computational methods can be used for the automatic resolution of abstract anaphors. First, we wonder is it plausible to formulate the task of finding the correct antecedent for a given abstract anaphor as learning the characteristics of the relation between the sentence with a given abstract anaphor and its antecedent? Second, can such characteristics be captured with an LSTM-Siamese neural model that: (i) ranks candidates for the antecedent on the basis of joint representations of the sentence that contains a given abstract anaphor and candidates for the antecedent, (ii) is trained with pairs of antecedents and anaphoric sentences? Can we reliably extract such pairs from parsed corpora? Next, can our models make use of sampled distances for extracted antecedents to narrow down the search scope of candidates for antecedents? Finally, can adversarial training filter artifacts that occur in extracted training data? Does multi-task learning give our models power to make predictions dependent on the anaphor type?

1.5 Contributions

The main contributions presented in the thesis are summarized below.

Sentiment inference. We develop novel models for tasks that presumably support propagation of sentiment in discourse: Opinion Role Labeling (ORL) and Abstract Anaphora Resolution (AAR).

Opinion Role Labeling (ORL). We exploit SRL data using different Multi-Task Learning (MTL) techniques and obtain significant improvements for ORL. Our improvements with MTL suggest that the scarcity of labeled training data is an obstacle for neural ORL models trained on the benchmark corpus MPQA. We propose a new experimental setup that is more suitable for the evaluation of neural models if we cannot afford to train our models many times. We conduct a thorough analysis that illustrates what is solved and what is next.

Abstract Anaphora Resolution (AAR). We approach AAR as learning characteristics of the relation between the sentence with the given abstract anaphor and its antecedent. We propose a stable baseline neural model for the core phenomenon that is able to find the abstract antecedent from a wider context. We learn how to reliably extract and integrate training data. We test adversarial training for addressing differences between harvested and natural data and multi-task learning for differences between anaphora types.

1.6 Thesis Overview

The remainder of the thesis is organized as follows.

Chapter 2 reviews background of sentiment inference, fine-grained opinion analysis, abstract anaphora resolution, and multi-task learning. In this thesis, we do not propose new models for sentiment inference. However, in Section 2.1, we thoroughly analyze the most prominent approaches to sentiment inference to illustrate (i) the reliance of the available sentiment inference models on existing approaches for detecting explicit opinion expressions and their roles, and (ii) how they could benefit from anaphora resolution. In Section 2.2 we focus on extraction and categorization of opinion expressions and their roles in news articles. Since an extensive overview of the literature related to the phenomenon of abstract anaphora has been written by Kolhatkar et al. (2018), the central topic of Section 2.3 is to establish (i) terminology related to the phenomenon of abstract anaphora and (ii) to investigate the reasons why the resolution of abstract anaphors is still relatively unexplored. Finally, Section 2.4 covers the basics of most deep learning approaches for representation learning of text and the well-received MTL techniques. Reading this section is necessary to follow the descriptions of our models in following chapters.

Chapter 3 describes how specificities, inconsistency, and the incompleteness of the MPQA annotations could be an obstacle for properly exploiting SRL data with MTL. In this chapter, we discuss how well-received MTL techniques could adapt to different cases of opinion role labeling. We then evaluate these MTL approaches and analyze what has been solved and what is next for the task.

Chapter 4 investigates our models for abstract anaphora resolution. We first reflect on the observed challenges from the related work, note other issues that emerge, and propose ways to address them. We then describe our ranking neural models and the training data extraction method in detail. We provide a thorough evaluation of our models.

Chapter 5 summarizes the findings of this thesis and discusses potential future directions of research.

1.7 Published Work

The majority of the research presented in this thesis is an extension of published research first-authored by the author of this thesis.

The majority of material presented in Chapter 3 is described in Marasović and Frank (2018). Chapter 4 is an extension of Marasović et al. (2017). In the paper, we present our core model and a method for extracting training data. We evaluated the model following the experimental setup of the prior work on shell noun resolution (Kolhatkar et al., 2013b). Their models rank all sub-constituents (even unlikely noun phrases) of the sentence that contains the antecedent. This is a restricted experimental setup compared to a more realistic setting where a system needs to find the antecedent in at least a few preceding sentences. In the restricted setup our models outperform the state of the art (Kolhatkar et al., 2013b) in shell noun resolution (see Appendix B). In the thesis, we evaluate our models in a realistic experimental setup in which models need to rank (embedded) sentences and verb phrases from the sentence with the anaphor as well as a few preceding sentences.

We addressed modal sense classification in Marasović et al. (2016) and Marasović and Frank (2016). This was the first step to identify senses of verbs, explore methods to automatically harvest training data, and to address further aspects of sentiment inference that relate to modal sense classification. In the initial phases, we investigated different deep learning approaches to learning textual representations. In this thesis, we consistently use recurrent neural networks but we also explored convolutional neural networks in Marasović and Frank (2016). Finally, in joint work with Zopf et al. (2018) we showed that sentiment annotations can serve as useful features for the notion of "information importance" in text summarization.

The code for processing the data as well as training and evaluating models presented in this thesis are listed in Appendix D.

Chapter 2

Background

This chapter reviews background relevant to the research questions we address in this thesis. We start by surveying the most prominent sentiment inference approaches in Section 2.1. The aim of this section is to demonstrate the importance of the proper detection of *opinion roles* (i.e. holders and targets) for the proposed sentiment inference approaches and how the resolution of anaphors can benefit them. We then survey and discuss approaches relevant for detection of opinion roles (Section 2.2). Next, we consider work on the resolution of abstract anaphors (Section 2.3). Finally, we review work on neural multi-task learning and adversarial training techniques because they are shown to be a powerful tool for addressing problems that arise in the presence of limited labeled data (Section 2.4).

2.1 Sentiment Inference

2.1.1 Overview of Deng and Wiebe’s Work

Opinions are omnipresent in written and spoken text ranging from editorials, reviews, blogs, guides, and informal conversations to written and broadcast news. As a result there is a large amount of research on automatic identification and characterization of opinions in text (see Section 2.2). However, past research in NLP has mainly addressed explicit opinion expressions, ignoring implicit opinions. As a result, research in opinion analysis has plateaued at a somewhat superficial level, providing methods that exhibit a fairly shallow understanding of subjective language.

Deng et al. (2013a, 2014) and Deng and Wiebe (2014, 2015a) are among the first to develop systems which detect both explicit and implicit sentiments expressed among entities

and events in the text. They focus on sentiment *implicatures*¹ that arise in the presence of explicit sentiments and events that positively or negatively affect entities. They refer to these events as BENEFACTIVE/MALEFACTIVE (Deng et al., 2013a), GOODFOR/BADFOR (Deng and Wiebe, 2014), or +EFFECT/-EFFECT events (Deng and Wiebe, 2015a). We use the GOODFOR/BADFOR notation.

Deng and Wiebe (2014) showcase two examples which illustrate how sentiment implicature rules may be utilized to propagate explicitly expressed sentiment to other entities.

- (5) a. The bill would lower health care costs, which would be a tremendous positive
agent BADFOR theme explicit sentiment
change across the entire health-care system.
- b. $\text{POSITIVEPAIR}(\text{writer}, \text{agent}) \wedge \text{BADFOR}(\text{event}) \Rightarrow \text{POSITIVEPAIR}(\text{writer}, \text{event}) \wedge \text{BADFOR}(\text{event}) \Rightarrow \text{NEGATIVEPAIR}(\text{writer}, \text{theme})$
- (6) a. The bill would curb skyrocketing health care costs.
agent BADFOR explicit sentiment theme
- b. $\text{NEGATIVEPAIR}(\text{writer}, \text{theme}) \wedge \text{BADFOR}(\text{event}) \Rightarrow \text{POSITIVEPAIR}(\text{writer}, \text{event}) \Rightarrow \text{POSITIVEPAIR}(\text{writer}, \text{agent})$

Since the writer describes the *bill* as a *tremendous positive change* in (5a), the writer is positive toward the *bill*. Given that the *bill* is also the agent of the event *lower* which negatively effects its theme *costs*, it may be concluded that the writer is negative toward the *costs*. The sentiment implicature rule supporting this inference is given in (5b).

Similarly, in Example (6a), the writer is explicitly negative toward the *costs* because the writer describes them as *skyrocketing*. Since the writer is negative toward the theme, it may be concluded that the writer is positive toward the event that negatively effects it, i.e. *curb*. Moreover, if the writer is positive toward the event, it is likely that the writer is positive toward the agent of the event, i.e. toward the *bill*. The sentiment implicature rule supporting this inference is given in (6b).

In their work, Deng and Wiebe represent and utilize sentiment implicature rules in three different kinds of learning models: a graph-based model with Loopy Belief Propagation (LBP) algorithm (Deng and Wiebe, 2014), an Integer Linear Programming (ILP) framework (Deng et al., 2014), and a Probabilistic Soft Logic (PSL) model (Deng and Wiebe, 2015a).

In the rest of this section we describe details of these models to illustrate (i) the reliance of the available sentiment inference models on existing approaches for detecting explicit opinion expressions and their roles, and (ii) how could they benefit from anaphora resolution.

¹*Implicature* denotes either (i) the act of meaning or implying one thing by saying something else, (ii) or the object of that act (Davis, 2014).

The summary at the end of this section recaps the most important aspects of Deng and Wiebe’s work pertaining to the research questions addressed in this thesis.

2.1.1.1 Data for Sentiment Inference

GOODFOR/BADFOR Corpus Deng and Wiebe (2014) and Deng et al. (2014) train and evaluate their model with the GOODFOR/BADFOR corpus (Deng et al., 2013a)² which consists of 134 opinionated documents (blogs and editorials) about a controversial topic: the Affordable Care Act. In the data, GOODFOR/BADFOR triples are annotated with the spans of the GOODFOR/BADFOR event, its agent, and its theme, as well as the effect class of the event (GOODFOR or BADFOR) and the writer’s attitude toward the agent and theme (positive, negative, or neutral). However, there are many false negatives of sentiments toward other entities that do not participate in GOODFOR/BADFOR relations and sentiments from entities other than the writer. Hence, the corpus does not support the training of a model that detects explicit sentiment. The effect of a GOODFOR event may be changed to BADFOR by a reverser (and the other way around). For example, in *The reform will not worsen the economy*, *not* is a reverser and it reverses the effect from BADFOR to GOODFOR. In contrast, a retainer is a word whose effect is to retain the effect of a GOODFOR/BADFOR event. In total, there are 1,762 annotated GOODFOR/BADFOR triples, out of which 692 are GOODFOR (or retainers) and 1,070 are BADFOR (or reversers). From the writer’s perspective, 1,495 noun phrases are annotated positive, 1,114 are negative, and the remaining 8 are neutral.

MPQA 3.0 Deng and Wiebe (2015a) train and evaluate their models using the MPQA 3.0 dataset which we describe in more detail in Section 2.2.1. Currently, there are 70 documents, 1,634 sentences, and 1,921 opinion expressions³ in total.

2.1.1.2 Graph-Based Sentiment Inference Model

Deng and Wiebe (2014) proposed a graph-based model of entities and the GOODFOR/BADFOR relations between them to enable sentiment propagation from entities associated with explicit sentiment to entities associated with implicit sentiment. In particular, nodes in a GOODFOR/BADFOR entity graph are noun phrase agents or theme spans which are linked with an edge if they co-occur in at least one ⟨agent, GOODFOR/BADFOR event, theme⟩ triplet. Finally, Loopy Belief Propagation (LBP) (Pearl, 1982; Yedidia et al., 2005) is applied

²<http://mpqa.cs.pitt.edu/corpora/gfbf/>

³Here opinion expressions can also be so-called expressive subjective elements such as *full of absurdities* in *"The report is full of absurdities," Xirao-Nima said*. See Section 2.2.1 for a detailed explanation.

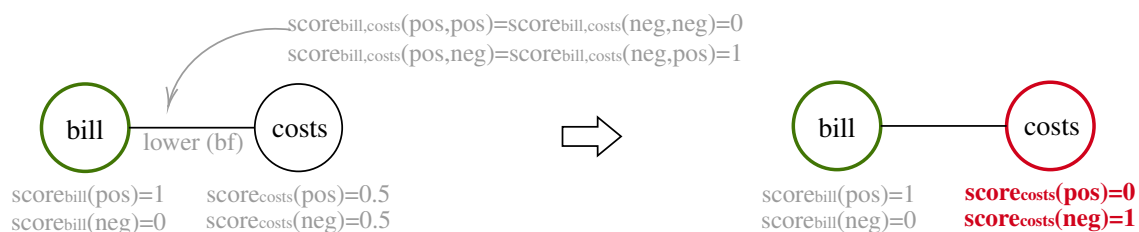


Fig. 2.1 The GOODFOR/BADFOR graph of Example (5a): *The bill would lower health care costs, which would be a tremendous positive change across the entire health-care system.*

to accomplish sentiment propagation to nodes without explicit sentiment. Fig. 2.1 illustrates the GOODFOR/BADFOR graph of Example (5a) on page 12.

Dependence on existing opinion analysis approaches. In their LBP formulation each node is associated with two scores, and each edge with four scores. A score of the positive label $s_i(\text{positive})$ of a node n_i is 1 if the writer explicitly expressed positive sentiment to the corresponding agent or theme, and 0 otherwise. A similar process applies for a score of the negative label $s_i(\text{negative})$. Since the GOODFOR/BADFOR corpus does not support training a model that detects explicit sentiments (see Section 2.1.1.1), they are detected using available resources. In particular, Deng and Wiebe (2014) use two opinion analysis systems: the OpinionFinder system⁴ and the system of Johansson and Moschitti (2013), and three lexica: OpinionFinder (Wilson et al., 2005), General Inquirer (Stone et al., 1966), and the connotation lexicon (Feng et al., 2013). Two systems and three lexica make for five votes. Before explicit sentiment classification, each node n_i has the positive value $s_i(\text{positive})$ of 0.5 and the negative value $s_i(\text{negative})$ of 0.5. The number of positive votes multiplied by 0.1 is added to the score of positive label $s_i(\text{positive}) = 0.5$ and the number of negative votes multiplied by 0.1 is added to the score of negative value $s_i(\text{negative}) = 0.5$. If the positive value is larger, the positive value is maintained, and the negative value is assigned to be $s_i(\text{negative}) = 1 - s_i(\text{positive})$ (similarly if the negative value is larger). The scores before the LBP inference are illustrated in Fig. 2.1 in light grey. The polarity of an opinion is assigned to each word that occurs in the *mod* or *obj* dependency relation with the opinion expression.⁵ For example, the word *skyrocketing* in (6a) on page 12 is negative and because it occurs in the *mod* dependency relation with the noun *costs*, negative sentiment is also assigned to *costs*. Finally, they assume that the writer’s sentiments toward the GOODFOR/BADFOR events in the clauses of the given sentence, the previous sentence, and the next sentence are the same and refer to this assumption as the *discourse heuristic*.

⁴http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_2/

⁵A system that identifies the opinion target was not available at the time.

Ruleset 1	SENTIMENT(GoodFor/BadFor event)	⇒	SENTIMENT(theme)
	POSPAIR(writer,gf event)		POSPAIR(writer,theme)
	NEGPAIR(writer,gf event)	⇒	NEGPAIR(writer,theme)
	POSPAIR(writer,bf event)		NEGPAIR(writer,theme)
	NEGPAIR(writer,bf event)		POSPAIR(writer,theme)
Ruleset 2	SENTIMENT(theme)	⇒	SENTIMENT(GoodFor/BadFor event)
	POSPAIR(writer,theme)		POSPAIR(writer,gf event)
	NEGPAIR(writer,theme)	⇒	NEGPAIR(writer,gf event)
	POSPAIR(writer,theme)		NEGPAIR(writer,bf event)
	NEGPAIR(writer,theme)		POSPAIR(writer,bf event)
Ruleset 3	SENTIMENT(GoodFor/BadFor event)	⇒	SENTIMENT(agent)
	POSPAIR(writer,gf event)		POSPAIR(writer,agent)
	NEGPAIR(writer,gf event)	⇒	NEGPAIR(writer,agent)
	POSPAIR(writer,bf event)		POSPAIR(writer,agent)
	NEGPAIR(writer,bf event)		NEGPAIR(writer,agent)
Ruleset 4	SENTIMENT(agent)	⇒	SENTIMENT(GoodFor/BadFor event)
	POSPAIR(writer,theme)		POSPAIR(writer,gf event)
	NEGPAIR(writer,theme)	⇒	NEGPAIR(writer,gf event)
	POSPAIR(writer,theme)		POSPAIR(writer,bf event)
	NEGPAIR(writer,theme)		NEGPAIR(writer,bf event)

Table 2.1 Sentiment implicature rulesets of Deng and Wiebe (2014). Sentiment in the first column is explicitly stated in text and implies sentiment in the second column. POSPAIR(writer,theme/event) denotes that the writer of a given document is positive toward theme of an event or event itself (and similarly for NEGPAIR).

The conducted error analysis shows that 21.32% of errors come from incorrectly detected explicit sentiments, and 31.89% of the errors are due to nodes not assigned polarity, but given incorrect values because their subgraph has an incorrect polarity, and 4.62% error come from the discourse heuristic. These results confirm that we are in need of better explicit sentiment analyzers and a better understanding of sentiment in discourse.

Representation and integration of sentiment implicature rules Scores $s_{ij}(y_k, y_l)$, $y_k, y_l \in \{positive, negative\}$ of an edge between two nodes n_i and n_j represent the likelihood that the node n_i has the polarity y_k and n_j the polarity y_l . From the sentiment implicature rules (Table 2.1) it may be noticed that regardless of the writer’s sentiment toward the event, if the event is GOODFOR, then the writer’s sentiment toward the agent and theme are the same, while if the event is BADFOR, the writer’s sentiment toward the agent and theme are opposite. Deng and Wiebe (2014) refer to this observation as a *sentiment constraint* and they decide that $s_{ij}(positive, positive) = s_{ij}(negative, negative) = 1$ if two nodes

```

Data: The GOODFOR/BADFOR corpus (Deng et al., 2013a).
1 initialize all  $m_{i \rightarrow j}(pos) = m_{i \rightarrow j}(neg) = 1$  while all  $m_{i \rightarrow j}$  stop changing do
2   foreach  $n_i \in N$  do
3     foreach  $n_j \in Neighbor(n_i)$  do
4       foreach  $y \in \{pos, neg\}$  do
5         calculate  $m_{i \rightarrow j}(y)$ 
6       end
7     normalize  $m_{i \rightarrow j}(pos) + m_{i \rightarrow j}(neg) = 1$ 
8   end
9 end
10 end
11 foreach  $n_i \in N$  do
12    $\operatorname{argmax}_{y \in \{pos, neg\}} \left( s_i(y) \cdot \prod_{n_k \in Neighbor(n_i)} m_{k \rightarrow i}(y) \right)$ 
13   neutral, in case of a tie
14 end

```

Algorithm 1: Sentiment Inference via Loopy Belief Propagation.

are linked by a GOODFOR edge and zero otherwise. Likewise, $s_{ij}(positive, negative) = s_{ij}(negative, positive) = 1$ if two nodes are linked by a BADFOR edge, and zero otherwise. It is important to note that this is the way rules are integrated into the system; the values of $s_{ij}(y_k, y_l)$ are based on sentiment implicature rules (Table 2.1) and sentiment is propagated with regard to the scores.

Finally, LBP is an iterative message passing algorithm and propagation of sentiment works as follows. A message from n_i to n_j over the edge between them has two values, (i) $m_{i \rightarrow j}(pos)$, how much information from node n_i indicates node n_j is positive, and (ii) $m_{i \rightarrow j}(neg)$, how much information from node n_i indicates node n_j is negative. To calculate $m_{i \rightarrow j}(pos)$ they use the score that node n_i is positive itself, the likelihood that the node n_i is positive and n_j is positive, and the positive message n_i 's neighbors (besides n_j) carry to it (Equation 2.1) (and similarly for $m_{i \rightarrow j}(neg)$). The full procedure is shown in Algorithm 1.

$$\begin{aligned}
 m_{i \rightarrow j}(pos) = & s_{ij}(pos, pos) \cdot s_i(pos) \cdot \prod_{n_k \in Neighbor(n_i)/n_j} m_{k \rightarrow i}(pos) + \\
 & s_{ij}(neg, pos) \cdot s_i(neg) \cdot \prod_{n_k \in Neighbor(n_i)/n_j} m_{k \rightarrow i}(neg)
 \end{aligned} \tag{2.1}$$

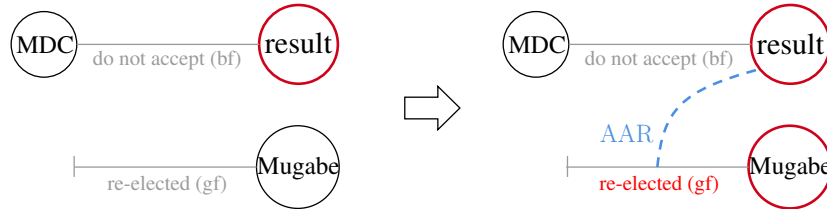
Sentence: We therefore as MDC do not accept this result.

Opinion holder: MDC

Opinion target: result

Abstract anaphor: result

Antecedent: Mugabe was re-elected with 1,685,212 votes against 1,258,758 votes for Tsvangirai



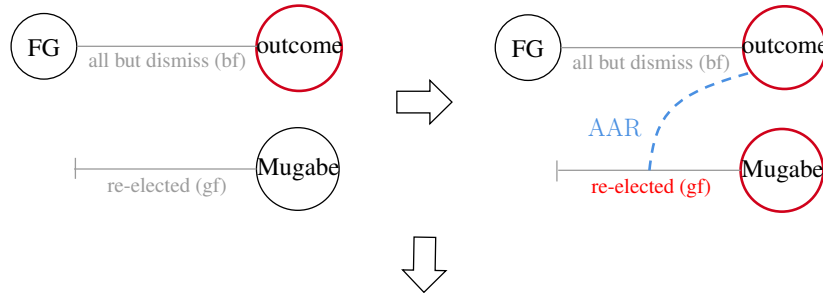
Sentence: Foreign governments all but dismissed the outcome even before it was announced, threatening to leave Mugabe internationally isolated despite his victory.

Opinion holder: foreign governments (FG)

Opinion target: outcome

Abstract anaphor: outcome

Antecedent: Mugabe was re-elected with 1,685,212 votes against 1,258,758 votes for Tsvangirai



POSITIVEPAIR(MDC, foreign governments)

Fig. 2.2 Example how resolution of anaphors results in more explicit sentiments relations and hence better sentiment propagation in the graph-based sentiment inference model.

Benefit from anaphora resolution. Fig. 2.2 demonstrates how the graph-based model could benefit from resolution of anaphors that refer to facts, events, plans, actions, or situations. First, assume that the model is extended to capture sentiments from entities other than the writer. Then the upper part of Fig. 2.2 illustrates the perspective of *MDC* and the lower part the perspective of the *foreign governments*. The showcased sentences occur in the same MPQA document. If we extend the model to detect that the *result* and the *outcome* refer to the fact that *Mugabe was re-elected with 1,685,212 votes against 1,258,758 votes for Tsvangirai*, we would be able to a priori determine that a negative label score for the *re-election* node $score_{re-election}(negative)$ is 1. Since the re-election is a GOODFOR event for its theme (the person who is being re-elected, *Mugabe*), the already implemented propagation

would assign the negative score to *Mugabe*. Such a propagation of sentiment in discourse is more plausible than the simplistic discourse heuristic.

Furthermore, consider an additional sentiment implicature rule that regards the fact that entities in supportive social relations tend to share similar sentiment toward each other and are often mutually positive (Lazarsfeld and Merton, 1954; Choi et al., 2016). When we resolve anaphors *result* and *outcome* we capture that the *MDC* and the *foreign governments* are negative toward the same entities. The new rule then says that it is likely that the *MDC* and the *foreign governments* support each other and hence are mutually positive.

2.1.1.3 Integer Linear Programming (ILP) Sentiment Inference Model

Deng and Wiebe (2014) did not adequately consider that utilizing GOODFOR/BADFOR information for sentiment inference requires resolving several ambiguities: (i) given a document, which spans are GOODFOR/BADFOR events?, (ii) given a GOODFOR/BADFOR text span, what is its effect, GOODFOR or BADFOR?, (iii) is the effect of a GOODFOR/BADFOR event being reversed?, (iv) which noun phrase in the sentence is the agent and which is the theme?, and (v) what are the writer’s sentiments toward the agent and theme, positive or negative?. Therefore, Deng et al. (2014) employed an ILP optimization framework which exploits these inter-dependencies to jointly resolve the ambiguities instead of having a pipeline approach.

In particular, they formulated the ILP objective function such that the model selects a set of labels that optimizes two goals. First, the selected set of labels maximizes the scores given by three local models: a GOODFOR/BADFOR event scorer, a GOODFOR/BADFOR reverser scorer, and a sentiment scorer. Second, the selected set of labels minimizes the cases where the sentiment implicature *constraints* of Deng and Wiebe (2014) are violated. Optimization is performed over two sets of variables. The first set S_{GfBf} contains a variable for each GOODFOR/BADFOR event in the document and each variable in this set is assigned either a GOODFOR or BADFOR effect label. The other set S_{Entity} contains a variable for each agent or theme candidate⁶. Only one agent candidate is assigned a positive or negative label; the other is considered to be an incorrect agent of the GOODFOR/BADFOR event (similarly for

⁶They extract two agent candidates and two theme candidates for each GOODFOR/BADFOR event (one of each will ultimately be chosen by the ILP model). The candidates are extracted using the output of the SENNA SRL tool (Collobert et al., 2011) and the dependency parser (Chen and Manning, 2014). If SENNA labels a span as A0 of the GOODFOR/BADFOR event, they consider it as the *semantic agent*; if there is no A0 but A1 is labeled, they consider A1; if there is no A0 or A1 but A2 is labeled, they consider A2. Similarly, to extract the *semantic theme* candidate, they consider A1, A2, A0 in order. To extract the *syntactic agent* candidate, they find the nearest noun in front of the GOODFOR/BADFOR span, and then extract any other word that depends on the noun according to the dependency parse. To extract the *syntactic theme* candidate, the same procedure is conducted as for the syntactic agent, but the nearest noun should be after the GOODFOR/BADFOR.

the theme candidates). The goal is to assign optimal labels to variables in $S_{Entity} \cup S_{GfBf}$ and at the same violate sentiment constraints as little as possible.

Dependence on existing opinion analysis approaches. Deng et al. (2014) adopt the local sentiment detector from Deng and Wiebe (2014), i.e. they use available resources to detect writer’s sentiments toward agent and theme candidates. Therefore, since we observed that the graph-based model is in need of better explicit sentiment analyzers, the same holds for the ILP model.

Representation and integration of sentiment implicature rules. Deng et al. (2014) use the same sentiment implicature constraints as Deng and Wiebe (2014): (i) the writer has the same sentiment toward entities in a GOODFOR relation (GOODFOR implicature constraint), and (ii) the writer has opposite sentiments toward entities in a BADFOR relation (BADFOR implicature constraint). However, the GOODFOR/BADFOR implicature constraints are now integrated as ILP constraints:

$$\left| \sum_{a, \langle a, e, t \rangle} u_{a, pos} - \sum_{t, \langle a, e, t \rangle} u_{t, pos} \right| + |u_{e, gf} - u_{e, r}| \leq 1 + \xi_{a, e, t}, \forall e \in S_{GfBf} \quad (2.2)$$

$$\left| \sum_{a, \langle a, e, t \rangle} u_{a, neg} - \sum_{t, \langle a, e, t \rangle} u_{t, neg} \right| + |u_{e, gf} - u_{e, r}| \leq 1 + \xi_{a, e, t}, \forall e \in S_{GfBf} \quad (2.3)$$

$$\left| \sum_{a, \langle a, e, t \rangle} u_{a, pos} + \sum_{t, \langle a, e, t \rangle} u_{t, pos} - 1 \right| + |u_{e, bf} - u_{e, r}| \leq 1 + \delta_{a, e, t}, \forall e \in S_{GfBf} \quad (2.4)$$

$$\left| \sum_{a, \langle a, e, t \rangle} u_{a, neg} + \sum_{t, \langle a, e, t \rangle} u_{t, neg} - 1 \right| + |u_{e, bf} - u_{e, r}| \leq 1 + \delta_{a, e, t}, \forall e \in S_{GfBf} \quad (2.5)$$

where $u_{a, pos}, u_{a, neg} \in \{0, 1\}$ ($u_{t, pos}, u_{t, neg} \in \{0, 1\}$) are the binary indicator variables which indicate whether the sentiment label *pos* or *neg* is assigned to the agent *a* (theme *t*) variable, $u_{e, gf}, u_{e, bf} \in \{0, 1\}$ are the binary indicator variables representing whether the event is GOODFOR/BADFOR, $u_{e, r}$ is the binary indicator variables representing whether the event is reversed, and $\xi_{a, e, t}$ ($\delta_{a, e, t}$) are binary variables that take value 1 if the corresponding triplet $\langle a, e, t \rangle$ violates the GOODFOR (BADFOR) implicature constraint.

In contrast to the graph-based model which utilizes only the GOODFOR/BADFOR sentiment implicature constraints, the ILP framework has additional constraints:

- (i) A GOODFOR/BADFOR event must be either GOODFOR or BADFOR, but the model is free to choose whether it is being reversed: $\sum_{l \in \{gf, bf\}} u_{e, l} = 1, \forall e \in S_{GfBf}$.

- (ii) Every GOODFOR/BADFOR event has two agent candidates and only one is indicated as agent and only one polarity is assigned to it: $\sum_{a \in S_{Entity}, (a,e,t) \in S_{Triple}} \sum_{l \in \{pos, neg\}} u_{a,l}$. In this way, the framework disambiguates the agent span and sentiment polarity simultaneously (similarly for themes).
- (iii) If there is more than one GOODFOR/BADFOR event in a sentence, the path between the two GOODFOR/BADFOR events in a dependency parse contains only the coordinating conjunction (*conj*) or open clausal complement (*xcomp*) dependency relation, and there is no other noun between the latter GOODFOR/BADFOR and the conjunction, then it is safe to assume that two agents are the same and the sentiments toward them should also be the same. These assumptions can be written as an ILP constraint similar to the GOODFOR constraint in Equations (2.2–2.3). Henceforth, we refer to this assumption as the *coreference constraint*.

Benefit from anaphora resolution. Similarly to the coreference constraint, we can capture additional sentiments with anaphors that refer to facts, events, situations, or actions. If there is an explicit sentiment toward a theme of a GOODFOR/BADFOR event and the theme co-refers with a fact, event, situation, or action, we can make assumptions about the sentiment label scores of the co-referring event’s agent and theme. Assume that the ILP is extended to handle multiple holders and consider again Example (1) on page 1. The local explicit sentiment analysis scorer assigns $u_{result, neg} = 1$ on the basis of the final sentence. Since the *result* and *re-election* are coreferent, then there is also negative sentiment associated with the *re-election*. Furthermore, *re-elected* is a GOODFOR event so negative sentiment is also directed to its theme i.e. $u_{Mugabe, neg} = 1$. As with the coreference constraint, these assumptions can be written as an ILP constraint similar to the GOODFOR constraint in Equations (2.2–2.3).

2.1.1.4 Probabilistic Soft Logic (PSL) Sentiment Inference Model

Deng and Wiebe (2015a) address some limitations of the prior sentiment inference approaches. They expand the set of sentiment implicature rules of Deng and Wiebe (2014) such that they allow opinion holders other than the writer and opinion targets that may be any of the entities or events (Table 2.2). Moreover, under their new rules, the targets of sentiments may be other sentiments (Ruleset 2 in Table 2.2). They encode inference rules in a Probabilistic Soft Logic (PSL) framework which supports assigning weights to first-order logical rules combining a diverse set of inference rules and performing sound probabilistic inference. Moreover, this framework enables them to work directly with sentiment implicature rules in contrast

Ruleset 1: Aggregation Rules		
1.1.	$\text{HOLDER}(y,h) \wedge \text{ETARGET}(y,t) \wedge \text{POS}(y)$	$\Rightarrow \text{POSPAIR}(h,t)$
1.2.	$\text{HOLDER}(y,h) \wedge \text{ETARGET}(y,t) \wedge \text{NEG}(y)$	$\Rightarrow \text{NEGPAIR}(h,t)$
Ruleset 2: Inference Rules		
2.1.	$\text{POSPAIR}(h_1,y_2) \wedge \text{HOLDER}(y_2,h_2)$	$\Rightarrow \text{POSPAIR}(h_1,h_2)$
2.2.	$\text{POSPAIR}(h_1,y_2) \wedge \text{ETARGET}(y_2,t_2) \wedge \text{POS}(y_2)$	$\Rightarrow \text{POSPAIR}(h_1,t_2)$
2.3.	$\text{POSPAIR}(h_1,y_2) \wedge \text{ETARGET}(y_2,t_2) \wedge \text{NEG}(y_2)$	$\Rightarrow \text{NEGPAIR}(h_1,t_2)$
2.4.	$\text{NEGPAIR}(h_1,y_2) \wedge \text{HOLDER}(y_2,h_2)$	$\Rightarrow \text{NEGPAIR}(h_1,h_2)$
2.5.	$\text{NEGPAIR}(h_1,y_2) \wedge \text{ETARGET}(y_2,t_2) \wedge \text{POS}(y_2)$	$\Rightarrow \text{NEGPAIR}(h_1,t_2)$
2.6.	$\text{NEGPAIR}(h_1,y_2) \wedge \text{ETARGET}(y_2,t_2) \wedge \text{NEG}(y_2)$	$\Rightarrow \text{POSPAIR}(h_1,t_2)$
Ruleset 3: Inference Rules over GOODFOR/BADFOR Event Information		
3.1.	$\text{POSPAIR}(h,e) \wedge \text{AGENT}(e,a)$	$\Rightarrow \text{POSPAIR}(h,a)$
3.2.	$\text{POSPAIR}(h,e) \wedge \text{THEME}(e,th) \wedge \text{GF}(e)$	$\Rightarrow \text{POSPAIR}(h,th)$
3.3.	$\text{POSPAIR}(h,e) \wedge \text{THEME}(e,th) \wedge \text{BF}(e)$	$\Rightarrow \text{NEGPAIR}(h,th)$
3.4.	$\text{NEGPAIR}(h,e) \wedge \text{AGENT}(e,a)$	$\Rightarrow \text{NEGPAIR}(h,a)$
3.5.	$\text{NEGPAIR}(h,e) \wedge \text{THEME}(e,th) \wedge \text{GF}(e)$	$\Rightarrow \text{NEGPAIR}(h,th)$
3.6.	$\text{NEGPAIR}(h,e) \wedge \text{THEME}(e,th) \wedge \text{BF}(e)$	$\Rightarrow \text{POSPAIR}(h,th)$
3.7.	$\text{POSPAIR}(h,a) \wedge \text{AGENT}(e,a)$	$\Rightarrow \text{POSPAIR}(h,e)$
3.8.	$\text{POSPAIR}(h,th) \wedge \text{THEME}(e,th) \wedge \text{GF}(e)$	$\Rightarrow \text{POSPAIR}(h,e)$
3.9.	$\text{POSPAIR}(h,th) \wedge \text{THEME}(e,th) \wedge \text{BF}(e)$	$\Rightarrow \text{NEGPAIR}(h,e)$
3.10.	$\text{NEGPAIR}(h,a) \wedge \text{AGENT}(e,a)$	$\Rightarrow \text{NEGPAIR}(h,e)$
3.11.	$\text{NEGPAIR}(h,th) \wedge \text{THEME}(e,th) \wedge \text{GF}(e)$	$\Rightarrow \text{NEGPAIR}(h,e)$
3.12.	$\text{NEGPAIR}(h,th) \wedge \text{THEME}(e,th) \wedge \text{BF}(e)$	$\Rightarrow \text{POSPAIR}(h,e)$

Table 2.2 Sentiment implicature rulesets of Deng and Wiebe (2015a). Information in the first column is explicitly stated in text and implies sentiment in the second column.

to Deng and Wiebe (2014) and Deng et al. (2014) who integrate the sentiment *constraints* gathered from the sentiment implicature rules.

Representation and integration of sentiment implicature rules. A PSL model is defined using a set of atoms to be grounded and a set of weighted if-then rules expressed in the first-order logic. For example, $\text{FRIEND}(x,y) \wedge \text{VOTESFOR}(y,z) \Rightarrow \text{VOTESFOR}(x,z)$, means that a person will likely vote for the same person as his or her friend. Each predicate in the rule is an *atom* (e.g. $\text{FRIEND}(x,y)$). A *ground atom* is produced by replacing variables with *constants* (e.g. $\text{FRIEND}(\text{Tom}, \text{Mary})$). Each rule is associated with a weight, indicating the importance of this rule in the whole rule set.

Given a dataset, a PSL model first constructs a set of ground atoms. Some of the atoms have fixed truth values and some have unknown truth values. For every assignment of truth

values to the unknown atoms, the model gets a set of weighted distances from satisfaction. A key distinguishing feature of PSL is that each ground atom has a soft, continuous truth value in the interval $[0, 1]$ rather than a binary truth value as in most other probabilistic logic frameworks. PSL seeks the interpretation (i.e. the mapping from atoms to soft truth values) with the minimum distance to satisfaction, $d(r)$, which defines how far a rule r is from being satisfied, and which satisfies all rules to the extent possible. For example, for the rule: $\text{FRIEND}(x,y) \wedge \text{VOTESFOR}(y,z) \Rightarrow \text{VOTESFOR}(x,z)$, the model prefers interpretations where a person’s vote agrees with many friends, that is, satisfies many groundings of the rule. If there is a rule with a higher weight, it will be prioritized.

Dependence on existing opinion analysis approaches. Deng and Wiebe (2015a) argue that recognizing and inferring both explicit and implicit sentiments toward entities and events requires solving three sub-problems for every opinion: (i) classifying its polarity, (ii) identifying its holder, and (iii) identifying its entity-level target (i.e., ETARGET). They intentionally utilize previous work on span-based sentiment analysis instead of building an entity/event-level sentiment system from scratch. Their first PSL model aggregates sentiments toward span-based targets into sentiments toward targets at the entity or event level (Ruleset 1 in Table 2.2). The additional two PSL models build upon the first to infer additional sentiments (Rulesets 2–3 in Table 2.2). We refer to these models as PSL1, PSL2, and PSL3, respectively.

For **PSL1**, Deng and Wiebe (2015a) use three systems for sentiment aggregation: (S1) the model of Yang and Cardie (2013) trained on MPQA 2.0⁷ to collect $\langle \text{holder span, opinion span, target span} \rangle$ triplets, (S2) the model of Yang and Cardie (2014a) trained on MPQA 2.0 to collect opinion expressions and classify their polarity, and (S3) the model of Socher et al. (2013) trained on movie review data to collect opinion expressions and classify their polarity. The set of opinions is union of opinion expressions collected with all three systems.

For each opinion y from this set, a ground atom $\text{POS}(y)$ or $\text{NEG}(y)$ is created as follows. If S2 assigns a polarity, then the polarity S2 predicts is used. If S2 does not assign a polarity, but S3 does, then the polarity S3 predicts is used. If neither S2 nor S3 assign a polarity, then the MPQA subjectivity lexicon (Wilson et al., 2005) is used.

For each opinion y , if S1 assigns a holder h for it, a ground atom $\text{HOLDER}(y,h)$ and score 1.0 is assigned to it. Otherwise, if S3 extracts opinion y , a ground atom $\text{HOLDER}(y,\text{writer})$ is created with the score 1.0 (since S3 assumes the holder is always the writer). Otherwise, the nearest named entity to the opinion span on the dependency graph is treated as the holder

⁷See Section 2.2 for a detailed description of the MPQA Corpus (Wiebe et al., 2005).

and the score is the reciprocal of the length of the path between the opinion span and the holder span in the dependency parse.

For each opinion y and for each ETARGET candidate t , the ground atom $\text{ETARGET}(y,t)$ is created in three ways: (ET1) all the nouns and verbs in the sentence are considered, (ET2) all the nouns and verbs in the target spans and opinion spans that are automatically extracted by systems S1, S2 and S3 are considered, and (ET3) the heads of the target and opinion spans that are automatically extracted by systems S1, S2 and S3 are considered. ET3 also considers the heads of siblings of target spans and opinion spans. In addition, for the ETARGET candidate set extracted by ET2 or ET3, Deng and Wiebe (2015a) run the Stanford coreference system (Recasens et al., 2013; Lee et al., 2013; Clark and Manning, 2015) to expand the set in two ways: (i) the referring entities of every ETARGET candidate are added to the candidate set and (ii) words which the Stanford system judges to be entities, regardless of whether they have any referent or not, are added to the candidate set as well. An SVM classifier with simple syntactic features is trained on MPQA 2.0 to assign a score to the ground atom $\text{ETARGET}(y,t)$.

PSL2 infers sentiments using the atoms and rules in PSL1 (Ruleset 2 of Table 2.2 on page 21).

For **PSL3**, GOODFOR/BADFOR event atoms and rules are added to PSL2 for the inference of additional sentiments (Ruleset 3 of Table 2.2 on page 21). First scores to atoms $\text{GOODFOR}(e)$, $\text{BADFOR}(e)$, $\text{AGENT}(e,a)$, and $\text{THEME}(e,th)$ are calculated. The scores of $\text{GOODFOR}(e)$ and $\text{BADFOR}(e)$ are assigned using the +/- sense-level lexicon (Choi and Wiebe, 2014). An SVM classifier is run to assign scores to $\text{AGENT}(e,a)$, and another SVM classifier is run to assign scores to $\text{THEME}(e,h)$. Both SVM classifiers are trained on the GOODFOR/BADFOR corpus (Deng et al., 2013a). SVM features include unigram, bigram, and syntax information.

The outcome is that the currently available span-based sentiment analysis systems alone (i.e. PSL1) do not provide enough accurate information for sentiment analysis at the entity or event level. Furthermore, when a certain opinion expression is not extracted by any span-based system, it is not input into PSL and PSL cannot possibly find its ETARGETS. Therefore, the simple baseline ET1 that considers all the nouns and verbs in the sentence results in better PSL accuracy than ET2 and ET3 that take outputs of S1, S2, and S3 as well as a comparable F-score. ET1 considers more ETARGET candidates and consequently gives PSL a greater opportunity to remove true negatives leading to an overall increase in accuracy.

Benefit from anaphora resolution. In this section, we have seen that a good recall of span-based opinion analysis systems is crucial for their PSL models. We have previously

model	implicature rules	propagation	explicit sentiment	anaphora resolution	discourse
graph	implicature constraints (edge scores)	LBP	2 systems & 3 lexica	-	heuristics
ILP	implicature constraints (ILP constraints)	global optimization	2 systems & 3 lexica	coreference constraint	-
PSL	first-order logic rules	PSL inference	3 systems & 1 lexicon	coreference system	-

Table 2.3 The recap of Deng and Wiebe’s work on sentiment inference with emphasis on the following: dependence on existing systems for capturing explicit sentiment, usage of anaphora resolution, and focus on discourse beyond a single sentence.

shown that we can infer more opinions once we resolve abstract anaphors. Therefore, their model would likely benefit from the resolution of anaphors that refer to facts, event, situations, and actions, in addition to resolving anaphors that refer to concrete objects.

2.1.1.5 Summary

In this section we reviewed the three most prominent learning approaches to sentiment inference having two questions in mind: (i) how reliant are available sentiment inference models on existing approaches for detecting explicit opinion expressions and their roles and (ii) how could they benefit from anaphora resolution. In this thesis, we will not integrate or evaluate our models for detecting explicit opinion expressions and abstract anaphora resolution into the described models for sentiment inference. The observations from this section serve as a motivation for our work.

In their studies, Deng and Wiebe intentionally exploit available systems and lexica for detection and categorization of opinion expressions and their roles (column 4 in Table 2.3). However, they repeatedly show that the imperfections of these systems and lexica considerably effect the performance of their sentiment inference approaches.

They acknowledge the benefit of anaphora resolution for the ILP and PSL frameworks, but only of anaphors that refer to concrete entities that exist in the real world i.e. coreference resolution (column 5 in Table 2.3). However, in the survey we recurrently notice that their models could also benefit from resolution of anaphors that refer to facts, events, situations, and actions i.e. from Abstract Anaphora Resolution (AAR). Moreover, we illustrated that once there is an available system that resolves such type of anaphors, its output can be easily integrated in their models. Finally, we showed that AAR enables a better propagation of sentiment in discourse beyond one sentence. The discourse-oriented propagation is nearly

completely ignored in their work (column 6 in Table 2.3), except in the simplistic discourse heuristic in the graph-based model which performs poorly.

Deng and Wiebe’s sentiment inference approaches (Section 2.1.1) are based on the GOODFOR/BADFOR lexicon (Deng et al., 2013a; Choi et al., 2014; Choi and Wiebe, 2014; Choi et al., 2017). In the meantime, alternative lexica have been proposed: effect functors (Ruppenhofer and Brandes, 2016), connotation frames (Feng et al., 2013; Rashkin et al., 2016), sentiframes (Klenner, 2015; Klenner and Amsler, 2016), to name a few.

2.2 Fine-Grained Opinion Analysis

Although there has been a lot of research on subjectivity detection and polarity classification, less attention has been paid to the extraction of *opinion roles*, i.e. opinion holders and targets. The exception is the extraction of opinion targets in the product review domain (Hu and Liu, 2004; Jakob and Gurevych, 2010; Liu et al., 2013a,b, 2014, 2015). However, Wiegand et al. (2015) argue that the review domain is not suitable for studying opinion role extraction. The first reason for this is that product reviews typically reflect only the author’s views on a particular product. Therefore, the majority of explicitly mentioned opinion holders refer to the author of the review. Second, opinion expressions in reviews are mostly adjectives. Third, the review domain is typically focused on products and hence only specific semantic types are eligible for opinion roles. For example, persons are less likely to be opinion targets. For all these reasons, opinion role extraction from product reviews is not a challenging task.

On the other hand, news corpora typically tend to be multi-topic. As a consequence, opinion targets can be of different semantic types. For example, persons can function both as opinion holders and targets. Moreover, we do not have a bias toward adjectives. In the MPQA corpus (Section 2.2.1), Wiegand et al. (2015) found that there are 10% more opinion verb mentions than opinion adjective mentions. In this section we survey the literature on extraction and categorization of opinion expressions and their roles in news articles.

2.2.1 MPQA Opinion Corpus

The Multi-Perspective Question Answering (MPQA) Opinion Corpus⁸ consists of news articles and other text documents manually annotated for opinions and other *private states* (Quirk et al., 1985): opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments. MPQA is the only commonly accepted corpus for English containing manual annotation of both opinion holders and targets.

⁸http://mpqa.cs.pitt.edu/corpora/mpqa_corpus

original MPQA	ULA	ULA-LU
economic collapse in Argentina	travel guides	emails related to the Enron case
Bush's 2002 State of the Union Address	transcriptions of spoken conversation	spoken language transcripts
detention of prisoners in Guantanamo Bay	fundraising letters	newswire text
U.S. State Department Human Rights Report	a chapter of the 9/11 report	Wall Street Journal texts
U.S. State Department report on human rights	a chapter from a linguistics textbook	translations of Arabic source texts
Israeli settlements in Gaza and the West Bank	articles from Slate magazine	
space missions of various countries		
relationship between Taiwan and China		
presidential coup in Venezuela		
2002 presidential election in Zimbabwe		

Table 2.4 MPQA 2.0 topic categories.

MPQA 1.0 (Wiebe et al., 2005). The first version consists of 535 documents from 187 different foreign and U.S. news sources, dating from June 2001 to May 2002. About two thirds of the documents were selected to be on one of ten specific topics (e.g. economic collapse in Argentina) listed in the first column in Table 2.4. The annotation scheme is centered on the notion of *private state* described as the state of an *experiencer* holding an *attitude*, optionally toward a *target*. They create two private state frames for three main types of private states: (i) explicit mentions of private states, (ii) speech events expressing private states, and (iii) expressive subjective elements. An example of each type can be found in (7).

- (7) a. "The U.S. fears a spill-over," said Xirao-Nima.
 explicit *speech*
- b. "The report is full of absurdities," Xirao-Nima said.
 expressive subjective

Direct subjective frames are used to represent Direct Subjective Elements (DSEs), i.e. both speech events expressing private states and explicitly mentioned private states. Expressive subjective element frames are used to represent Expressive Subjective Elements (ESEs). The agent frame is used to mark noun phrases that refer to holders of private states and speech events. For example, agent frames are created for *U.S.* and *Xirao-Nima* in sentence (7a).

MPQA 2.0 (Wilson, 2008). The important differences between the first and second versions are that MPQA 2.0 contains additional sets of documents, it is extended to be a multi-genre corpus, and it also covers attitudes other than sentiments. Additional sets of documents are the Opinion Question Answering (QpQA) subset of 98 documents, 85 Wall Street Journal texts from the Penn TreeBank (XBank), the Unified Linguistic Annotation (ULA) subset of 48 documents, and the Language Understanding (ULA-LU) subset of 24 documents. ULA and ULA-LU subsets contain documents other than news articles (columns 2–3 in Table 2.4).

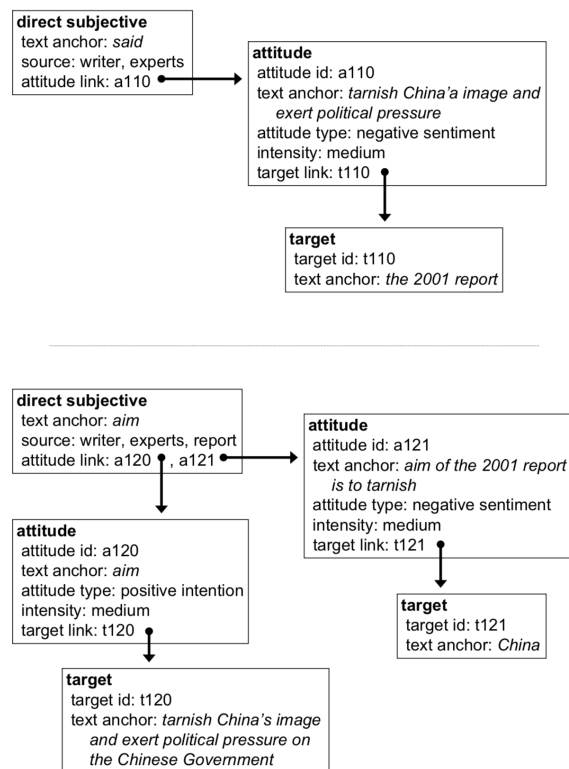


Fig. 2.3 MPQA annotation scheme. Figure from Wilson (2008).

Private states in language are often quite complex in terms of the attitudes they express and the targets of those attitudes. For example, Wilson (2008) showcased the private state represented by the direct subjective phrase *are happy*:

(8) I think people **are happy** because Chavez has fallen.

From this sentence we understand that *people* expressed a positive sentiment toward the fall of *Chavez*. However, there is a second attitude, a negative sentiment toward *Chavez* himself, which can be inferred from the phrase *happy because Chavez has fallen*. Just as a private state may involve more than one type of attitude, an attitude may be directed toward more than one target. In the original annotation scheme, attitudes are represented with the attitude type attribute in direct subjective and expressive subjective element frames, and targets are represented with the target attribute in direct subjective. In the new representation, attitudes and targets are annotation frames, with target frames linking to attitude frames and attitude frames linking to private state frames. Figure 2.3 gives the various direct subjective, attitude, and target frames for the sentence (9) and shows how they are all linked together.

- (9) Its **aim** of the 2001 report is to tarnish China’s image and exert political pressure on the Chinese Government, human rights experts **said** at the seminar held by the China Society for Study of Human Rights (CSSHR) on Friday.

The new representation also includes a new set of attitude types: sentiment, agreement, arguing, intention, speculation, and other attitudes. Sentiment, agreement, arguing, and intention may be further broken down into positive and negative variants.

MPQA 2.0 is the platform for experiments in Chapter 3. We report detailed pre-processing of MPQA⁹ and data statistics in Appendix A. Table 3.3 on page 62 in Chapter 3 summarizes data statistics.

MPQA 3.0 (Deng and Wiebe, 2015b) . The second and third versions differ in their target annotations. Since the exact boundaries of the spans are hard to define even for human annotators (Wiebe et al., 2005; Yang and Cardie, 2013), the target span in MPQA 2.0 could be a single word, an NP or VP, or a text span covering more than one constituent. In contrast, in MPQA 3.0 each target is anchored to the head of an NP or VP, which is a single word. It is called an ETARGET since it is an entity or an event. In MPQA 2.0, only attitudes have target-span annotations. In MPQA 3.0 both attitudes and expressive subjective elements have ETARGET annotations. Importantly, the ETARGET includes the targets of both explicit and implicit sentiments. Unfortunately, MPQA 3.0 contains only 70 documents.

2.2.2 Phrase-level Opinion Extraction and Categorization

The area of opinion extraction—detecting the boundaries of opinion expressions—ranges from identifying the sentiment-bearing expressions at the level of individual words, phrases, sentences, or even documents (see Pang and Lee (2008) and Liu (2015) for a thorough survey). However, since a certain sentence may contain more opposing views from different entities, Hovy (2011) argues that extraction and categorization (determining polarity) of opinions at the sentence or document level is not sufficient for proper understanding of opinionated text.

At the phrase level, opinion extraction has traditionally been tackled as a sequence labeling problem where a system needs to label a word with the conventional BIO tagging scheme: B for the beginning of an opinion expression, I for tokens inside the opinion expression, and O to indicate tokens outside any opinion expression. The older work manually designed relevant features and utilized different variants of CRF approaches to

⁹Examples how to use our scripts can be found at https://github.com/amarasovic/naacl-mpqa-srl4orl/blob/master/generate_mpqa_jsons.py.

output labels (Breck et al., 2007; Yang and Cardie, 2012). Nowadays it is common to produce an input's representation with a stack of bi-directional LSTM neural networks while still using a CRF layer to output labels (Irsoy and Cardie, 2014b; Katiyar and Cardie, 2016). The work mentioned so far considers only extraction but not the categorization of opinions.

To the best of our knowledge, Wilson et al. (2005) are the first to address both the phrase-level opinion expression extraction and categorization with a learning model. Choi and Cardie (2010) are among the first to *jointly* address the opinion expression extraction and the opinion attribute categorization, i.e., determining values for polarity and intensity. In particular, they integrate the hierarchical parameter sharing technique (Zhao et al., 2008) in a CRF. Yang and Cardie (2014b) embedded this approach in semi-Markov CRF that allows contiguous spans in the input sequence (e.g., a noun phrase) to be treated as a group rather than as distinct tokens. After the "CRF-phase", Irsoy and Cardie (2014a) were the first to propose a neural approach.

2.2.3 Extraction of Opinion Holders

To the best of our knowledge, Choi et al. (2005) and Choi et al. (2006) proposed the first approaches for labeling of opinion holders in a given sentence from the MPQA corpus. They also discuss the differences and possible benefits of related tasks: named entity recognition and semantic role labeling.

Choi et al. (2005) proposed a hybrid approach that combines two types of learning methods: graphical models and extraction pattern learning. In particular, they employ a linear-chain CRF that labels every word in a sentence with the BIO notation. They utilize a variation of AutoSlog (Riloff, 1996) to learn extraction patterns that rely on both the syntax and lexical semantics of a sentence. The extraction patterns provide two kinds of information: whether a word activates any holder extraction pattern and whether a word is extracted by any holder pattern. Each extraction pattern has frequency and probability values which are used as CRF features. Other features for the CRF are designed considering three properties of opinion holders: (i) they are mostly noun phrases, (ii) the holder phrases should be semantic entities that can bear or express opinion, and (iii) the holder phrases should be directly related to an opinion expression. The third condition is what delineates the labeling of opinion holders from named entity recognition (Bikel et al., 1997; Peters et al., 2018a). To be precise, holder labeling requires the recognition of opinion expressions and a more sophisticated encoding of sentence structure to capture relationships between holder phrases and opinion expressions. Although the proposed system utilizes a feature that checks whether the parent chunk of the current tokens includes an opinion word, it is still not able to specify which opinion expression the labeled holder relates to.

For this reason, Choi et al. (2006) proposed a model that jointly labels both opinion expressions and holders, and specifies which holders correspond to which opinion expressions (i.e. link prediction). Their approach consists of three components. They train two sequence-tagging linear-chain CRFs for the labeling of opinion expressions and holders without the knowledge of any nearby or neighboring entities, or relations. The input to these CRF models are local syntactic and lexical features. The third component is a binary maximum entropy classifier trained using only local syntactic information to identify the link relation. The global inference procedure is implemented via Integer Linear Programming (ILP) to produce an optimal and coherent extraction of entities and relations, inspired by Roth and Yih (2004).

Choi et al. (2006) state that 60% of opinion-holder relations in the MPQA 1.0 dataset appear as predicate-argument relations. Consider the following sentence.

- (10) [Taipei]_{H₁}^{A₁} [was [angered]^P]_{O₁} when [then-president Bill Clinton]_{H₂}^{A₀} [pledged]_{O₂}^P the "three NOs" during his China visit in 1998.

Opinion holders: *Taipei* and *then-president Bill Clinton* are semantic roles of the corresponding predicates: *angered* and *pledged*. Thus, it is expected that Semantic Role Labeling (SRL), the task of predicting predicate-argument structure of a sentence will be beneficial for the labeling of opinion holders. Choi et al. (2006) integrated SRL information in two ways. First, they used boolean features to inspect whether the span of an SRL argument and an opinion entity match exactly. Second, they used an additional ILP constraint based on SRL. The SRL constraint forces the extraction of verb(V)-agent(A0) frames such that the model assigns very high weights for links that match V-A0 frames generated by SRL. They show that SRL features for the link classifier further improve the performance but forcing the extraction of such frames via extra ILP constraints hurts performance by not allowing the extraction of non-V-A0 pairs that could have been better choices.

Johansson and Moschitti (2013) also show that relational features derived from dependency-syntactic and semantic role structures can significantly improve the performance of automatic approaches for a number of fine-grained opinion analysis sub-tasks: extraction of opinion expressions, categorization of opinion expressions, and identifying opinion holders. These features make it possible to model the way opinions expressed in natural-language discourse interact in a sentence over arbitrary distances.

2.2.4 Extraction of Opinion Targets

Once the models for extraction of opinion expressions and their holders were proposed, the most reasonable next step is to extend them to extract targets as well. Since opinion

target extraction followed opinion expression and opinion holder extraction, there are far less studies with focus on *only* extraction of targets. Exceptions are the studies of Stoyanov and Cardie (2008) and Qiu et al. (2011).

Stoyanov and Cardie (2008) treat target extraction as a topic coreference resolution problem. The key to their approach is to cluster opinions sharing the same target together.

On the other hand, Qiu et al. (2011) propose a semi-supervised bootstrapping approach that extracts opinion words (or targets) iteratively using opinion words and targets that are extracted in previous iterations through the identification of syntactic relations. These relations can be identified using a dependency parser and then utilized to expand the initial opinion lexicon and to extract targets (hence bootstrapping). They call it double propagation as it propagates information between opinion words and targets. Since the model only needs an initial opinion lexicon to start the bootstrapping process, the method is semi-supervised due to the use of opinion word seeds.

2.2.5 Joint Approaches to Fine-Grained Opinion Analysis

Presently, in addition to detecting opinion expressions and their holders, we expect from models that analyze opinionated text to also detect entities or propositions at which the attitude is directed.

Initially, pipeline models were investigated. These models first predict opinion expressions and then, given an opinion, label its opinion roles, i.e. holders and targets. The most notable pipeline approach was proposed by Kim and Hovy (2006). They explored how an opinion holder and a target are semantically related to an opinion bearing word in a sentence. In particular, given a sentence and an opinion bearing word, their method identifies FrameNet (Fillmore et al., 2003) elements in the sentence and searches which frame element corresponds to the opinion holder and which to the target. Since FrameNet has a limited number of words in its annotated corpus, for a broader coverage, they used a clustering technique to predict a most probable frame for an unseen word.

Yang and Cardie (2013) argue that uncertainty regarding the spans of opinion entities can harm the prediction of opinion relations. On the other hand, evidence of opinion relations might provide clues to guide the accurate extraction of opinion entities. Since then, pipeline models have been substituted with so-called joint models that simultaneously identify all opinion entities and predict which opinion role is related to which opinion. Yang and Cardie (2013) extended the work of Choi et al. (2006). In particular, their model (i) labels opinion targets and predicts to which opinion expression they relate to, (ii) handles opinion expressions that do not have an explicit holder or target, and (iii) simplifies the ILP formulation. This model is still the state-of-the-art fine-grained opinion analysis model.

Since 2015, the majority of NLP research has been motivated by the fact that the feature-based state-of-the-art models depend on the availability of lexica, dependency parsers, named-entity taggers, and other resources typically available for English and certain domains. Thus, they are not effortlessly applicable to other languages or domains even if the labeled data is available. For this same reason, Katiyar and Cardie (2016) proposed a neural version of the state-of-the-art model of Yang and Cardie (2013). They investigated the use of deep bi-directional LSTMs for joint labeling of a given sentence with opinion entities and predicting which opinion expressions holders and targets correspond to.

The simplest way to train a LSTM for sequence labeling is to make a prediction at each word independent of predictions for other words. This kind of training is known as *word-level log-likelihood*. However, it is well-acknowledged that it is beneficial to consider the correlations between labels in neighborhoods for sequence labeling (or general structured prediction) tasks and jointly decode the best chain of labels for a given input sentence. The common way to utilize the dependencies between labels is to use a CRF model. Luckily for modern NLP techniques based on neural networks, integrating a CRF predictor in a neural architecture is straightforward. Collobert et al. (2011) and Huang et al. (2015) were among the first to recognize that the outputs of the final network's layer can be the input to a CRF and that the whole architecture can be trained by minimizing the negative log-likelihood obtained by the CRF predictor. This kind of loss is known as *sentence-level log-likelihood*.

Katiyar and Cardie (2016) recognized that the sentence-level log-likelihood cannot directly be used for modeling relations between non-adjacent words in the sentence. For this reason they proposed a new loss that calculates the distance of each word to its corresponding opinion entity to the left and right, as well as a transition score for jumping from the opinion entity label i and distance d to the opinion entity label j and distance d' of adjacent words.

However, their model lags behind a state-of-the-art feature-based CRF inference approach for the labeling of opinion holders and targets. Since their model is trained using only word embeddings and a complex objective that aims to simultaneously optimize two tasks (determining opinion entities and predicting which opinion role is related to which opinion), it is hard to trace and understand what is solved and what is next for neural models trained on the MPQA corpus. Moreover, both the neural and the CRF joint models achieve about 55% F1 score for predicting which targets relate to which opinions in MPQA. Thus, these models are not yet ready to answer the question this line of research is generally motivated with: "Who expressed what kind of sentiment toward what?", and consequently assist applications such as sentiment inference and opinion summarization.

2.2.6 Linguistic Studies

So far we review studies with focus on developing novel machine learning models for understanding of opinionated text which is also the focus of this thesis. However, there are studies with linguistic focus.

Wiegand and Ruppenhofer (2015) present an approach for the labeling of opinion holders and targets of verbal predicates. They assume that verbs that can be found in a common sentiment lexicon can be divided into three different types. Each type has a characteristic mapping between semantic roles and opinion roles (i.e opinion holders and targets). Thus, they reduce the problem of opinion role labeling to the automatic categorization of opinion verbs. For the induction of opinion verb types, they consider semi-supervised graph clustering with an appropriate similarity metric.

Wiegand et al. (2016a) examine opinion roles that are realized in noun compounds whose head is an opinion noun such as *user rating* or *victim support*. This task is challenging because the immediate context of compounds does not contain explicit cues as the relation between head and modifier. Apart from examining traditional features from noun compound analysis, they also introduce novel features specially designed for the analysis of opinion compounds. They solve the given task as a supervised classification problem and employ Markov Logic Networks.

Wiegand et al. (2016b) focus on the views that an opinion expression evokes. They distinguish between the two most common types: (i) expressions conveying sentiment of the entities participating in the event denoted by the opinion word (*actor views*) and (ii) expressions conveying sentiment of the speaker of the utterance (*speaker views*). They show that the distinction between those categories is relevant for the opinion role extraction. They employ Markov Logic Networks since they allow them to both define features and global constraints between different instances.

2.2.7 Opinion Summarization

The fine-gained opinion analysis is usually motivated to assist other tasks such as opinion summarization. A common approach to opinion summarization is to list aspects as well as the number of positive and negative opinions for each aspect.

For example, the notion of summary of Stoyanov and Cardie (2011) is fundamentally different from the textual summaries usually used in NLP. They expect that users will use summaries of fine-grained opinion information in two distinct ways, giving rise to two distinct summary formats. The first is an aggregate opinion summary in which multiple opinions from a holder on a target are merged into a single aggregate opinion that represents

the accumulated opinions of the holder on that target considering the document as a whole. In another type of summary, a so-called opinion set summary, multiple opinions from a holder on a target are collected into a single set (without analyzing them for the overall trend). This type of summarization is similar to concept-map-based summarization (Falke et al., 2017). However, Stoyanov and Cardie (2011) do not aim to learn the best concept-map-based summary but only to construct its representation from the output of a FGOA system.

While this format gives an overall idea of people’s opinion, reading the actual text might be necessary to gain a better understanding of specific details. Textual opinion summaries are created following mostly extractive methods and various formats ranging from lists of words (Popescu and Etzioni, 2005), phrases (Lu et al., 2009), sentences (Mei et al., 2007; Blair-Goldensohn et al., 2008; Lerman et al., 2009; Wang and Ling, 2016), and documents (Angelidis and Lapata, 2018). However, even the most recent approach of Angelidis and Lapata (2018) is still focused on the review domain. For each product from a list of products, their goal is to produce a summary of the most salient opinion expressions by selecting a small subset of opinions from the set of all opinions expressed in all reviews of the product. To collect those opinions they use a pre-defined list of the aspects pertaining to a specific domain. Thus, it is not trivial to extend their model to challenging domains such as newswire.

In this thesis, we do not evaluate impact of our models on opinion summarization.

2.3 Abstract Anaphora Resolution

An extensive overview of the literature related to the phenomenon of abstract anaphora has been written by Kolhatkar et al. (2018). The central topic of this section is to (i) establish terminology related to the phenomenon of abstract anaphora and to (ii) investigate the reasons why the resolution of abstract anaphors is still relatively unexplored even though they are very frequent across languages (Eckert and Strube, 2000; Vieira et al., 2005; Dipper and Zinsmeister, 2012). Moreover, an understanding of them is valuable for computational systems in machine translation (Le Nagard and Koehn, 2010; Hardmeier et al., 2015), summarization (Steinberger et al., 2005; Orăsan, 2007), question answering (Quarteroni, 2007; Vicedo and Ferrández, 2008), and, as we have hypothesize in Section 2.1, probably in sentiment inference as well.

2.3.1 Terminology

What is anaphora? Various definitions of anaphora have been proposed. One way of thinking about anaphora is based on the notion of cohesion (Mitkov, 2014, p. 4–5). In Mitkov’s own words,

Anaphora is cohesion which points back to some previous item. The "pointing back" word or phrase is called an *anaphor* and the entity to which it refers or for which it stands is its *antecedent*. The process of determining the antecedent of an anaphor is called *anaphora resolution*.

Another way to describe anaphora is to say that anaphora refers to the relation between two linguistic entities, an *anaphor* and an *antecedent*, in which the interpretation of anaphor depends upon the meaning of the antecedent (Huddleston and Pullum, 2002, p. 1453) (Kolhatkar et al., 2018). The actual meaning of an antecedent is referred to as its *referent*.

Coreference vs. anaphora resolution. An anaphor and its antecedent are *coreferential* if the anaphor refers to the antecedent and they have the same referent in the real world (Mitkov, 2014, p. 5). Alternatively, in an *anaphoric relation*, the anaphor and its antecedent may refer to different entities in the real world. For example, *bridging anaphors* are anaphoric noun phrases that are not coreferent, but instead linked via associative relations to the antecedent (Hou et al., 2018). Therefore, an anaphoric relation may be coreferential or not. Likewise, an expression might be coreferential without being anaphoric. This is the case of subsequent mentions of self explaining expressions such as *the champion of the 2002 World Cup—the team that won the 2002 world cup championship* (Vieira et al., 2005). In this thesis, we are interested in anaphoric relations.

Types of anaphora. There are different ways of categorizing instances of anaphora. If we examine the syntactic type of either the anaphor or its antecedent we can distinguish pronouns and full noun phrases (NPs) as anaphors as well as nominal and non-nominal antecedents. Non-nominal antecedents can be further examined from the semantic perspective.

Semantically, non-nominal antecedents typically denote *abstract objects* (Asher, 1993) such as facts, events, actions, situations, or propositions, in contrast to the *concrete objects* that denote real existing entities such as people, places, institutions, and locations. The majority of previous work focuses on the semantic aspects of non-nominal antecedents, but they utilize different theoretical approaches. As a result different terminologies emerged: *abstract anaphora* (Asher, 1993; Navarretta and Olsen, 2008; Dipper et al., 2011), *discourse deixis* (Webber, 1988; Eckert and Strube, 2000; Recasens, 2008), *indirect anaphora* (Gundel

et al., 2004; Botley, 2006), and *complex anaphora* (Consten et al., 2007), *inter alia*. Kolhatkar et al. (2018) note that these approaches can be divided into two classes depending on whether they distinguish two types of abstract objects, eventualities and factualities, or not.

Eventualities are events and states which have spatial, temporal, and causal properties and can be observed by the senses. They are similar to concrete objects in that they can be directly introduced into the discourse model by syntactic constructions (Asher, 1993, p. 86). *Factualities* are facts and propositions which do not have a spatiotemporal location and are not perceivable by the senses but are only mentally conceivable (Asher, 1993, p. 57). In contrast to eventualities, they are introduced by the semantic constraints imposed by nouns or verbs which require their arguments to be of a certain type (Asher, 1993, p. 116, p. 175).

Furthermore, to be precise, *deixis* is not a type of anaphora. If non-nominal antecedent anaphora is viewed as an instance of deixis, then there is no antecedent involved. Instead the anaphor's referent is determined by pointing to a region of the discourse or discourse model (Webber, 1988). As Kolhatkar et al. (2018) we use the term *antecedent* to refer to the linguistic constituent that most closely represents the intended interpretation of the anaphor, so far as it is overtly realized, and the term *referent* to refer to the interpretation itself. Likewise, when we say that an anaphor refers to a non-nominal antecedent, we mean that the anaphor refers to some abstract referent that is represented in the text by the non-nominal antecedent.

Finally, although each of these terminologies reflect slight variations in how the authors define the phenomenon, they all have in common that they describe anaphora where (i) antecedents are usually not NPs and where (ii) referents are abstract objects. In this thesis we use the term abstract anaphora since our ultimate goal is to assist sentiment inference models. Therefore, whether the entity to which an antecedent refers is abstract or concrete is a crucial distinction. However, we support Kolhatkar et al. (2018) in the claim that the fact that antecedents are non-nominal present a distinctive and challenging problem from a computational perspective. Therefore, it is plausible to define abstract anaphora in terms of the syntactic shape of the antecedent, i.e. as a syntactic notion.

In this thesis we focus on non-nominal antecedents, but we do not make any assumptions about their semantic type nor about the semantic or syntactic type of the anaphor. Therefore we say we are addressing *unrestricted* abstract anaphora resolution.

2.3.2 Linguistic Properties of Abstract Anaphora

Realization of abstract anaphora. Abstract anaphors can be signaled with a variety of expressions. For example, a common way is to use a simple demonstrative pronoun (*this* as

in (11a)) or *that*.¹⁰ Moreover, the personal pronoun *it* may also be used as in (11b). We use the term *pronominal* anaphors as an umbrella term for demonstrative pronouns (*this*, *that*) and the personal pronoun *it*.

- (11) a. China might stave off a crisis if it acts as forcefully as it did to arrest the 1985 decline, when Beijing slammed the brakes on foreign-exchange spending and devalued the currency. But this time, China faces a more difficult battle because of economic forces that have come into play since the Tiananmen Square killings June 4. For example, China's hard-currency income is expected to suffer from the big drop in tourist arrivals since June 4. [*Antec* **Revenue from tourism this year is projected to total \$1.3 billion, down from \$2.2 billion last year.**] [*AnaphS* Because of **this**_{AA} and the huge trade gap, the deficit in China's current account, which measures trade in goods and services plus certain unilateral transfers of funds, is expected to widen sharply from the \$ 3.8 billion deficit last year.]
- b. It is this circle that makes sure that women remain inferior, that democracy is unthinkable and that exposure to the outside world is minimal. It is also that circle that leads the way in blaming everybody outside the Moslem world, for the miseries of the region. The outer circle is largely financed by Saudi Arabia, but also by donations from certain Moslem communities in the United States and Europe and, to a smaller extent, by donations of European Governments to various NGO's and by certain United Nations organizations, whose goals may be noble, but they are infested and exploited by agents of the outer circle. The Saudi regime, of course, will [*Antec* **be the next victim of major terror, when the inner circle will explode into the outer circle**]. [*AnaphS* The Saudis are beginning to understand **it**_{AA}, but they fight the inner circles, while still financing the infrastructure at the outer circle.]

Another common way to realize abstract anaphors is with *shell nouns* (Schmid, 2000; Kolhatkar et al., 2013b; Kolhatkar, 2015). These are highly abstract nouns, such as *plan* in (12a) and *reason* in (12b) which can only be interpreted jointly with their *shell content* i.e. the text that provides the interpretation of the shell noun phrase. The concept of shell content is similar to the concept of an antecedent except that the term antecedent is not appropriate in some constructions of shell nouns. Nevertheless, we use the term antecedent for shell content. Kolhatkar et al. (2013b) refer to shell nouns whose antecedent occurs in the prior discourse

¹⁰In all examples in this section, the anaphoric expressions are shown in boldface and indicated with the subscript *AA*. Their antecedents are shown in boldface and put in brackets that start with the subscript *Antec*. The sentence with an anaphor (i.e. anaphoric sentence) is put in brackets that start with the subscript *AnaphS*. Examples (11a–11b) are drawn from the CoNLL-12 corpus (Pradhan et al., 2012) which is based on OntoNotes (Pradhan et al., 2007) and the ARRAU corpus (Uryupina et al., 2016), respectively. Examples (12a–12c) are drawn from the ARRAU corpus, the ASN corpus (Kolhatkar et al., 2013b), and the CSN corpus (Kolhatkar et al., 2013b), respectively.

as Anaphoric Shell Nouns (ASNs), e.g. *plan* in (12a) and *reason* in (12b). Otherwise we are talking about Cataphoric Shell Nouns (CSNs), e.g. *fact* in (12c).

- (12) a. In an open letter that will run today in the trade journal *Automotive News*, Ron Tonkin, president of the National Car Dealers Association, says [*Antec* **dealers should cut their inventories to no more than half the level traditionally considered desirable**]. Mr. Tonkin, [...], said that with half of the nation's dealers losing money or breaking even, it was time for "emergency action". U.S. car dealers had an average of 59 days supply of cars in their lots at the end of September, according to Ward's *Automotive Reports*. But Mr. Tonkin said dealers should slash stocks to between 15 and 30 days to reduce the costs of financing inventory. [*AnaphS* Ford Motor Co. and Chrysler Corp. representatives criticized **Mr. Tonkin's plan_{AA}** as unworkable].
- b. "It's a debate that rages in the absence of any data", said Dr. Lawrence Brody, a geneticist at the National Center for Human Genome Research in Bethesda, Md. Dr. O'Brien is on the side of those who think the mutations benefited populations in generations past. He noted that the great population geneticist, J.B.S. Haldane, said in the 1940's that [*Antec* **probably the greatest selection pressure of all is not a changing environment or a scarce food supply but the harsh culling of infectious disease**]. [*AnaphS* Dr. O'Brien said he would not be surprised if mutations like the gene that protects against AIDS were preserved for **this reason_{AA}**].
- c. Congress has focused almost solely on **the fact_{AA}** that [special education is expensive - and that it takes away money from regular education.]

Finally, when an anaphoric expression is a full noun phrase such as *Mr. Tonkin's plan* in (12a) or *this reason* in (12b) we refer to it as a *nominal* abstract anaphor. In this thesis we do not examine reflexive pronouns (e.g. *itself*), pro-verb constructions (e.g. the anaphor of a form of *do*), pro-action (e.g. *does it*), or adverbs (e.g. *so*).

Lexical and semantic properties of abstract anaphora. Kolhatkar et al. (2018) categorize semantic properties of abstract anaphora into three types: (i) those imposed by the context of the anaphor, (ii) by the anaphor or the antecedent itself, and (iii) by the context of the antecedent. We reflect only on properties that are important for distinguishing different antecedent candidates. Consult Kolhatkar et al. (2018) for properties that differentiate abstract and concrete anaphors.

The context of an anaphoric expression determines the semantic type of the referent where semantic type is an abstract object (e.g. fact, proposition, or event). For example, a verb *happen* can only be used with subjects denoting some sort of event (Asher, 1993, p. 22,

p. 192). Furthermore, the anaphor's semantic type preferences can constrain the syntactic type of antecedents. Kolhatkar et al. (2018) showcase this observation in Example (13). The anaphor refers to a concept and therefore a verb phrase is the antecedent instead of the full sentence.

(13) John [*Antec* **crashed the car**]. Jane did **that**_{AA} too. (concept)

Finally, the antecedent puts an important preference on tense. For example, antecedents whose semantic type is *fact* have a strong preference for present or past tense since future facts are unknown (Schmid, 2000, p. 104–105).

Syntactic properties of abstract antecedents. Antecedents of abstract anaphors differ considerably in syntactic type (Asher, 1993, p. 226). Vieira et al. (2005) conducted a small corpus study to investigate anaphoric and coreferential properties of demonstrative noun phrases (i.e. *this* NPs) in French and Portuguese written texts. They categorized relations between a demonstrative noun phrase and its antecedent in three categories: direct coreference, indirect coreference, and other anaphora. Other anaphora contains cases when the antecedent is not a nominal expression or the relation between a demonstrative NP and its antecedent is not a coreference relation. They show that the antecedents of demonstrative NPs were identified in 38% cases as one single sentence, part of a sentence, or paragraphs. That is, systems that work on anaphor resolution based on NP structures are likely to fail on about 38% of the cases. Concrete demonstratives were related to NP antecedents for the majority of the cases in both languages (94 to 100%). In contrary, for abstract head nouns they had difficulty drawing conclusions, since they seem to be generally distributed over different syntactic types. Moreover, Kolhatkar et al. (2018) note on page 43 that non-nominal antecedents are not always clearly delimited and identifiable stretches of text.

Distance between anaphor and antecedent. Information about the distance between the anaphor and the antecedent plays an important role in anaphora resolution systems since it can narrow down the search scope of candidates for antecedents (Mitkov, 2014, p. 17). The distance can be measured in number of tokens, sentences, the number of edges between nodes in some discourse structure, etc. The distance varies with respect to the anaphoric expressions. In particular, the more semantic information an anaphor has the larger the average distance to its antecedent (Byron, 2003, p. 34–35). Therefore, the pronominal anaphors (*this, that, it*) are typically closer to their antecedents than *this* NPs (Schmid, 2000; Kolhatkar, 2015). This is also evident from examples (11a) and (12a).

2.3.3 Corpora for Abstract Anaphora Resolution

There is a number of lexical, semantic, and syntactic properties associated with abstract anaphora (see Section 2.3.2). This makes Abstract Anaphora Resolution (AAR) a great challenge from a computational perspective. Moreover, identifying the exact boundaries of antecedents is difficult for humans as well. Botley (2006) points out that the difficulty of identifying abstract antecedents is due to the lack of clear boundaries as well as the complex or unclear inference process for finding antecedents. Artstein and Poesio (2006) conducted an annotation experiment on the TRAINS91 dialog corpus (Allen and Heeman, 1995) with 20 inexperienced annotators. The annotators agreed with the most popular choice of the beginning of the antecedent in only 42% of the cases and in 64% of the cases for the ending. That is, annotators often disagree on the exact boundaries of antecedents but they agree more on the ends of the segments than on their beginning. Due to all these reasons, mostly small-scaled corpus studies have been carried out to investigate AAR (Kolhatkar et al., 2018, p. 74). Therefore, there are only a few corpora for resolution of abstract anaphors that are interesting from a machine learning perspective. We will describe them in more detail in the rest of this subsection. This corpora is the platform for experiments in Chapter 4. We report the number of anaphors across these corpora in Table 4.2 on page 87.

The ASN and CSN corpora. Kolhatkar et al. (2013a,b) annotated the Anaphoric Shell Nouns (ASN) corpus and the Cataphoric Shell Nouns (CSN) corpus for shell noun resolution.

The CSN corpus. As already mentioned in Section 2.3, Kolhatkar et al. (2013b) distinguish two types of shell nouns: (i) Anaphoric Shell Nouns (ASNs) whose antecedent occurs anywhere in the prior discourse (the case we are interested in), and (ii) Cataphoric Shell Nouns (CSNs) whose antecedent occurs in the same sentence in an embedded position following the shell noun. The important difference between ASNs and CSNs is that antecedents of CSNs can be extracted using simple predefined rules based on the syntactic structure and the categorization of shell nouns in the literature. Kolhatkar et al. (2013b) made a good use of this property of CSNs to create training data for resolving shell nouns. They apply the predefined rules to extract the embedded antecedent of CSNs, e.g. the pattern "*<shell noun> that*" in (14a).

- (14) a. **Sentence with the pattern *<shell noun> that*:** Congress has focused almost solely on the fact that special education is expensive - and that it takes away money from regular education.
- b. **Anaphoric sentence:** Congress has focused almost solely on this fact.

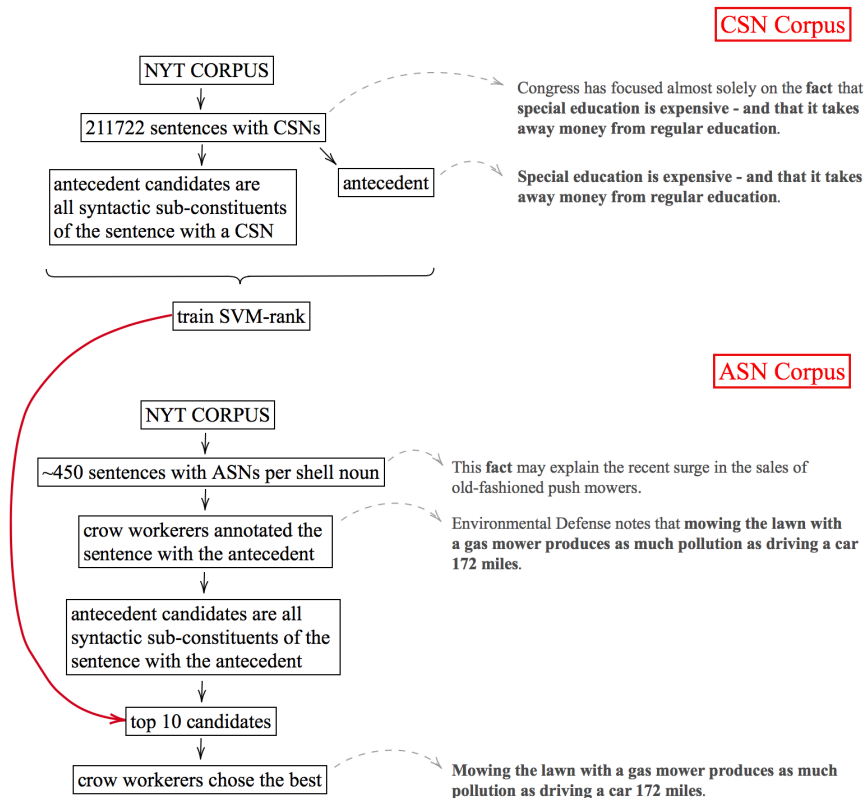


Fig. 2.4 The creation of the ANS and CSN corpora (Kolhatkar et al., 2013b).

- c. **Antecedent:** Special education is expensive - and that it takes away money from regular education.

They extracted such pairs of antecedents and sentences with CSNs from the New York Times (NYT) corpus (Sandhaus, 2008) for six shell nouns: *decision*, *fact*, *issue*, *possibility*, *reason*, and *question*. Using this data, they trained six ranking SVM models for six shell nouns. These SVM models rank candidates for the antecedent such that the highest ranked candidate is the predicted antecedent. All syntactic constituents of the constructed anaphoric sentence as well as constituents of the extracted antecedent were considered as candidates for training a ranking model. They assume that the ranking SVM models trained on the CSN corpus will also be able to perform resolution of ASNs since CSN antecedents and ASN antecedents share linguistic properties. We provide details of the ranking SVM model in Section 2.3.4. The described construction of the CSN corpus is illustrated with the upper part of Figure 2.4.

The ASN corpus. The ASN corpus was also constructed from the NYT corpus by selecting anaphoric instances with the pattern "*this* <shell noun>" for six selected shell nouns. Kolhatkar et al. (2013a) crowdsourced annotations for the sentence which contains the antecedent which

they refer to as a *broad region*. Candidates for the antecedent were obtained again by using all syntactic constituents of the broad region. They ranked them using the ranking SVM model trained on the CSN corpus. The top 10 ranked candidates were presented to the crowd workers and they chose the best answer that represents the ASN antecedent. The workers were encouraged to select *None* when they did not agree with any of the displayed answers and they could provide information about how satisfied they were with the displayed candidates. We consider this dataset as gold even though it may be biased toward the offered candidates. Figure 2.4 illustrates the creation of the ASN corpus and how it is connected to the CSN corpus.

OntoNotes and the CoNLL-12 shared task dataset. OntoNotes (Pradhan et al., 2007) is a five year multi-site collaboration between BBN Technologies, Information Sciences Institute of University of Southern California, University of Colorado, University of Pennsylvania and Brandeis University. OntoNotes provides rich annotations including syntactic parse, predicate-argument structure, coreference, and word senses linked to an ontology. As a result, two CoNLL shared tasks are organized based on the dataset (Pradhan et al., 2012). However, they annotated only a simple form of abstract anaphora, i.e. event anaphora. Jauhar et al. (2015) collected pronouns (*this, that, it*) from the CoNLL-12 shared task dataset whose preceding mention in the coreference cluster is a verb phrase. They used this subset to train a feature-based classifier. We provide details of their classifier in Section 2.3.4.

The ARRAU Corpus. The ARRAU corpus (Uryupina et al., 2016; Poesio et al., 2018) is one of very few anaphoric annotation projects that have attempted annotating abstract anaphora in its entirety (Artstein and Poesio, 2006; Dipper and Zinsmeister, 2012). The most distinctive feature of the corpus is the annotation of a wide range of anaphoric relations, including bridging references and abstract anaphora in addition to coreference. We use instances from the WSJ subpart of the ARRAU corpus that are marked with the category *abstract* or *plan*. The ARRAU corpus is too small to assist the training of a neural model. However, since the ARRAU corpus is challenging due to a mixture of nominal and pronominal anaphors as well as a great range of confounders, it is an excellent dataset for evaluation. Therefore, we use the ARRAU corpus to evaluate our models in resolution of *unrestricted* abstract anaphora and provide the first benchmark results on this corpus.

2.3.4 Computational Approaches to AAR and Related Tasks

Thus far we have already mentioned a few reasons why it is challenging to apply computational methods to resolve abstract anaphors automatically. First, the search space of

Document-level features	
1	document number
NP-level features	
2	grammatical function of the antecedent (subject, object, other)
3	form of antecedent (definite NP, indefinite NP, personal pronoun, demonstrative pronoun, possessive pronoun, proper name)
4	antecedent's agreement in person, gender, and number
5	semantic class of antecedent (human, concrete object, abstract object)
6	grammatical function of anaphor (subject, object, or other)
7	form of antecedent (definite NP, indefinite NP, personal pronoun, demonstrative pronoun, possessive pronoun, proper name)
8	anaphor's agreement in person, gender, and number
9	semantic class of anaphor (human, concrete object, abstract object)
Coreference-level features	
10	distance between anaphor and antecedent in words
11	distance between anaphor and antecedent in sentences
12	distance between anaphor and antecedent in markables
13	indicator whether anaphor and antecedent have the same grammatical function (yes or no)
14	indicator whether anaphor and antecedent consist of identical strings (yes or no)
15	indicator whether one string contains the other (yes or no)

Table 2.5 Features for coreference resolution in Strube et al. (2002).

non-nominal antecedent candidates is large. Second, there is a number of lexical, semantic, and syntactic properties associated with abstract anaphora. Third, non-nominal antecedents are not always clearly delimited and identifiable stretches of text. Fourth, for all these reasons, identifying the exact boundaries of antecedents is difficult even for humans and therefore the labeled data for unrestricted abstract anaphora resolution is quite scarce.

As a result, there are only a few attempts to resolve abstract anaphors using machine learning. These approaches differ significantly from the models proposed in this work in two respects. First, the prior work makes assumptions about the anaphor's semantic or syntactic type as well as about the semantic type of the antecedent. Second, the prior work proposed feature-based models, while in this thesis we use neural models capable of learning the relevant features from data. In the rest of this subsection, we review prior machine learning approaches. For the overview of rule-based methods (e.g. Eckert and Strube, 2000; Byron, 2004; Pappuswamy et al., 2005) consult Kolhatkar et al. (2018), p. 81–87.

Coreference features. Strube et al. (2002) investigated how to design a feature space for the resolution of anaphoric expressions in general, including definite noun phrases (i.e., entities that are identifiable in a given context, in English commonly marked with *the*, *this*,

syntactic type of the candidate	
1	fine-grained syntactic type (e.g. NP-TMP, RRC)
2	coarse-grained syntactic type (e.g. NP, VP, S, PP)
context of the candidate	
3	syntactic type of left and right siblings of the candidate
4	part-of-speech tag of the preceding and following words of the candidate
embedding level of the candidate within its sentence	
5	top embedding level (i.e. with respect to its top clause (the root node))
6	immediate embedding level (i.e. with respect to the closest ancestor of type S or SBAR)
subordinating conjunctions	
7	indicator whether the candidate follows the pattern SBAR -> (IN sconj) (S ...)
verb features of the candidate	
8	presence of verbs in general
9	whether the main verb is finite or non- finite
10	presence of modals
length of the candidate	
11	length of the candidate in words
12	relative length of the candidate with respect to the sentence containing the antecedent
lexical features of the candidate	
13–62	occurrence of the 50 highly ranked unigrams for specific shell noun

Table 2.6 Features for shell noun resolution in Kolhatkar et al. (2013b).

every, and *both* determiner), proper names as well as personal, possessive, and demonstrative pronouns (*this*, *that*). They started with a standard coreference feature set (Table 2.5). A closer examination of the performance for different forms of anaphoric expressions showed that a decision tree classifier performed poorly on definite NPs and demonstrative pronouns. Since definite NPs occur more frequently than demonstrative pronouns in their corpus, they developed better features for definite NPs. Since definite NPs can not be abstract anaphors as they refer to entities, resolution of definite NPs is not addressed in this thesis.

Shell noun resolution. Among machine-learning based resolution systems for non-nominal antecedents the most prominent is work by Kolhatkar et al. (2013b) on resolution of shell nouns. In particular, they addressed resolution of six shell nouns: *decision*, *fact*, *issue*, *possibility*, *reason*, and *question*. To circumvent the training data bottleneck they extracted training data automatically using the specific properties and categorization of Cataphoric Shell Nouns (CSNs). We described the CSN corpus in Section 2.3.3. For each shell noun, Kolhatkar et al. (2013b) trained a separate *ranking* SVM model using instances of that shell noun in the CSN corpus. Candidates for the antecedent are all syntactic constituents of the sentence that contains the CSN as well as constituents of the extracted antecedent. The highest ranked candidate is the predicted antecedent. Their features (Table 2.6) consider

the syntactic type of the candidate, the context of the candidate, the embedding level of the candidate within its sentence, the presence of subordinating conjunctions in the candidate, properties of verbs that occur in the candidate, length of the candidate, and its lexical features. Note that the context of the shell noun is never used. The ranking SVM models trained on the CSN corpus are then used to resolve anaphoric shell nouns in the ASN corpus. That is, the ASN corpus serves only as the test set and never as a training set. We train our models on the CSN dataset, evaluate them on the ASN dataset, and compare with the reported result in Kolhatkar et al. (2013b). We also train our models on the ASN dataset and evaluate on the ARRAU dataset which we use exclusively for evaluation due to its small size.

Event coreference. A system for event coreference needs to determine which event mentions in a text refer to the same real-world event. In Example (15), such a system needs to detect that *had hanged* and *suicide* refer to the same event of *Lo Presti killing himself*.

(15) Police said Lo Presti had hanged himself. His suicide appeared to be related to clan feuds.

Models for this task can be trained on relatively large data sets with thousands of event coreference chains (Lu and Ng, 2017). The crucial difference between event coreference and unrestricted abstract resolution is that event coreference is restricted to a subclass of events and typically focuses on coreference between verb (phrase) and noun (phrase) mentions of similar abstractness levels. In contrast, abstract anaphora typically involves a non-nominal antecedent that is referred to by a highly abstract noun or a simple pronoun. In this thesis we look only in specific events described in the following paragraph.

Resolving *this*, *that*, and *it*. Jauhar et al. (2015) (following Müller, 2007) propose a two-stage feature-based approach that first classifies pronouns as abstract or concrete and then selects their antecedents accordingly, where antecedents can occur both before the anaphor (anaphoric pronouns) or after (cataphoric pronouns). Their model is trained on the subset of the CoNLL-12 shared task data described in Section 2.3.3 using features presented in Table 2.7. We train and evaluate our models on the part of this subset of the CoNLL-12 shared task data with anaphoric pronouns. We do not compare our results to results reported in Jauhar et al. (2015) since they also resolve cataphoric pronouns and therefore our test sets differ significantly.

Sluice resolution. Sluice or *wh*-fronted ellipses are questions where the specification of what is asked for (beyond the *wh*-word) is elided (and thus needs to be retrieved from context).

feature	description	selected
pronoun word	word of p	-
demonstrative	p is <i>this</i> or <i>that</i>	-
parent and label	lemma of parent and dependency label of p	•
pronoun path	dependency label path of p to root	-
sentence distance	number of sentences between v and p	•
token distance	log-distance between v and p in tokens	•
verb distance	number of verbs between v and p	-
relative position	v precedes p (anaphora or cataphora)	•
direct dominance	v is the immediate parent of p	•
dominance	v is an ancestor of p	•
candidate path	dependency label path of v to root	•
negated candidate	v is negated	•
candidate transitivity	transitivity of v (i.e. it has a child with a direct object label)	•
clause-governing candidate	probability of v to govern a clause	-
right frontier	v is in the right frontier of p	•
I-incompatibility	attribute of p is a <i>non-concrete</i> adjective	•
verb association strength	NPMI between v and parent verb of p	-
selectional preference	preference between v and parent verb of p	-

Table 2.7 Features for pronoun p and candidate v for the pronoun resolution in Jauhar et al. (2015). Features marked with • were selected, and those marked with - were discarded by feature selection.

For example, in (16a) the missing sentential complement of the wh-phrase *why* is understood to mean *why did he resort to that*. This is an example of an *embedded sluice*. Example (16b) is a *root sluice* because the wh-word is not embedded in a larger structure.

(16) a. He resorted to that. I don't know why.

b. A: Jennifer is looking for you. B: Why?

Anand and Hardt (2016) proposed a feature-based model for sluice resolution. They evaluated their model on a corpus of 3103 embedded sluices from news articles (Anand and McCloskey, 2015). Recently, Rønning et al. (2018b) were the first to apply a neural model for sluice resolution. They also annotated 2000 examples¹¹ from the OpenSubtitles corpus (Tiedemann, 2009). In Rønning et al. (2018a), they investigate the linguistic knowledge that models presented in Rønning et al. (2018b) learned implicitly. We leave sluice resolution for the future work.

¹¹1000 examples are root sluices, and 100 are embedded sluices.

2.4 Neural Multi-Task Learning and Adversarial Training

Multi-Task Learning (MTL) is a machine learning paradigm for solving two or more related tasks by leveraging the information shared between them. It is particularly helpful for the over-parameterized machine learning regime where the number of training examples is fewer than the number of parameters in the model. In other words, MTL acts as a powerful regularizer for solving tasks with limited labeled data using models with many parameters such as popular deep learning models which build on neural networks.

The synergy between MTL and neural networks dates back to at least Caruana (1997). Since neural models are easy to engineer to jointly solve multiple tasks, MTL is successfully applied to different machine learning application areas such as computer vision (Zhang et al., 2014), speech recognition (Deng et al., 2013b), and NLP (Collobert et al., 2011), *inter alia*.

The main challenge of successfully applying MTL is establishing what the related tasks are and how to share the network's parameters between them. As a result, MTL can generally be divided into two major categories: *hard parameter sharing* and *soft parameter sharing*. Hard parameter sharing stands for manually selecting which parts of the network are shared between tasks and which are task-specific. For example, the most widely used hard parameter sharing technique shares layers that produce a fixed-size dense vector representation of the input and separates task-specific layers that make final predictions. However, forcing learning features relevant for all tasks might be harmful to the learning features specific to a particular task. Thus, soft parameter sharing techniques are proposed which have only task-specific parameters but the loss at the final layer is regularized with the distance between the certain parts of the main and the auxiliary task parameters. The distance may be defined as the L_2 distance (Duong et al., 2015) or the trace norm (Yang and Hospedales, 2017).

In the rest of the section, we introduce different deep learning approaches for producing a fixed-size dense vector to represent text and the most prominent hard-parameter sharing architectures. We do not provide details of how these models are trained. For more information, consult Goodfellow et al. (2016), Goldberg (2017), or Ruder (2017).

2.4.1 Representing Text with Supervised Deep Learning

In recent years deep learning has improved performance on many kinds of machine learning problems in NLP such as classification (Kim, 2014; Tang et al., 2015), tagging (Collobert et al., 2011), and ranking (Clark and Manning, 2016), to name a few.

The basis of solving these problems with deep learning is initializing the first layer of a neural network with pre-trained word embeddings and learning a fixed-size dense vector to represent the input text without the use of external resources such as dependency parsers,

named entity recognizers, or lexica. Due to this deep learning is easily applicable to other languages, domains, and genres under the assumption that labeled data is available.

The most prominent networks for learning text representations are Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). RNNs process text token by token and hence are suitable for representing sequential data. They have a feedback recurrent connection between adjacent tokens which feeds back outputs of the model at the current token back into itself and consequently constraints the output at the next token on the previous output. On the other hand, CNNs are initially intended for processing a grid of values such as an image. They are known for employing a simple mathematical operation called *convolution* in place of the general matrix multiplication in at least one of their layers. The most important property of convolution is that it allows for the detection of appropriate features irrelevant of their position in the input (i.e. spatial invariance).

In the rest of this section we describe in more detail the learning of word embeddings and the workings of RNNs and CNNs. A reader experienced in deep learning may skip this part.

Word representation. How do we convert words to a numerical form in order to present them to a machine learning model? The strong recent trend is to use *word embeddings*¹²—dense vectors whose relative similarities correlate with semantic similarity. The training of models that produce word embeddings is based on the so-called *distributional hypothesis* (Harris, 1954) which states that linguistic items with similar distributions have similar meanings, i.e. words in similar contexts have similar meanings. They are popularized with the release of the WORD2VEC toolkit¹³(Mikolov et al., 2013a,b) which was the first to enable easy training and the use of pre-trained word embeddings.

WORD2VEC can utilize two model architectures to produce a distributed representation of words: *continuous bag-of-words* or *continuous skip-gram*. In the continuous bag-of-words architecture, the model predicts the current word from a window of surrounding context words. The order of context words does not influence prediction (bag-of-words assumption). In the continuous skip-gram architecture the model uses the current word to predict the surrounding window of context words.

Nowadays using pre-trained word embeddings "off-the-shelf" and initializing the first layer of neural networks with them is a key building block of developing a deep learning model when dealing with natural language, or in Christopher Manning's words: "WORD2VEC

¹²Alternative terms are *distributional semantic vectors*, *distributed representations*, or *semantic vectors*.

¹³<https://code.google.com/archive/p/word2vec/>

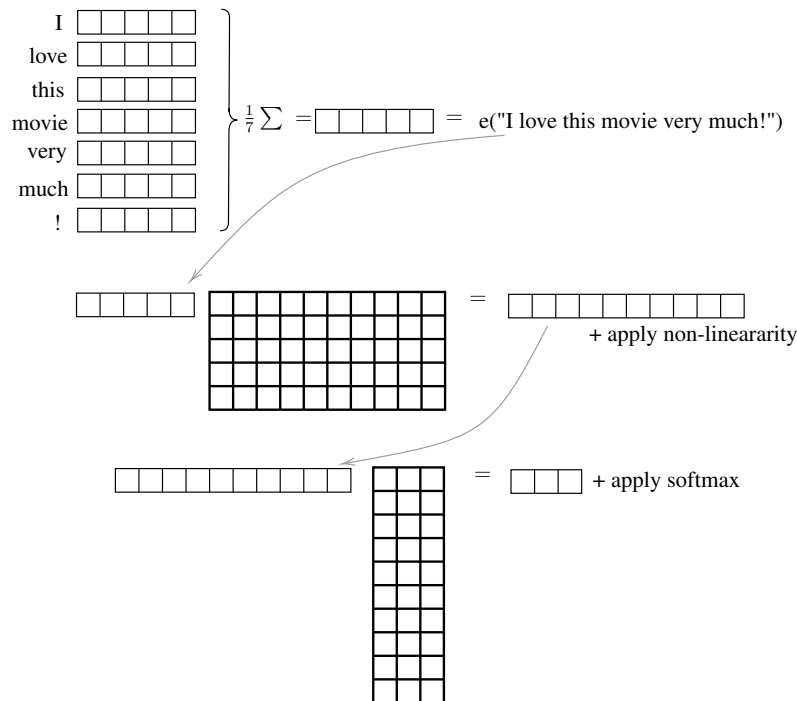


Fig. 2.5 Feed-Forward Neural Network (FFNN).

is the sriracha sauce of deep learning!"¹⁴. For more information see Sebastian Ruder's description¹⁵ and Goldberg (2017), p. 89–105.

Feed-Forward Neural Networks (FFNNs). FFNNs or Multi-Layer Perceptrons (MLPs) are the basic deep learning models. FFNNs are called feed-forward because information flows from the input to the output through the intermediate computations without feedback connections in which outputs of the model are fed back into itself.

Figure 2.5 illustrates a FFNN in a matrix form. Imagine we want to classify the sentence *I love this movie very much!* into three classes: positive, negative, and neutral. First, since the input sentences may vary in length, we construct a fixed-sized vector of the input sentence by averaging the embeddings of the words that occur in the sentence: $\{e(\text{I}), e(\text{love}), e(\text{this}), e(\text{movie}), e(\text{very}), e(\text{much}), e(!)\}$. Then we multiply the resulting vector by the 5×10 dimensional weight matrix where 5 is the word embedding size and 10 is the hidden vector size. Then we apply a non-linear *activation* function to the resulting vector and produce a hidden representation of the input. The most successful and widely popular activation function is the Rectified Linear Unit (ReLU) i.e. $\sigma(\mathbf{x}) = (\max\{0, x_1\}, \dots, \max\{0, x_n\}) \in \mathbb{R}^n, \mathbf{x} \in \mathbb{R}^n$. The hidden representation is multi-

¹⁴<https://nlp.stanford.edu/manning/talks/NAACL2015-VSM-Compositional-Deep-Learning.pdf>

¹⁵<http://ruder.io/word-embeddings-1/>

plied by a 10×3 dimensional output weight matrix where 3 is the number of classes. Finally, we apply the softmax function to this vector and produce a probability distribution over three different possible class values. The output of the softmax function is used to calculate the cross-entropy loss and the gradients that are needed in the tuning of the network's parameters (the matrices in bold).

Recurrent Neural Networks (RNNs). A FFNN's expressive capacity is limited and forces us to disregard the word order. CNNs offer some sensitivity to order but they are still restricted to mostly local patterns and they disregard the order of patterns that are far apart in the sequence. RNNs (Elman, 1990) allow representing arbitrarily sized sequential inputs in a fixed-size vector, while paying attention to the global word order. For example, to finish the sentence *I grew up in France, I speak fluent with French* the model needs to remember the mention of *France*. In theory, RNNs are absolutely capable of handling such long-term dependencies, but in practice they do not seem to be able to learn them due to the vanishing gradient problem (Hochreiter, 1991; Bengio et al., 1994). For this reason gated RNNs such as a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and a Gated Recurrent Unit Network (GRU) (Cho et al., 2014) are proposed. They typically incorporate memory mechanisms to be capable of learning long-term dependencies. See Christopher Olah's description¹⁶ for a good visualization of RNNs and LSTMs.

Uni- and Bi-Directional Long Short-Term Memory Networks (LSTMs). All RNNs have the form of a chain of repeating modules of a single neural network layer. LSTMs also have this chain-like structure, but instead of applying a single neural network layer at each token, LSTMs apply three neural network layers that interact in a very special way.

These special neural network layers are called *gates*: forget gate, input gate, and output gate. They regulate the amount of information to add or remove from the so-called *cell state* which serves as a memory. Gates are composed of a sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}} \in \langle 0, 1 \rangle$, $x \in \mathbb{R}$, and the pointwise multiplication operation. The output of a gate is a scalar between 0 and 1, where 0 stands for completely remove and 1 for completely add.

The forget gate decides what information is going to be thrown away from the cell state

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \in \langle 0, 1 \rangle^n, \quad (2.6)$$

where n is the dimension of the cell state vector and the sigmoid function is applied element-wise.

¹⁶<http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

The input gate decides which values to update. We use the tanh function $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \in \langle -1, 1 \rangle$, $x \in \mathbb{R}$, to create a vector of new candidates values \tilde{C} for the cell state. Then we update the old cell state C_{t-1} into the new cell state C_t as follows,

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_f) \in \langle 0, 1 \rangle^n, \quad (2.7)$$

$$\tilde{C} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \in \langle -1, 1 \rangle^n, \quad (2.8)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C} \in \mathbb{R}^n, \quad (2.9)$$

where \odot is used for the pointwise multiplication operation and the tanh function is applied element-wise.

Finally, we need to decide what to output. First, a sigmoid function decides which parts of the cell state we will output,

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \in \langle 0, 1 \rangle^n. \quad (2.10)$$

Then a tanh function pushes the values of the cell state between -1 and 1. Finally, the output gate multiplies the squashed cell state with the output of the sigmoid function. This way we output only the parts we choose,

$$h_t = o_t \odot \tanh(C_t) \in \langle -1, 1 \rangle^n. \quad (2.11)$$

Presently it is common to process the input sequence both starting from the beginning (forward) and starting from the end (backward). This ensures that the information from both the past and the future is captured. Representations from the forward and backward pass are usually concatenated and passed to the next layer. This kind of an LSTM is called a *bi-directional* LSTM (Graves and Schmidhuber, 2005). Moreover, it is common to stack a few LSTM layers such that the output of one LSTM is the input to the next LSTM. This kind of an LSTM is called a *deep* LSTM (Zhou and Xu, 2015).

Convolutional Neural Networks (CNNs). Although CNN models are designed for processing a grid of values such as an image, they have been shown to achieve excellent results for many NLP tasks: semantic parsing (Yih et al., 2014), sentence modeling (Kalchbrenner et al., 2014), question answering (Dong et al., 2015), event extraction (Chen et al., 2015), modal sense classification (Marasović and Frank, 2016), language modeling (Pham et al., 2016), semantic role labeling (Marcheggiani and Titov, 2017), text generation (Semeniuta et al., 2017), dependency parsing (Yu and Vu, 2017), machine translation (Gehring et al., 2017), and aspect-based sentiment analysis (Xue and Li, 2018), *inter alia*.

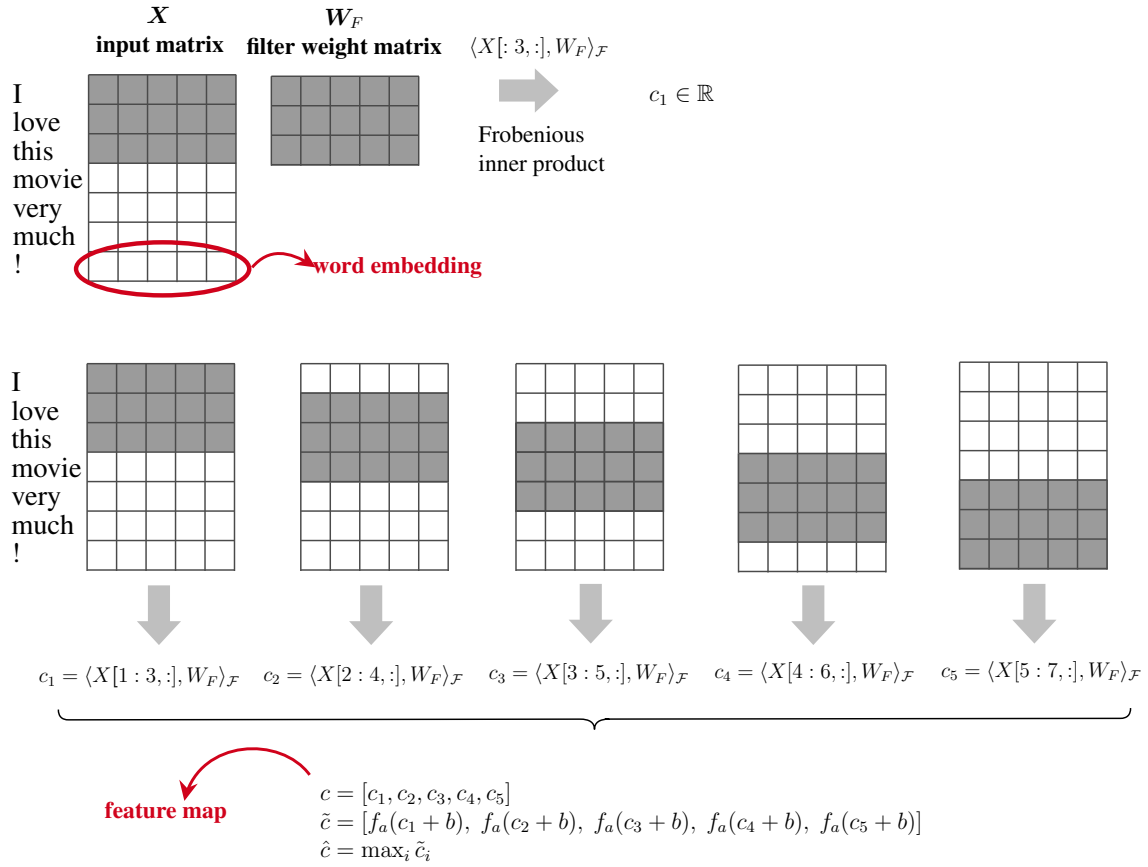


Fig. 2.6 Convolution.

The input to a CNN is a matrix $X \in \mathbb{R}^{s \times d}$ whose rows correspond to d -dimensional word embeddings of words in the sentence of length s . Based on the input layer, a CNN builds up one or more convolutional layers. A convolution is an operation between sub-matrices of the input matrix $X \in \mathbb{R}^{s \times d}$ and a filter parametrized by a weight matrix $W_F \in \mathbb{R}^{n \times d}$, where n is the so-called *filter size*. Convolution returns a vector that is usually referred to as a *feature map*. The convolution is illustrated in Figure 2.6.

Formally, let $X[i : i + n - 1, :]$ be the sub-matrix of the input matrix X from the i -th row to the $(i + n - 1)$ -th row and let $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ denote the sum of elements of the element-wise inner product of two matrices, known as the Frobenius inner product. The i -th component of the feature map $c \in \mathbb{R}^{s-n+1}$ is obtained by taking the Frobenius inner product of the sub-matrix $X[i : i + n - 1, :]$ and the filter weight matrix W_F ,

$$c_i = \langle X[i : i + n - 1, :], W_F \rangle_{\mathcal{F}} = \sum_{k,l} X[i : i + n - 1, :]_{kl} \cdot W_{F_{kl}}, \quad (2.12)$$

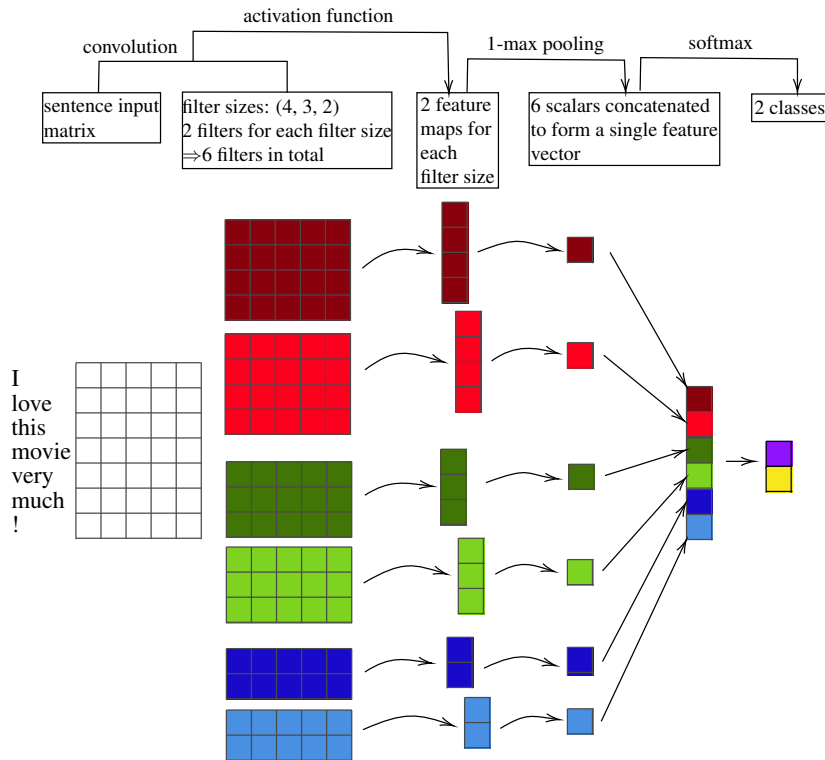


Fig. 2.7 The Convolutional Neural Network for sentence classification. Figure replicated from Zhang and Wallace (2017).

for $i \in \{1, \dots, s - n + 1\}$. Afterward, a bias term $b \in \mathbb{R}$ is added to every component of the feature map and an activation function f_a is applied to obtain a new feature map \tilde{c} , where $\tilde{c}_i = f_a(c_i + b)$. Finally, *max-over-time pooling* (Collobert et al., 2011), is applied over a single feature map that extracts the maximum value $\hat{c} = \max_i \tilde{c}_i \in \mathbb{R}$, which represents the most important feature detected with the corresponding filter.

We have described the process by which one feature $\hat{c} \in \mathbb{R}$ is extracted from one filter. However, in practice we use multiple filters with different filter sizes n , resulting in multiple feature maps. Features obtained through max-pooling from each feature map are concatenated to a vector representation of the input sentence that is passed to the softmax layer (see Figure 2.7). Parameters to learn are elements of the filter matrices.

Filters are trained to be especially active when they encounter a sequence of words relevant for the given classification task. Kalchbrenner et al. (2014) present lists of n-grams (detected by different filters) that capture positive or negative sentiment phrases as well as more abstract semantic categories, such as negation or degree particles (e.g. *too*) that are relevant in compositional sentiment detection. For the modal sense classification task, we showed that the feature maps capture semantic categories found to be relevant in prior work,

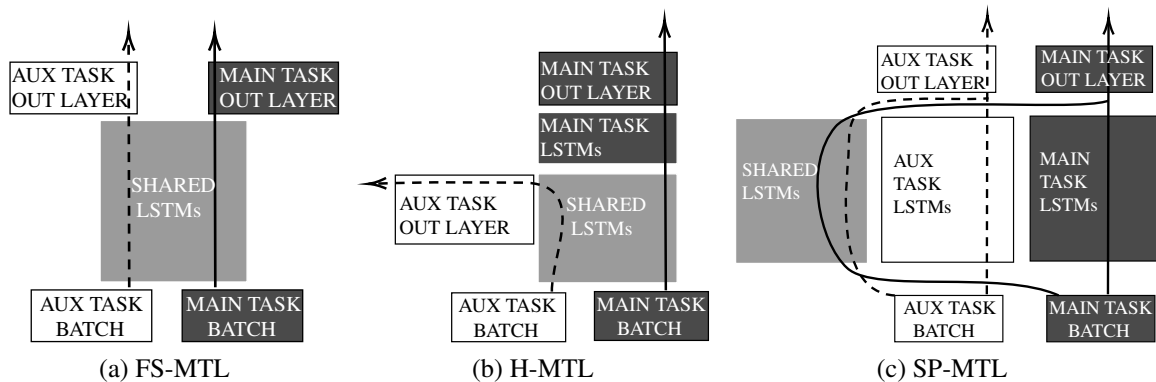


Fig. 2.8 Hard-parameter sharing architectures.

such as tense, negation, and semantic properties of verbs and phrases as well as features that model the wider syntactic context, especially subject and embedded verb, and their semantics (abstractness, semantic class, aspect, tense) (Marasović and Frank, 2016).

Although CNNs achieve great results for many NLP tasks and allow a faster training using parallel and distributed strategies (unlike RNNs), LSTMs are still a more frequent choice for a text representation learner in NLP.

2.4.2 Hard-Parameter Sharing

Here we review the most prominent hard-parameter sharing techniques and use a deep LSTM as a textual representation learner. We illustrate all techniques with figures. Shared parameters are illustrated in light grey, the auxiliary task’s parameters in white, and the main task’s parameters in dark grey. The forward pass from the input to output of the auxiliary task is illustrated with a dashed line and for the main task with a solid line. These lines may be used to trace which parts of the network are changed by a data batch of the auxiliary task and which parts are changed by a data batch of the main task. We assume that the reader is familiar with the gradient descent optimization algorithm; if needed consult Goodfellow et al. (2016), p. 79–83.

The Fully Shared MTL model (FS-MTL). FS-MTL (Caruana, 1997; Collobert et al., 2011) shares all parameters of the general model between the main and the auxiliary task except the output layer (Figure 2.8 (a)). The main task’s output layer is never tuned using the auxiliary data, and the other way around.

The Hierarchical MTL model (H-MTL). For NLP applications, often some given (high-level) task is supposed to benefit from another (low-level) task more than the other way

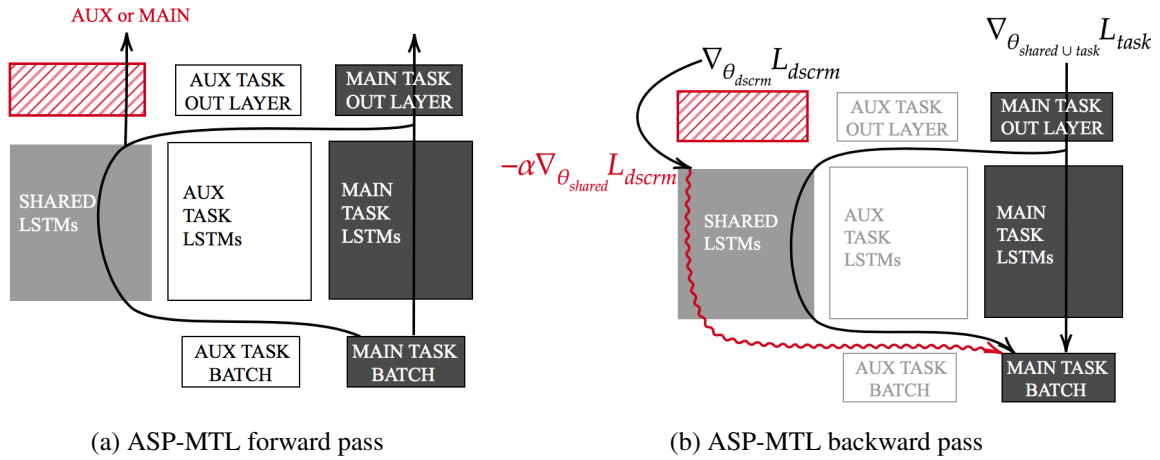


Fig. 2.9 Adversarial training of the SP-MTL model.

around, e.g. chunking from POS tagging (Shen and Sarkar, 2005). This intuition leads to designing hierarchical MTL models (Søgaard and Goldberg, 2016; Hashimoto et al., 2017) in which predictions for low-level tasks are not made on the basis of the representation produced at the final LSTM, but on the representation produced by a lower-layer LSTM (see Figure 2.8 (b)). The task-specific layers atop the shared layers enable the network to capture task-specific features while being aware of the knowledge shared between tasks. This architecture can be seen as a way to combine multi-task and cascaded learning.

The Shared-Private MTL model (SP-MTL). The key idea behind the SP-MTL model is to fully separate shared and task-specific (private) parameters. As a result, the model may decide on its own whether the shared or the task-specific information is relevant for the given instance. To be precise, in the SP-MTL model in addition to the stack of shared LSTMs, each task has a stack of task-specific LSTMs (Liu et al., 2017) (Figure 2.8 (c)). Representations at the outermost shared LSTM and the task-specific LSTM are concatenated and passed to the task-specific output layer.

Adversarial Training. The limitation of the SP-MTL model is that it does not prevent the shared layers from capturing task-specific features. To ensure this, a *task discriminator* is integrated in the SP-MTL model (Ganin and Lempitsky, 2015; Liu et al., 2017). The task discriminator predicts to which task the current batch of data belongs based on the representation produced by the shared LSTMs (the hashed rectangle in Figure 2.9). If the shared LSTMs are indeed task-invariant the discriminator will perform badly. Thus, we update the shared parameters such that they maximize the discriminator’s cross-entropy loss.

Since the direction of the positive gradient points to the local minimum, we approach a local maximum by making a gradient descent step in the direction of the negative gradient (the wavy line in Figure 2.9 (b)). At the same time we want the discriminator to challenge the shared LSTMs, so we update the discriminator’s parameters to minimize its cross-entropy loss, i.e. we make a step in the direction of the positive gradient (the leftmost solid line in Figure 2.9 (b)). This minmax optimization is known as *adversarial training* and recently it has gained a lot of attention for NLP applications (Liu et al., 2017; Chen et al., 2017; Kim et al., 2017; Qin et al., 2017; Wu et al., 2017; Gui et al., 2017; Li et al., 2017; Zhang et al., 2017; Joty et al., 2017; Yasunaga et al., 2018; Chen and Cardie, 2018; Cai and Wang, 2018; Kang et al., 2018; Kurita et al., 2018).

Chapter 3

Sentence-Level Neural Fine-Grained Opinion Analysis

Sentence-level Fine-Grained Opinion Analysis (FGOA) aims to extract and label *opinion-holder-target* structures from a given sentence to answer: "Who expressed what kind of sentiment toward what?". This task is crucial for understanding opinionated text (Hovy, 2011) and therefore solving applications such as sentiment inference, opinion summarization, and opinion-oriented question answering. We focus on detecting opinion holders and targets i.e. Opinion Role Labeling (ORL) since it is a vital but less explored opinion analysis task (see Section 2.2 in Chapter 2). Our goal is to (i) investigate the limitations of neural ORL models, and (ii) to gain a better understanding of what is solved and what is next for ORL. We focus on neural models since they are adaptive to heterogeneous sources because they do not depend on external resources typically available only for English and for certain domains.

In this chapter we investigate the first research question: can we improve neural opinion role labeling models by using Multi-Task Learning (MTL) with a related task which has substantially more data, i.e., Semantic Role Labeling (SRL), even though there are divergences in the annotation schemes of opinion and semantic role labeling? We empirically show it is beneficial to use MTL with SRL.

In the NAACL-HLT paper "SRL4ORL: Improving Opinion Role Labeling Using Multi-Task Learning with Semantic Role Labeling" (Marasović and Frank, 2018), we discuss divergences in the annotation schemes of opinion and semantic role labeling, different MTL techniques, and we integrate them in an ORL model that predicts opinion holders and targets given a gold (oracle) opinion expression. This chapter represents an extension to this paper. In particular, we provide more results to compare different MTL techniques.

The chapter is organized as follows. First we illustrate that SRL cannot solve ORL for all cases although it is a good MTL candidate. Furthermore, we demonstrate how specificities,

predicate	predicate-argument structure
said:	[A0: Australia] [V: said] [A1: it feared violence if voters thought the election had be stolen] .
feared:	Australia said [A0: it] [V: feared] [A1: violence if voters thought the election had be stolen] .
thought:	Australia said it feared violence if [A0: voters] [V: thought] [A1: the election had be stolen] .
had:	Australia said it feared violence if voters thought the election [V: had] be stolen .
be:	Australia said it feared violence if voters thought the election had [V: be] stolen .
stolen:	Australia said it feared violence if voters thought [A1: the election] had be [V: stolen] .

Table 3.1 Output of the AllenNLP SRL demo (He et al., 2017) for the sentence *Australia said it feared violence if voters thought the election had be stolen*.

inconsistencies and the incompleteness of the MPQA annotations could be an obstacle to properly exploit SRL data with MTL (Section 3.1). Then, in Section 3.2 we discuss how well-received MTL techniques outlined in Chapter 2 (Section 2.4) could adapt to different cases of ORL from Section 3.1. In Section 3.3 we describe the experimental setup, datasets, evaluation metrics, hyperparameters, and other training details. We then evaluate and compare different MTL approaches (Section 3.4). Additionally, we analyze what has been solved and what is next for ORL (Section 3.5). Finally, we give a summary of the chapter (Section 3.6).

3.1 Similarities and Divergences in the Opinion and Semantic Role Labeling Annotation Schema

Semantic Role Labeling (SRL) is the task of predicting predicate-argument structure of a sentence, which answers the question: "Who did what to whom, where and when?". Automatic semantic role labeling may be useful to identify opinion holders and targets in context, since holders and targets of many lexical items that are typically used as opinion expressions can be mapped to semantic roles. For example, consider the following sentence and, if necessary, refer to the description of the MPQA annotation scheme on page 26.

(17) Australia said [it]_H^{A0} [**feared**]_{O_{neg}} [violence]_T^{A1} if voters thought the election had been stolen.

Table 3.1 illustrates its predicate-argument structure according to the output of the SRL demo¹ (a reimplementation of He et al. (2017)) following the PropBank SRL scheme (Palmer et al., 2005). The semantic roles (agent A0, theme A1) of the predicate *fear* (marked blue bold) overlap with the opinion roles H and T, according to MPQA. Since this is not an isolated case, the output of SRL systems has been commonly used for feature-based FGOA

¹<http://demo.allennlp.org/semantic-role-labeling/NDA4MjYz>

models (Kim and Hovy, 2006; Johansson and Moschitti, 2013; Choi et al., 2006; Yang and Cardie, 2013). Additionally, a considerable amount of training data is available for training SRL models (see Table 3.3 on page 62). This resulted in the successful training of neural SRL models (Zhou and Xu, 2015; He et al., 2017; Peng et al., 2018).

Although SRL is similar in nature to labeling of *opinion roles* (i.e. holders and targets), it cannot solve Opinion Role Labeling (ORL) for all cases (Ruppenhofer et al., 2008). In Example (18) holder and target of the predicate *please* correspond to theme (A1) and agent (A0) semantic roles respectively, whereas for the predicate *fear* in (17) holder and target correspond to agent (A0) and theme (A1) respectively. We took into account this observation when deciding on an appropriate MTL model by splitting its parameters into shared and task-specific categories.

- (18) [I]_H^{A1} am very [**pleased**]_{O_{pos}} that [the Council has now approved the Kyoto Protocol thus enabling the EU to proceed with its ratification]_T^{A0}.

A further obstacle for properly exploiting SRL training data could be specificities, inconsistencies and the incompleteness of the MPQA annotations. In Example (19), *Rice* expressed his negative sentiment toward *the three countries in question* by *setting the criteria* which states something negative about those countries: *they are repressive and grave human rights violators, and aggressively seeking weapons of mass destruction*. In this case, the model should not pick any local semantic role for the target.

- (19) [The criteria]_H^{A1} [**set by**]_{O_{neg}} [Rice]_H^{A0} are the following: [the three countries in question]_T are repressive and grave human rights violators, and aggressively seeking weapons of mass destruction.

In Examples (20–21), the same opinion expression *concerned* realizes different scopes for the target. A model which exploits SRL knowledge could be biased to always label targets as complete SRL role constituents, as in example (21).

- (20) Rice told us [the administration]_H^{A1} was [**concerned**]_{O_{neg}} that [[Iraq]_T would take advantage of the 9/11 attacks]_{A0}.

- (21) [The Chinese government]_H^{A1} is deeply [**concerned**]_{O_{neg}} about [the sudden deterioration in the Middle East situation]_T^{A0}, Tang said.

Regarding the incompleteness, prior work (Katiyar and Cardie, 2016) has shown that their model makes reasonable predictions in sentences which do not have annotations at all,

e.g. [mothers]_H [care]_O for [their young]_T, in: *From the fact that mothers care for their young, we can not deduce that they ought to do so, Hume argued.*

Examples discussed in this section demonstrate that leveraging SRL knowledge with Multi-Task Learning (MTL) is a reasonable idea. At the same time, they alert us that given the specificities of MPQA and ORL annotations in general, it is not obvious whether MTL can overcome divergences in the annotation schemes of opinion and semantic role labeling. We investigate this research question by adopting one of the recent successful SRL neural models (Zhou and Xu, 2015) and experimenting with different MTL frameworks.

3.2 Neural MTL for SRL and ORL

Neural MTL currently receives a lot of attention and new MTL architectures emerge regularly. However, there is no clear consensus which MTL architecture to use in which conditions. In this section we discuss well-received architectures outlined in Chapter 2 (Section 2.4) that could adapt to different cases of ORL from Section 3.1.

As a general neural architecture for single- and multi-task learning we use the SRL model of Zhou and Xu (2015) (henceforth, Z&X-STL, where STL stand for "Single-Task Learning" in contrary to Multi-Task Learning) which successfully labels semantic roles without any syntactic guidance.² This model consists of a stack of bi-directional LSTMs and a CRF layer which makes the final prediction. Every sentence is processed as many times as there are predicates in it (see Table 3.2). The inputs to the first LSTM are word embeddings as well as three additional features: embedding of the predicate, the continuous bag-of-words representation of the context of the predicate, and an indicator feature (1 if the current token is in the predicate context, 0 otherwise). Adapting this model for labeling of opinion roles is straightforward, the only difference being that opinion expressions can be multi-words and only two opinion roles are assigned: holder (H) and target (T).

As previously mentioned in Chapter 2 (Section 2.4), the standard approach to MTL is to establish in advance which layers should have tied parameters and which should be task-specific (i.e. hard-parameter sharing). In the following text we will discuss the most prominent models outlined in Chapter 2 (Section 2.4.2) in the context of ORL.

Since a **Fully-Shared (FS-MTL)** model shares all parameters of the general model except the output layer, this model should be effective for constructions with a clear mapping between opinion and semantic roles such as $\{A0 \mapsto H, A1 \mapsto T\}$ as in Example (17).

²New neural SRL models are proposed (He et al., 2017; Peng et al., 2018) since we started this work. Bingel and Sjøgaard (2017) show that MTL works when the main task has a flattening learning curve, but the auxiliary task curve is still steep. We notice such behavior in our learning curves. Therefore the best SRL model does not benefit us since we do not even need to train an SRL model to convergence.

predicate	supervision														
	Australia	said	it	feared	violence	if	votes	thought	the	election	had	be	stolen	.	
said:	B-A0	V	B-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	O
feared:	O	O	B-A0	V	B-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	O
thought:	O	O	O	O	O	O	B-A0	V	B-A1	I-A1	I-A1	I-A1	I-A1	I-A1	O
had:	O	O	O	O	O	O	O	O	O	O	V	O	O	O	O
be:	O	O	O	O	O	O	O	O	O	O	O	V	O	O	O
stolen:	O	O	O	O	O	O	O	O	B-A1	I-A1	O	O	V	O	O

Table 3.2 Oversampling for the Z&X-STL model (Zhou and Xu, 2015). The sentence *Australia said it feared violence if voters thought the election had be stolen* is oversampled six times for six different predicates that occur in it. Each row illustrates the label sequence of the corresponding predicate.

The **Hierarchical MTL (H-MTL)** model is based on the assumption that for NLP applications often some high-level task is supposed to benefit from another low-level task more than the other way around. This is the case for ORL (high-level task) and SRL (low-level task). For this reason, in this MTL model only the first two LSTMs are shared and the final LSTM is ORL-specific. Task-specific layers atop shared layers could potentially give the model more power to distinguish or ignore certain semantic roles. If so, the H-MTL model is more suitable for examples like (18) and (19).

The final MTL model is motivated by observations from different cases in Section 3.1. For example, in (17) it is sufficient to know the semantic roles. In (18) it is useful to know potential semantic roles but realize that in this ORL case the simple mapping $\{A0 \mapsto H, A1 \mapsto T\}$ is incorrect. Hence, we need knowledge relevant for both tasks. To the contrary, only ORL knowledge is relevant for Example (19). For this reason, in the **Shared-Private MTL (SP-MTL)** model, in addition to the stack of shared LSTMs, each task has a stack of task-specific LSTMs. As noted in Chapter 2 (Section 2.4), the limitation of the SP-MTL model is that it does not prevent the shared layers from capturing task-specific features. To ensure this, we train the SP-MTL model with *adversarial training* that we described in detail on page 55 in Chapter 2 (Section 2.4.2).

3.3 Experimental Setup

3.3.1 Datasets

For SRL we use the newswire CoNLL-2005 shared task dataset (Carreras and Màrquez, 2005), annotated with PropBank (Palmer et al., 2005) predicate-argument structures. Sections 2–21 of the WSJ corpus are used for training and section 24 as the development set. The test set consists of section 23 of WSJ and 3 sections of the Brown corpus.

	task	train size	dev size	test size	$ \mathcal{Y} $
CoNLL'05	SRL	90750	3248	6071	106
MPQA (4-CV)	ORL	3141.25	1055	1036.75	7
MPQA (10-CV)	ORL	3516.3	1326	349.3	7

Table 3.3 Datasets with the number of SRL predicates/ORL opinions in train, dev & test set, size of label inventory.

For ORL we use the manually annotated MPQA 2.0 corpus (Wiebe et al., 2005; Wilson, 2008) described in Chapter 2 (Section 2.2.1). We report detailed pre-processing of MPQA³ and data statistics in the Appendix.

3.3.2 Training Settings

We evaluate our models using two evaluation settings. First, we follow Katiyar and Cardie (2016) which set aside 132 documents for development and used the remaining 350 documents for 10-fold CV. However, in the 10-fold CV setting, the size of the test sets is 3 times smaller than the dev set size (Table 3.3, row 3), and, consequently, one run of 10-fold CV results in high-variance estimates on the test sets. If we run 10-fold CV at least 10 times we get a reasonable estimate. However, training a neural model 100 times is expensive and unfortunately the common practice nowadays is to report only a single run. We additionally evaluate our models with 4-fold CV which we repeat twice. We set aside 100 documents for development and use 25% of the remaining documents for testing. The resulting test sets are comparable in size to the dev set (Table 3.3, row 2). We run a 4-fold CV twice with two different random seeds.

We do not tune hyperparameters (HPs), but follow suggestions proposed in the thorough HP study for sequence labeling tasks (Reimers and Gurevych, 2017).

3.3.3 Evaluation Metrics

For both tasks we adopt evaluation metrics from prior work. For SRL, precision is defined as the proportion of semantic roles predicted by a system which are correct, recall is the proportion of gold roles which are predicted by a system, and F1 score is the harmonic mean of precision and recall. We do not report performance for SRL.

³Examples how to use our scripts can be found at https://github.com/amarasovic/naacl-mpqa-srl4orl/blob/master/generate_mpqa_jsons.py.

In case of ORL, we report 10-fold CV⁴ and repeated 4-fold CV with *binary F1 score* and *proportional F1 score* for holders and targets separately. *Binary precision* is defined as the proportion of predicted holders (targets) that overlap with the gold holder (target). *Binary recall* is the proportion of gold holders (targets) for which the model predicts an overlapping holder (target):

$$\text{binary precision} = \frac{1}{|\mathcal{H}_P|} \sum_{h_p \in \mathcal{H}_P} \mathbb{1}_{\{h_p \text{ overlaps with the gold holder}\}}, \quad (3.1)$$

$$\text{binary precision} = \frac{1}{|\mathcal{H}_G|} \sum_{h_g \in \mathcal{H}_G} \mathbb{1}_{\{h_g \text{ overlaps with a predicted holder}\}}, \quad (3.2)$$

where \mathcal{H}_P is the set of predicted holders and \mathcal{H}_G of the gold holders.

Proportional recall measures the proportion of the overlap between a gold holder (target) and an overlapping predicted holder (target). *Proportional precision* measures the proportion of the overlap between a predicted holder (target) and an overlapping gold holder (target).

$$\text{proportional precision} = \frac{1}{|\mathcal{H}_P|} \sum_{h_p \in \mathcal{H}_P} \frac{\# \text{tokens}(\text{overlap}(h_p, h_g))}{\# \text{tokens}(h_p)}, \quad (3.3)$$

$$\text{proportional recall} = \frac{1}{|\mathcal{H}_G|} \sum_{h_g \in \mathcal{H}_G} \frac{\# \text{tokens}(\text{overlap}(h_p, h_g))}{\# \text{tokens}(h_g)}. \quad (3.4)$$

F1 scores are the harmonic means of the corresponding precision and recall.

3.3.4 Hyperparameters

Input representation. We used 100d GloVe word embeddings (Pennington et al., 2014) pre-trained on Gigaword and Wikipedia and did not fine-tune them. For MTL models, vocabulary was built from all the words in the training data of both tasks, and out-of-vocabulary words were replaced with an UNK token. The embedding of the context of a predicate or an opinion is the average of the embeddings of the predicate (or the opinion phrase), 2 preceding words, and 2 words after.

Weights initialization. The size of all LSTM hidden states was set to 100. The number of the backward and the forward LSTM layers is set to 3, which counts for 6 LSTM layers in Z&X-STL. Z&X-STL achieved circa 2% higher SRL F1 score with 8 LSTM layers, but such a deep model would cause overfitting on the small-sized ORL data. In the H-MTL model, SRL is supervised at the second LSTM layer. We initialized the LSTM weights with random

⁴We used the same splits as the prior work (Katiyar and Cardie, 2016). We thank the authors for providing the splits.

	dev (MPQA)				test (MPQA)				
	holder		target		holder		target		
	binary F1	prop. F1	binary F1	prop. F1	binary F1	prop. F1	binary F1	prop. F1	
Z&X-STL	80.15 _{1.10}	76.87 _{1.26}	74.62 _{0.67}	70.23 _{1.04}	Z&X-STL	80.24 _{2.91}	77.98 _{2.90}	76.30 _{2.55}	71.18 _{2.55}
FS-MTL	83.68 _{0.44}	81.45 _{0.58}	76.23 _{0.75}	73.01 _{0.93}	FS-MTL	83.47 _{2.26}	81.80 _{2.26}	77.60 _{2.52}	73.77 _{2.28}
H-MTL	84.14 _{0.72}	81.86 _{0.48}	76.11 _{0.61}	72.55 _{0.73}	H-MTL	84.03 _{2.65}	82.34 _{2.51}	77.41 _{2.14}	73.10 _{1.96}
SP-MTL	82.18 _{0.89}	79.66 _{0.72}	74.99 _{1.17}	71.32 _{1.81}	SP-MTL	82.19 _{2.49}	80.11 _{2.36}	76.01 _{3.03}	71.51 _{3.34}
ASP-MTL	82.63 _{0.84}	80.20 _{0.99}	74.24 _{0.58}	70.16 _{1.29}	ASP-MTL	83.15 _{2.92}	81.12 _{2.66}	75.89 _{2.66}	71.21 _{2.78}

Table 3.4 ORL 10-fold CV results.

orthogonal matrices (Henaff et al., 2016), and all other weight matrices with the so-called *He initialization* (He et al., 2015). LSTM forget biases were initialized with ones (Jozefowicz et al., 2015), all other biases with zeros.

Optimization. We trained our model in mini-batches of size 32 using Adam (Kingma and Ba, 2015) with the learning rate of 10^{-3} . For MTL we alternate batches from different tasks. We clip gradients by global norm (Pascanu et al., 2013) with a clipping value set to 1. Single-task models were trained for 10K iterations and MTL models for 20K. One epoch counts for $\lceil \frac{\text{train size}}{\text{batch size}} \rceil$ iterations. We stop training if the arithmetic mean of proportional F1 scores of holders and targets is not improved in 25 epochs. For the minmax optimization we use a gradient reversal layer (Ganin and Lempitsky, 2015). The discriminator’s cross-entropy loss is scaled with 0.1.

Regularization. Variational dropout (Gal and Ghahramani, 2016) with a keep probability $k_p \in 0.85$ was applied to the outputs and the recurrent connections of the LSTMs. Standard dropout (Srivastava et al., 2014) was applied to the output classifier weights with a keep probability $k_p \in 0.85$ and to the input embeddings with $k_p \in 0.7$.

The code for training and evaluating our models can be found at <https://github.com/amarasovic/naacl-mpqa-srl4orl>.

3.4 Experiments

3.4.1 Evaluating Benefits of MTL for ORL

This section describes the evaluation of different Multi-Task Learning (MTL) models proposed in Section 3.2. We evaluate all models after every $\lceil \frac{\text{train size}}{\text{batch size}} \rceil$ iteration on the ORL dev set and save them if they achieve a higher arithmetic mean of proportional F1 scores of holders and targets on the ORL dev set. The saved models are used for testing. We report the mean of F1 scores over 10 folds and the standard deviation (appears as a subscript) of all models in Table 3.4. We report the mean of F1 scores over 4 folds and 2 different seeds (8

	dev (MPQA)				test (MPQA)				
	holder		target		holder		target		
	binary F1	prop. F1	binary F1	prop. F1	binary F1	prop. F1	binary F1	prop. F1	prop. F1
Z&X-STL	79.73 _{1.19}	77.06 _{1.14}	76.09 _{0.94}	70.45 _{1.07}	Z&X-STL	80.42 _{1.92}	77.48 _{2.06}	73.84 _{1.17}	67.03 _{2.13}
FS-MTL	83.58 [•] _{0.69}	82.16 [•] _{0.59}	78.32 [•] _{1.57}	75.09 [•] _{2.27}	FS-MTL	83.67 [•] _{1.52}	81.59 [•] _{1.50}	77.04 [•] _{1.45}	73.01 [•] _{2.53}
H-MTL	82.36 [◇] _{0.81}	80.84 [◇] _{0.98}	78.11 [•] _{0.82}	74.89 [•] _{1.33}	H-MTL	82.80 [•] _{1.87}	80.40 [•] _{1.91}	77.12 [•] _{1.34}	73.16 [•] _{1.78}
SP-MTL	82.21 [◇] _{0.79}	80.23 [•] _{0.88}	76.14 [◇] _{1.18}	71.14 [◇] _{0.97}	SP-MTL	82.51 [•] _{2.17}	80.03 [•] _{2.00}	74.61 [◇] _{1.32}	68.70 [◇] _{2.32}
ASP-MTL	81.41 [◇] _{1.27}	79.39 [•] _{1.45}	76.49 _{1.39}	72.13 [•] _{1.87}	ASP-MTL	81.77 [◇] _{1.74}	79.32 [◇] _{1.62}	74.92 [•] _{0.84}	69.89 [•] _{1.80}

Table 3.5 ORL repeated 4-fold CV results.

evaluations), as well as the standard deviation of all models in Table 3.5. We mark significant difference between MTL models and the single-task (Z&X-STL) model, observed using a Kolmogorov-Smirnov significance test ($p < 0.05$) (Massey Jr, 1951), with [•] in superscript and between the FS-MTL model and other MTL models with [◇].

Single-Task Learning (STL) vs. Multi-Task Learning (MTL). In the 10-fold CV evaluation setting (Table 3.4), the FS-MTL and the H-MTL models improve over the Z&X-STL model in all evaluation measures, for both holders and targets. When evaluated in the repeated 4-fold CV setting (Table 3.5), all MTL models improve over the Z&X-STL model in all evaluation measures, for both holders and targets.

The FS-MTL and the H-MTL models improve *significantly* in all evaluation measures, for both holders and targets, on both dev and test sets, when evaluated with repeated 4-fold CV. With 10-fold CV the improvements are also significant, except for targets on the test set. This is probably due to the small size of the test sets (Table 3.3, row 3), which results in a high-variance estimate. Indeed, standard deviations on the 10-fold CV test sets are always much higher compared to the development set or to the test sets of 4-fold CV.

It is not surprising that larger improvements are visible in the labeling of holders. They are usually short, less ambiguous, and often presented with the A0 semantic role, whereas annotating targets is a challenging task even for humans.⁵

Larger improvements are visible for proportional F1 score than for binary F1 score. That is, more data and SRL knowledge helps the model to better detect the scope of opinion roles.

Comparing MTL models. In Section 3.2 we introduced MTL models with task-specific LSTM layers hypothesizing that these layers should give MTL models more power to adapt to a variety of potentially problematic cases that we illustrated in Section 3.1. However, our results show that the FS-MTL model performs significantly better or comparable to MTL

⁵Wilson (2008) reports annotator agreement for target labeling of 86.00 binary F1 score.

models that include task-specific LSTM layers. Reimers and Gurevych (2017) show that MTL is especially sensitive to the selection of hyperparameters (HPs). Thus, a firm and solid comparison of the different MTL models requires thorough HP optimization, to properly control the number of parameters and the regularization of the models.

Therefore, we randomly sample 20 HP configurations and use each configuration to evaluate FS-MTL, H-MTL, and SP-MTL with one run of 4-fold CV on the development set. Sampled HPs are reported in Table 3.6; h_{LSTM} stands for the LSTM hidden size, g_{clip} for the gradient clip value, o_w for the minimum occurrence of a vocabulary word, k_c , k_i , and k_o for the keep probability of the cell state, the input, and the output, respectively, n_{task} for the number of task-specific LSTM layers, n_{share} for the number of shared LSTM layers, b for the batch size, and w for the window size.

Figure 3.1 illustrates a violin plot for each model. In each violin plot, the white dot represents the median of 20 evaluations, the thick gray bar in the center represents the interquartile range, and the thin gray line represents the 95% confidence interval. Wider sections of a violin plot represent a higher probability that the corresponding model will result in the given value; the skinnier sections represent a lower probability. All violin plots are wider in the middle indicating that the results are concentrated around the median. However, among all MTL models SP-MTL is the most concentrated around its median while H-MTL has the longest tail. The results mostly follow the insights from Table 3.5. The exception is comparison between H-MTL and SP-MTL for targets. We observe that SP-MTL performs slightly better for targets with respect to the binary F1 score and comparable with respect to the proportional F1 score. This observation is not consistent with results in Table 3.5. This evaluation confirms that the simplest MTL model performs the best and more sophisticated MTL models (H-MTL, SP-MTL) are mutually comparable. We speculate why FS-MTL performs the best in two ways. First, although we have showcased that SRL cannot solve ORL for all cases it could be the case that these scenarios are under-represented in the MPQA corpus. Second, since FS-MTL is designed such that it learns a representation that is equally good for both tasks, it is a stronger regularizer and therefore generalizes better.

3.5 Analysis of What Works and What is Next

The central topic of this section is to analyze what the proposed models are good at, in which ways MTL improves over the single-task ORL model, and what could be done to achieve further progress.

We evaluate the FS-MTL and the Z&X-STL models on the ORL dev set using 4-fold CV repeated twice with different seeds (8 evaluation trials). We say that a model predicts a role

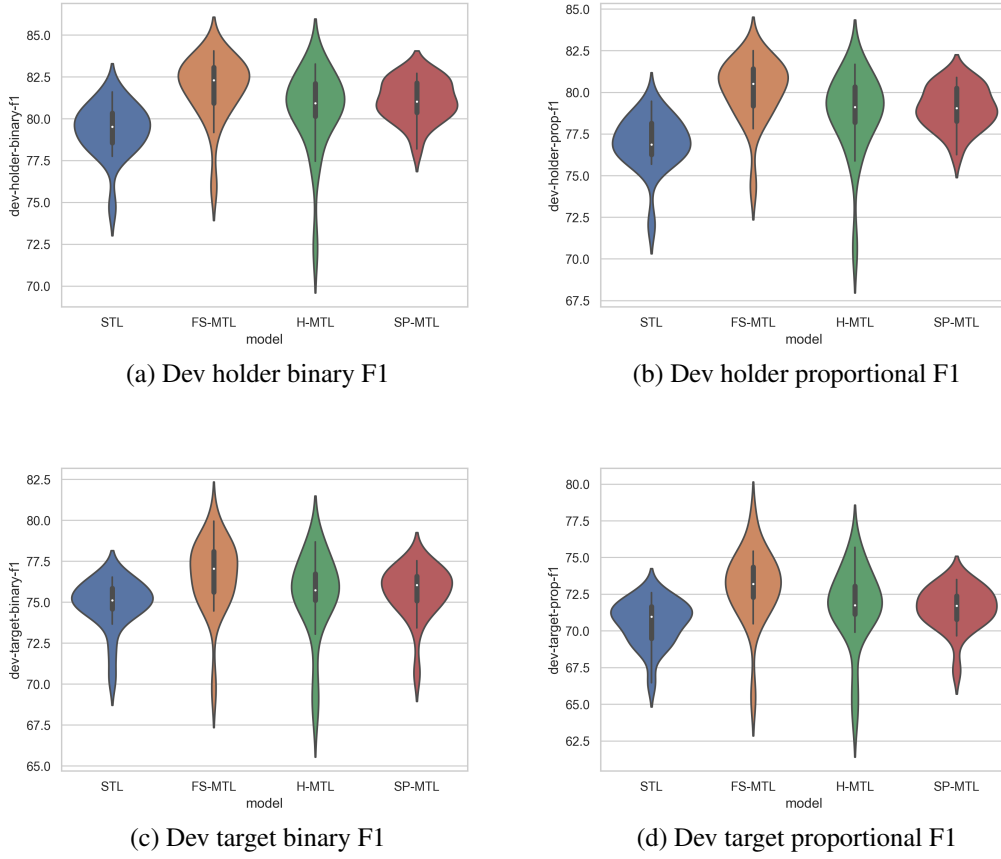


Fig. 3.1 Comparison of MTL models using violin plots.

id	h_{LSTM}	g_{clip}	o_w	k_c	k_i	k_o	n_{task}	n_{share}	b	w
1	61	88.73	2	0.74	0.98	0.96	3	1	6	2
2	44	6.38	1	0.79	0.94	1.00	1	3	6	1
3	105	3.83	2	0.75	0.79	0.77	3	2	44	2
4	33	19.58	1	0.78	0.72	0.86	2	1	43	2
5	57	87.03	1	0.84	0.82	0.90	3	3	23	2
6	36	5.14	2	0.73	0.54	0.69	1	2	8	2
7	49	76.52	1	0.84	0.79	1.00	3	2	4	2
8	110	18.23	3	0.89	0.53	0.67	1	3	7	1
9	103	36.73	3	0.82	0.86	0.73	1	2	3	1
10	117	46.92	3	0.73	0.74	0.95	3	1	36	1
11	54	76.51	3	0.98	0.82	0.88	2	1	5	2
12	101	68.67	3	0.69	0.92	0.65	1	3	12	2
13	79	61.24	3	0.99	0.78	0.94	3	2	22	2
14	71	68.22	3	0.95	0.82	0.75	3	3	22	1
15	111	76.14	2	0.73	0.80	0.77	3	3	19	1
16	97	78.08	2	0.81	0.79	0.90	2	3	19	1
17	103	21.97	1	0.66	0.77	0.70	1	1	10	2
18	119	58.14	1	0.75	0.55	0.87	1	3	45	2
19	123	57.17	1	0.81	0.78	0.87	1	1	20	2
20	104	51.32	1	0.74	0.75	0.88	1	1	10	1

Table 3.6 Randomly sampled hyperparameters for comparing MTL models.

1	Malinga _{FS,ZX} said according to the guidelines in the booklet, the election had been legitimate .
2	movie um-hum that 's interesting so that was a good movie too well do you _{FS,ZX} think we've covered baseball i think so okay well have a good night
3	The nation _{FS,ZX} should certainly be concerned about the plans to build a rocket launch pad , work on the infrastructure for which is due to start in 2002 , with launches beginning from 2004 .
4	Bam on Sunday said she _{FS,ZX} believed Zimbabwe's election was not free and fair , adding they were not in line with international standards as well as those of her organisation .
5	The majority report , endorsed only by the ANC , said the observer mission _{FS,ZX} had noted that over three million Zimbabweans had cast their votes and this substantially represented the will of the people .

Table 3.7 The dev examples for which both models (FS-MTL, Z&X-STL) *correctly* predict the *holder* in 6/8 trials.

	easy	hard
% opinions that are predicates	91.32	93.33
% holders that are subjects	77.84	38.79
% holders that are A0 roles	74.10	33.33
avg. distance between holders & opinions	1.54	7.56

Table 3.8 Statistics of *holder* prediction.

of a given opinion expression *correctly* if the model predicts a role that overlaps with the correct role in at least 6 out of 8 evaluation trials. If a model predicts a role that overlaps with the correct role in at most 2 out of 8 trials, we say that the model predicts the role *incorrectly*. The requirement on 6-8 (in)correct predictions reduces the risk of analyzing inconsistent predictions and enables us to draw firmer conclusions. We analyze the following scenarios:

- (i) Both the FS-MTL model and the Z&X-STL model make correct predictions (Tables 3.7 and 3.9).
- (ii) Both models make wrong predictions (Tables 3.11 and 3.12).
- (iii) The FS-MTL model makes a correct prediction, while the Z&X-STL makes an incorrect prediction (Tables 3.13 and 3.14).

In the following, we categorize predictions in case (i) as *easy cases*, and predictions in case (iii) as *hard cases*. In Tables 3.7, 3.9, 3.11, 3.12, 3.13, and 3.14 the opinion expression is bolded, the correct role is italicized, predictions of the FS-MTL model are colored blue (subscript *FS*), predictions of the Z&X-STL model are colored yellow (subscript *ZX*), and green marks predictions where both models agree. For simplicity, we show only holders or targets, although the models predict both roles jointly.

1	Indonesia has come under pressure from several quarters to <i>take tougher action against alleged terrorist leaders</i> but has played down the threat <small>ZXF_S</small> .
2	Mugabe even talked about his desire to <i>keep safeguarding Zimbabwe 's sovereignty and land</i> <small>ZX</small> <i>in spirit</i> <small>FS</small> when he dies , a dream which the veteran leader said forced him to sacrifice a bright teaching career in the 1950s to lead [...].
3	Under his blueprint , the government hopes to <i>stabilize the economy</i> through curtailing state expenditure , reforming public enterprises and expanding agriculture <small>FS,ZX</small> .
4	He said those who thought the election process would be rigged were supporters of the MDC party , adding that they were prejudging and wanted to <i>direct the process</i> <small>FS,ZX</small> .
5	People in the rural areas support the ruling party because our party has been genuine on its policy on <i>land reform</i> <small>FS,ZX</small> .

Table 3.9 The dev examples for which both models (FS-MTL, Z&X-STL) *correctly* predict the *target* in 6/8 trials.

	easy	hard
% opinions that are predicates	92.58	89.20
% target's heads that are objects	22.12	14.77
% targets that are A1 roles	70.62	42.61
% targets that are A2 roles	9.00	0.57
avg. distance between targets & opinions	2.29	8.46

Table 3.10 Statistics of *target* prediction.

What works well? There are 668/1055 instances in the dev set for which both models predict holders correctly, and 663/1055 for targets.

Examples 1–5 in Table 3.7 suggest that holders that can be properly labeled by both models (*easy* cases) are subjects of their governing heads or agent (A0) semantic roles. The statistics in Table 3.8 (col. 1, rows 2–3) supports this observation.⁶ In contrast, holders that both models predict incorrectly (*hard* cases) are less frequently subjects or agent (A0) roles (col. 2, rows 2–3). Also, *easy* holders are close to the corresponding opinion expression: the average distance is 1.54 tokens (Table 3.8, row 4), contrary to the *hard* holders with the average distance of 7.56.

Examples 1–5 in Table 3.9 suggest that targets that can be properly labeled by both models are objects of their governing heads or theme (A1) semantic roles. Table 3.10, row 3, shows that the majority of the *easy* targets are indeed A1 roles, in contrast to the *hard* targets. Similar to holders, the *easy* targets are on average 7 tokens closer to the opinion expression.

⁶The statistics is calculated using the output of mate-tools (Björkelund et al., 2010).

1	It would be entirely improper if , in <i>its defense of Israel</i> _{FS} , the United States continues to exert pressure on [...] .
2	Indonesia _{FS,ZX} has come under pressure from <i>several quarters</i> to take tougher action against alleged terrorist leaders but has played down the threat .
3	Australia should adhere to the <i>Cardinal Principle of International Law</i> , which states that all nations in the world must first respect and promote the humanitarian interests and progress of all humankind .
4	<i>The department</i> said that it will cost \$ 600 for an HIV/AIDS patient per year at this time , and the following years this cost is expected to stand at just \$ 400/year for one patient as the production of such drugs becomes stable .
5	The Organisation of African Unity OAU _{ZX} also backed Zimbabwean President Robert Mugabe 's re-election , with its observer team _{FS,ZX} describing the poll as " transparent , credible , free and fair " .
6	Regarding the American proposed Anti-Missile Defense System too , neither Russia , China , Japan , nor even the European Union , had shown any enthusiasm ; rather they _{FS} had all _{FS,ZX} expressed their reserves on the project .
7	The president renewed his pledge to thwart terrorist groups _{FS,ZX} <i>who want to</i> " mate up " with regimes hoping to acquire weapons of mass destruction and said " nations will come with us " if the US-led war on terrorism is extended .

Table 3.11 The dev examples for which both models (FS-MTL, Z&X-STL) *incorrectly* predict the *holder* in 6/8 trials.

1	<i>State-sanctioned land invasions</i> , several times declared illegal by Zimbabwe 's courts , as well as a drought have disrupted Zimbabwe 's food production and famine is already looming in much of the country .
2	But he told the nation _{FS,ZX} that in spite of <i>stiff opposition to the agrarian reforms from powerful Western countries , especially the country 's former colonial power of Britain</i> , he would press ahead to seize farms from whites and [...] .
3	If the Europeans wish to influence Israel in the political arena – in <i>a direction</i> that many in Israel would support wholeheartedly – they will not be able to promote their positions in such a manner .
4	They _{FS,ZX} are fully aware that these are dangerous <i>individuals</i> , he said during a press conference [...] .
5	And her little girl just complained , " I don't want to <i>wash the dishes</i> " .
6	During <i>President Bush's speech</i> , I thought of heckling _{ZX} : What are you going to do with the Kyoto Protocol? _{FS}
7	At first I didn't want to apply for it _{FS,ZX} , but the principal called me during the summer months and said , " Sandra the time is running out , you need to apply " .

Table 3.12 The dev examples for which both models (FS-MTL, Z&X-STL) *incorrectly* predict the *target* in 6/8 trials.

What to do for further improvement? There are 165/1055 instances in the dev set for which both models predict holders incorrectly, and 176 for targets.

As we have seen so far, many holders that are subjects or agent (A0) semantic roles and targets that are theme (A1) semantic roles are properly labeled by both models. However, a considerable amount of such holders and targets are not correctly predicted (Table 3.8–3.10, col. 2, rows 2–3). Thus our models do not work flawlessly for all such cases. A distinguishing property of the *hard* cases is the distance of the role from the opinion. Thus, future work should advance the model’s ability to capture long-range dependencies.

Examples in Table 3.11 demonstrate that holders that the FS-MTL has difficulty capturing occur with the corresponding opinions in more complicated syntactic constructions. In the first example, the FS-MTL model does not recognize the possessive and is possibly biased towards picking the country (*Israel*) which occurs immediately after the opinion. In the second example, the opinion expression is a nominal predicate and the holder is its object. The sentence is in passive voice but the models probably interpret it in the active voice and thus make the wrong prediction. In the third example, the opinion expression is the head of the relative clause that modifies the holder. These observations raise the following question: would training a neural dependency parser together with SRL help the ORL model to handle syntactically harder cases?

Example 4 shows that holders specific to the MPQA annotation schema are hard to label since they require inference skills: from *the department said*, we can defeasibly infer that it is *the department who expects [this cost] to stand at just \$400/year [...]*. To handle such cases, it would be worth trying to train a model to recognize textual entailment together with the ORL model.

Examples 6–7 illustrate that some gap in performance stems from the difficulties in processing MPQA. Example 5 has no gold holder, but the models make plausible predictions. For example 6, FS-MTL predicts the discontinuous holder *they ... all*, while MPQA allows only contiguous entities. Therefore our evaluation scripts interpret *they* and *all* as two separate holders and deem *all* as incorrect which results in a lower precision. Finally, for example 7 our models make plausible predictions. However, the gold holder is always the entity from the coreference cluster that is the closest to the opinion.⁷ The evaluation scripts needs to be extended such that predicting any entity from the coreference cluster is considered to be correct. In conclusion, to better evaluate future developments, it would be worth curating MPQA instances with missing roles and extending the evaluation scripts to account for coreferent holders and discontinuous roles.

⁷We followed the prior work (Katiyar and Cardie, 2016).

The examples in Table 3.12 demonstrate that difficulties in labeling the targets originate from similar reasons as for the holders. Examples 1–3 demonstrate complex syntactic constructions, examples 4–6 MPQA-specific annotations that require inference and example 7 exemplifies a missing target.

1	Yoshihisa Murasawa , a management consultant for Booz-Allen & Hamilton Japan Inc. , said his firm _{FS,ZX} will likely be recommending acquisitions of Japanese companies more _{ZX} often to foreign clients in the future .
2	The source _{FS} , interviewed by Interfax in Grozny , expressed confidence that that the command of the Russian forces in Chechnya would soon “ be able to obtain documentary confirmation ” that Khattab was dead .
3	The Commonwealth team earlier this week _{FS} said that " the conditions in Zimbabwe did not adequately allow the free and fair expression of will by the electorate " .
4	Publishing such biased reports will only create mistrust among nations _{FS} regarding the objectives and independence of the UN Commission on Human Rights .
5	The Inkatha Freedom Party , Democratic Alliance , New National Party , African Christian Democratic Party , the Pan Africanist Congress and the United Christian Democratic Party _{ZX} had disagreed with the ANC _{FS} conclusion .
6	The Nigerian leader , President Olusegun Obasanjo _{ZX} , had urged the minister _{FS,ZX} not to attack Blair frontally over Britain ’s negative position regarding Zimbabwe , but to deal [...] .
7	US diplomats _{ZX} say Bush _{FS,ZX} will seek to support Kim ’s Nobel Prize winning policy by offering new talks with the North , while remaining firm about North Korea ’s missile sales and its feared chemical and biological weapons programmes.

Table 3.13 The dev examples for which the FS-MTL model *correctly* predicts the *holder* in 6/8 trials, whereas the Z&X-STL model predicts *incorrectly* in 6/8 trials.

1	In most cases he described the legal punishments _{FS} <i>like floggings and executions of murderers and major drug traffickers that are applied based on the Shria , or Islamic law as human rights violations</i> .
2	In another verbal attack Kharazi accused the United States _{FS} of wanting to exercise " world dictatorship " since the " horrible attacks " of September 11 .
3	He said those who thought the election process would be rigged were supporters of the MDC party , adding that they were prejudging and wanted to direct the process _{ZX} .
4	However , the fact that certain countries have a more balanced view of the conflict _{ZX} is not the only reason to doubt that anti-Israeli decisions _{FS} will , in fact , be adopted .
5	But his tough stand on P’yongyang _{FS} has provoked concern in Seoul _{ZX} , where President Kim Tae-chung , who is in the last year of his five-year term , has been trying to prise the hermit state out of isolation .

Table 3.14 The dev examples for which the FS-MTL model *correctly* predicts the *target* in 6/8 trials, whereas the Z&X-STL model predicts *incorrectly* in 6/8 trials.

How does MTL help? There are 18/1055 instances in the dev set for which the FS model predicts the holder correctly and the Z&X-STL model does not, and 19/1055 for targets.

For holders, for 9 out of 18 of such examples, the Z&X-STL model does not predict anything (as in examples 2–4 in Table 3.13). From examples 1–5 we notice that SRL data helps to handle more complex syntactic constructions. From examples 5–7 we observed that using MTL with SRL helps to handle cases when more than one person or organization is present in the close neighborhood of the opinion. For targets, in 11 out of 18 cases the Z&X-STL model does not predict anything as in examples 1–2 in Table 3.14. We conclude that the greatest improvements from the FS-MTL model come from having far fewer missing roles.

3.6 Summary

In this chapter, we presented how to address the problem of scarcity of annotated training data for labeling of opinion holders and targets (ORL) using multi-task learning (MTL) with Semantic Role Labeling (SRL). We adapted a recently proposed neural SRL model for ORL and enhanced it with different MTL techniques. Two MTL models achieve significant improvements with all evaluation measures, for both holders and targets, on both dev and test set, when evaluated with repeated 4-fold CV. We recommend evaluation with comparable dev and test set sizes for future work, as this enables more reliable evaluation.

With deeper analysis we show that future developments should improve the ability of the models to capture long-range dependencies, investigate if consistency with syntax can improve ORL, and consider other auxiliary tasks such as dependency parsing or recognizing textual entailment. We emphasize that future improvements can be measured more reliably if opinion expressions with missing roles in the MPQA corpus are curated and if the evaluation considers all mentions in opinion role coreference chains as well as discontinuous roles.

In the next chapter, we turn our attention to the second task that should assist sentiment inference: unrestricted abstract anaphora resolution.

Chapter 4

Resolving Abstract Anaphors in a Relational Neural Model

The sentence-level opinion analysis model presented in Chapter 3 is able to detect from the sentence *We therefore as MDC do not accept this result* that the *MDC* expressed a negative attitude toward the *result*. However, having all this information without knowing what the *result* refers to is not informative since we still do not understand what the *MDC* is truly negative about. That is, we can not understand this sentence in isolation. We need to go beyond the sentence-level and find what the *result* refers to in the prior discourse. In particular, we need to solve Abstract Anaphora Resolution (AAR). This is a challenging task of finding the antecedent of nominal expressions (e.g. *this result*) and pronominal expressions (e.g. *this*, *that*, and *it*) that refer to the so-called *abstract objects* such as facts, events, plans, actions, and situations. In contrary to AAR, significant research efforts have been invested in entity anaphora resolution (or coreference resolution) which resolves multiple ambiguous mentions of a single entity representing a person, a location, or an organization that we can imagine in the real world. In Chapter 2 (Section 2.3), we presented why the resolution of abstract anaphors is still relatively unexplored even though abstract anaphors are very frequent across languages as well as important to computationally solve many NLP tasks. In Section 4.1 we will reflect on the observed challenges, note other issues that emerge, and propose ways to address them. At the end of Section 4.1 we will give the outline of the rest of the chapter and mention the published work this chapter is based upon.

4.1 Challenges and Working Toward the Models

Here we outline the main challenges to resolution of abstract anaphors using a machine learning approach and our solutions to these obstacles.

Feature design. In Chapter 2 (Section 2.3.2 on pages 36–40 and Section 2.3.4 on pages 43–44) we showed that (i) there is a number of lexical, semantic, and syntactic properties associated with abstract anaphora and that (ii) the standard coreference resolution features cannot be applied to abstract anaphora resolution. Therefore, we propose a **neural model** that learns relevant features from data on its own and that does not force us to make certain assumptions that might be limiting.

Limited labeled data. In Chapter 2 (Section 2.3.3 on pages 40–42) we observed that even humans struggle with identifying the exact boundaries of abstract antecedents. This is due to the lack of clear boundaries as well as the complex or unclear inference process for finding antecedents. Therefore the labeled data for unrestricted abstract anaphora resolution is quite scarce. To circumvent the scarcity of labeled data, motivated by Kolhatkar et al. (2013b), we **extract training examples** from parsed corpora using a single common syntactic construction in the following way: (i) we search for constructions with a verb with an embedded sentential argument (22a), (ii) apply a simple transformation that replaces the embedded sentence with an abstract anaphor (such as *this* in (22b)), and (iii) use the cut-off embedded sentence as the antecedent (22c).¹

- (22) a. While few lawmakers [VP anticipated [S' that [S the humanitarian aid would be cut off next month]]], Mr. Ortega's threat practically guarantees that the humanitarian aid will be continued.
- b. **Anaphoric sentence:** While few lawmakers anticipated this, Mr. Ortega's threat practically guarantees that the humanitarian aid will be continued.
- c. **Antecedent:** The humanitarian aid would be cut off next month.

Henceforth, we refer to this extraction method as VC-SS-Extract² and to the extracted data as *silver data*.

¹This is a real example used in experiments.

²For verb/conjunction-mediated anaphoric relation extraction.

Limitations of extracted data. Extracted antecedents do not occur in a natural context once they are extracted from their original sentences. For this reason, we cannot use sequence labeling techniques that label each word of a given text as the part of the antecedent or outside of it. Therefore, we use **ranking models** that rank candidates for the antecedent i.e. syntactic constituents of the constructed sentence with the anaphor, a few preceding sentences as well as the constituents of the extracted antecedent.

Moreover, we note in Chapter 2 (Section 2.3.2 on page 39) that information about the distance between the anaphor and the antecedent plays an important role in anaphora resolution systems. However, an extracted antecedent does not have a distance from the anaphor. Therefore, for each extracted antecedent we **sample a distance** on the basis of a distribution calculated from the natural (gold) abstract anaphora cases in a development set.

Finally, the extracted and natural abstract anaphora cases might differ in some complex properties that cannot be captured with pre-processing decisions. For instance, in Example (23b) the anaphor does not occur in a natural position in the sentence. To ensure that our models do not fit such artifacts we propose **adversarial training** (described in Section 2.4.2 on page 54). It could potentially force our models to capture only features that are relevant for both silver and gold data.

- (23) a. "The main feature of the new organization [VP is [S' that [S each local manager will have both the authority and accountability for profitable and technically sound operations"]]], said Charles E. Spruell, president of the Mobil Unit.
- b. **Anaphoric sentence:** "The main feature of the new organization is this", said Charles E. Spruell, president of the Mobil Unit.
- c. **More natural anaphoric sentence:** "This is the main feature of the new organization", said Charles E. Spruell, president of the Mobil Unit.
- d. **Antecedent:** Each local manager will have both the authority and accountability for profitable and technically sound operations.

Different types of abstract anaphors. We aim to propose a single model for many types of abstract anaphors: pronouns, noun phrases, shell nouns, or events. However, in Chapter 2 (Section 2.3.2) we observed that perhaps different decisions need to be taken for scoring candidates of different types of abstract anaphora i.e. different types of relations between the anaphor and its antecedent. Due to this, resolving each type could be considered as a different but related task. For this reason, we extend our models with **multi-task learning**.

Evaluation. Since abstract antecedents lack clear boundaries, it is hard to properly evaluate performance of systems for resolution of abstract anaphors. To gain an accurate understanding of capabilities of our models we provide results in **supplementary evaluation measures**.

The research presented in this section evolved from the EMNLP paper "A Mention-Ranking Model for Abstract Anaphora Resolution" (Marasović et al., 2017). In this paper, we addressed the first two challenges: feature design and limited labeled data. First, we proposed a baseline ranking neural model which does not use the information about the distance and it is not trained using adversarial training or multi-task learning. Second, we proposed the VC-SS-Extract method for extracting training data and tested the limits of the training exclusively on the silver data. In the paper, we evaluated the model following the experimental setup of the prior work on shell noun resolution (Kolhatkar et al., 2013b). Their models rank *all* syntactic constituents of only *one* sentence—the sentence that contains the antecedent. This is a restricted experimental setup compared to a realistic task setup where a system needs to find the antecedent in at least few preceding sentences. In the restricted task setup our models outperform the state-of-the-art system (Kolhatkar et al., 2013b) in shell noun resolution (see Appendix B). In this thesis, we evaluate our models in the realistic experimental setup in which models need to rank (embedded) sentences and verb phrases from the sentence with the anaphor as well as few preceding sentences. The research results presented in this chapter has not been published yet.

The chapter is organized as follows. First we describe our ranking neural models in detail in Section 4.2. Then, in Section 4.3 we discuss how did we extract training (silver) data. In Section 4.4 we describe the experimental setups, datasets, evaluation metrics, hyperparameters, and other training details. We then evaluate and compare different variants of our core model (Section 4.5). In Section 4.6 we explain how we extend our core model with adversarial training and how does it affect its performance. Likewise, in Section 4.7 for multi-task learning. Finally, we give a summary of this chapter (Section 4.8).

4.2 MR-LSTM: Mention-Ranking LSTM-Siamese Neural Network

We build our approach upon the intuition that we can find what is the antecedent of a given abstract anaphor if we can capture characteristics of the relation that holds between the sentence with the anaphor (i.e. *anaphoric sentence*) and the antecedent. In the running Example (1) on page 1, the characteristics of this relation are the characteristics of something

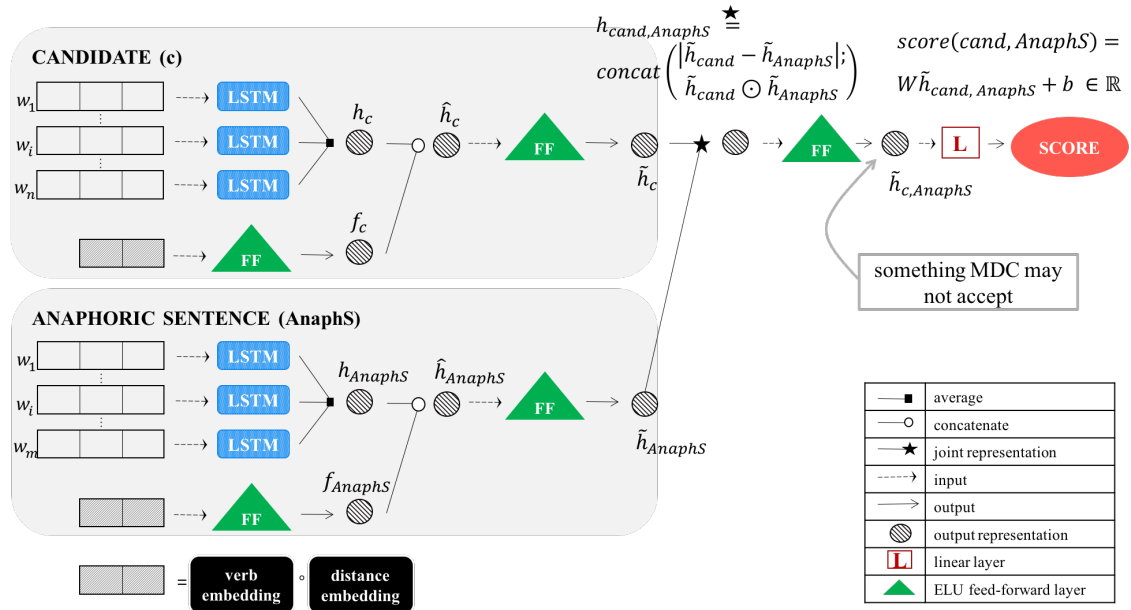


Fig. 4.1 MR-LSTM: Mention-Ranking LSTM-Siamese Neural Network.

that the *MDC* might not accept. In this specific example, the fact that *Mugabe was re-elected with 1,685,212 votes against 1,258,758 votes for Tsvangirai* fits the characteristics of something that the *MDC* might not accept.

We propose a Mention-Ranking LSTM-siamese neural network (MR-LSTM) to learn the characteristics of the mentioned relation. For each candidate for the antecedent, MR-LSTM produces a joint representation of the anaphoric sentence and the candidate. On the basis of this representation each candidate is assigned a relatedness score. The candidates are then ranked with respect to the calculated score. The first-ranked candidate is the predicted antecedent. Figure 4.1 displays the MR-LSTM model. In the remaining part of this section, we describe its components.

Individual representations. MR-LSTM first learns a representation of the anaphoric sentence $AnaphS$ and a representation of a candidate for the antecedent c using a bi-directional LSTM as well as two linguistic features. These features will be described in detail in the following paragraph. The same bi-LSTM is applied to the anaphoric sentence $AnaphS$ and to a candidate for the antecedent c . Hence, the term *siamese* (Chopra et al., 2005; Das et al., 2016; Mueller and Thyagarajan, 2016; Neculoiu et al., 2016).

For both $AnaphS$ and c , the fixed pre-trained word embeddings w_i are sequentially fed into a bi-LSTM which produces the outputs from the forward pass \vec{h}_i and the outputs \overleftarrow{h}_i from the backward pass. The final LSTM's output at the i -th word h_i is produced by concatenating

the vectors from the forward and backward passes i.e. $h_i = [\overleftarrow{h}_i; \overrightarrow{h}_i]$. All outputs h_i are averaged (except for the outputs that correspond to padding tokens) to obtain a representation of the anaphoric sentence h_{AnaphS} . A similar process applies for a representation of the candidate h_c . The representation of the anaphoric sentence h_{AnaphS} is then concatenated with the output of the feed-forward layer of Exponential Linear Units (ELUs) (Clevert et al., 2016) which produces a representation of the linguistic features f_{AnaphS} . The resulting vector is denoted by \hat{h}_{AnaphS} . The vector \hat{h}_c for a candidate for the antecedent is obtained in a similar fashion. Finally, the vectors \hat{h}_{AnaphS} and \hat{h}_c are fed into the second feed-forward layer of ELUs to produce the final representation of the anaphoric sentence \tilde{h}_{AnaphS} and the final representation of the candidate for the antecedent \tilde{h}_c . Since the same parameters are applied to the anaphoric sentence and the candidate, the vectors \tilde{h}_{AnaphS} and \tilde{h}_c are elements of the same vector space.

Linguistic features. We use two linguistic features: the fixed pre-trained word embedding of the verb that governs the head of the anaphor and the learned embedding of the distance between a candidate for the antecedent and the anaphor. We will describe how we extract these feature from data in Section 4.4. In Marasović et al. (2017), we investigated using other features too, but we found them ineffective. See Appendix B for more.

In Chapter 2 (Section 2.3.2 on page 38), we noted that the context of an anaphoric expression can determine the semantic type of the referent. For example, a verb *happen* can only be used with subjects denoting some sort of event (Asher, 1993, p. 22, p. 192). Second, the anaphor’s semantic type preferences can constrain the syntactic type of antecedents. The anaphor in Example (13) on page 39 refers to a concept and therefore a verb phrase is the antecedent instead of the full sentence. Thus, we suspect that the verb that governs the head of the anaphor may serve as an indicator for these preferences (e.g. the verbs *understand* in (11b) on page 37 and *criticized* in (12a) on page 38).

In Chapter 2 (Section 2.3.2 on page 39) we noted that the distance between a candidate for the antecedent and the anaphor is an important feature from a computational perspective. That is, it can narrow down the search scope of candidates for antecedents.

These two features are input to a feed-forward layer of ELUs. The output of the ELU-FF layer is concatenated with the representation produced by the LSTM. It is common to combine a representation produced by an LSTM with additional manually-designed features in this fashion (see e.g. Collins et al., 2017).

Joint representation. Thus far, we produced individual representations of the candidate \tilde{h}_c and the anaphoric sentence \tilde{h}_{AnaphS} . Since these vectors occur in the same vector space,

their dimensions are comparable. Therefore, we can produce a joint representation of the pair $(c, AnaphS)$ using simple mathematical operations over individual representations. In particular, a joint representation $\mathbf{h}_{c, AnaphS} = [|\tilde{\mathbf{h}}_c - \tilde{\mathbf{h}}_{AnaphS}|; \tilde{\mathbf{h}}_c \odot \tilde{\mathbf{h}}_{AnaphS}]$ (Tai et al., 2015), where $|\cdot|$ denotes the absolute values of the element-wise subtraction and \odot the element-wise multiplication. The vector $\mathbf{h}_{c, AnaphS}$ is then fed into a feed-forward layer of ELUs to obtain the final joint representation $\tilde{\mathbf{h}}_{c, AnaphS}$ of the pair $(c, AnaphS)$. We expect that $\tilde{\mathbf{h}}_{c, AnaphS}$ captures characteristics of the relation between the antecedent and the anaphoric sentence (i.e. characteristics of everything that the *MDC* might not accept).

Score. Finally, we compute the score for the pair $(c, AnaphS)$ that represents how well does the candidate for the antecedent c fits the characteristics of the relation between the antecedent and the anaphoric sentence $AnaphS$. The score is calculated by applying a single fully connected linear layer to the joint representation:

$$\text{score}(c, AnaphS) = W\tilde{\mathbf{h}}_{c, AnaphS} + b \in \mathbb{R}, \quad (4.1)$$

where W is a $1 \times d$ weight matrix and d the dimension of the vector $\tilde{\mathbf{h}}_{c, AnaphS}$.

We train the described mention-ranking model with the max-margin training objective from Wiseman et al. (2015) which is used for the antecedent ranking subtask in their study. In particular, suppose that the training set $\mathcal{D} = \{(a_i, s_i, \mathcal{T}(a_i), \mathcal{N}(a_i))\}_{i=1}^n$, where a_i is the i -th abstract anaphor, s_i the corresponding anaphoric sentence, $\mathcal{T}(a_i)$ the set of antecedents of a_i , and $\mathcal{N}(a_i)$ the set of candidates that are not antecedents (negative candidates). Let $\tilde{t}_i = \operatorname{argmax}_{t \in \mathcal{T}(a_i)} \text{score}(t, s_i)$ be the highest scoring antecedent of a_i . Then the loss is:

$$L_{\text{max-margin}} = \sum_{i=1}^n \max(0, \max_{c \in \mathcal{N}(a_i)} \{1 + \text{score}(c, s_i) - \text{score}(\tilde{t}_i, s_i)\}). \quad (4.2)$$

4.3 Training Data Extraction

An obvious question emerges: how can we train a neural model given that the annotated corpora for abstract anaphora resolution is small? We can train our models to learn what characterizes mentioned relations between the anaphor and its antecedent, if we train them on many instances of antecedent–anaphoric sentence pairs. Fortunately, such pairs can be extracted from parsed corpora by searching for a common construction which consists of a verb with an embedded sentence (complement or adverbial). This pattern is displayed in

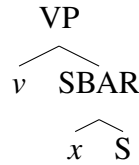


Fig. 4.2 A general pattern for creating anaphoric sentence–antecedent pairs.

embedding sentence type	head of SBAR	suitable anaphoric expressions
argument	\emptyset	this/that/it
argument	that, this, it	that/this/it
argument interrogative	whether	this/that/it
adjunct causal	because	because of this/that/it, due to this/that/it
adjunct causal	since-prp	because of this/that/it, due to this/that/it
adjunct causal	as-prp	because of this/that/it, due to this/that/it
adjunct conditional	if-adv	because of this/that/it, due to this/that/it
adjunct temporal	since-tmp	since then
adjunct temporal	after	after this/that/it

Table 4.1 The embedding sentence types (column 1), the head of SBAR (column 2), and anaphoric expressions they induce (column 3).

Figure 4.2. We refer to this extraction method as **VC-SS-Extract**³ and to the extracted data as *silver data*.

In particular, we detect the extraction pattern in a parsed corpus, e.g. see (24a). Then we find the value of the SBAR’s head in the second column in Table 4.1. In the third column of the corresponding row, we find a choice of suitable anaphoric phrases. We randomly choose one, e.g. *this*. We then "cut off" the SBAR constituent and replace it with the chosen anaphoric phrase to create the anaphoric sentence (24b). Finally, the extracted embedded sentence S serves as the antecedent (24c).

- (24) a. He [VP doubts [SBAR \emptyset [S a Bismarckian super state will emerge that would dominate Europe]]], but warns of "a risk of profound change in the heart of the European Community from a Germany that is too strong, even if democratic".
- b. **Anaphoric sentence:** He doubts this, but warns of "a risk of profound change in the heart of the European Community from a Germany that is too strong, even if democratic".
- c. **Antecedent:** A Bismarckian super state will emerge that would dominate Europe.

³For verb/conjunction-mediated anaphoric relation extraction.

Example (24) illustrates that the pattern applies to verbs that embed sentential arguments. In such cases, the verb establishes a specific semantic relation between the embedding sentence and its sentential complement. However, the pattern also applies to sentential adjuncts (causal or temporal). Adverbial clauses encode specific discourse relations with their embedding sentences which is often indicated by their conjunctions. In (25a), for example, the causal conjunction *as* relates a cause (embedded sentence) and its effect (embedding sentence). We randomly replace causal conjunctions *because*, *as*, *if*, and *since* with suitable anaphoric expressions such as *because of that* and *due to this* (see Table 4.1) that make the causal relation explicit in the anaphor such as in (25b) .

- (25) a. There is speculation that property casualty firms [VP will sell even more munis [SBAR as [S they scramble to raise cash to pay claims related to Hurricane Hugo and the Northern California earthquake]]].
- b. **Anaphoric sentence:** There is speculation that property casualty firms will sell even more munis because of this.
- c. **Antecedent:** They scramble to raise cash to pay claims related to Hurricane Hugo and the Northern California earthquake.

The VC-SS-Extract method is motivated by the construction of the CSN corpus for the resolution of shell nouns by Kolhatkar et al. (2013b) described in Section 2.3.3 on page 40. However, although their extraction method is clever, it heavily exploits the specific properties and categorization of shell nouns which cannot be generalized to other types of abstract anaphors, pronominal or nominal. On the other hand, since the VC-SS-Extract method is based upon a general extraction pattern (Figure 4.2), it covers a much wider range of anaphoric types. Compared to Kolhatkar et al. (2013b), who made use of a carefully constructed set of extraction patterns, a downside of our method is that our extracted antecedents are uniformly of type S. However, since our models are intended to induce semantic representations, we expect the syntactic form to be less critical when compared to a feature-based model. This also alleviates problems with languages like German, where (non-)embedded sentences differ in surface position of the finite verb. We can either adapt the order or ignore it when producing anaphoric sentence–antecedent pairs.

In Section 4.4, we will describe how the VC-SS-Extract method is carried out, provide pre-processing steps, and discuss the quality of silver data.

4.4 Experimental Setup

4.4.1 Data

Details of the corpora (other than silver data) mentioned in this subsection can be found in Chapter 2 (Section 2.3.3 on pages 40–42). We provide data statistics after we introduce training settings.

Silver data. We applied the VC-SS-Extract method to the manually parsed WSJ part of the Penn TreeBank corpus (Marcus et al., 1993) and the automatically parsed NYT corpus from the annotated Gigaword corpus (Napoles et al., 2012). The WSJ and NYT documents which occur in the development or test data were excluded from the extraction.

The VC-SS-Extract method used to extract silver data for experiments in this thesis differs from the extraction method in Marasović et al. (2017). First, in this thesis we use a slightly different set of values of head of the SBAR clause and corresponding anaphoric phrases (Table 4.1 on page 82). This is due to the evaluation of the quality of examples produced by the VC-SS-Extract method. We evaluated 10 randomly selected WSJ examples for every considered value of the SBAR head in Marasović et al. (2017). The first trial of the data quality evaluation showed that the VC-SS-Extract pattern frequently applies to relative clauses and consequently results in poor examples. We constrained the VC-SS-Extract pattern to ensure a more secure extraction and again evaluated 10 newly extracted WSJ examples for every considered value of the SBAR head. The second evaluation trial confirmed that the extraction is safe for a newly defined subset of the SBAR heads (reported in Table 4.1 on page 82). In Appendix C we provide examples we have judged for quality.

Another flaw of silver data produced by Marasović et al. (2017) is that a possible candidate for the antecedent is the syntactic constituent of the original sentence that differs from the extracted antecedent only in the embedding verb such as the candidate in (26d). Our neural model can easily exploit this artifact by scoring high candidates whose first word matches the verb that governs the head of the anaphor in the anaphoric sentence. This is problematic, since the evaluation measure $\text{success}@n$ measures whether the antecedent or a candidate that differs in one word are in the first n ranked. For this reason, the model that solves the task using this simplistic heuristic will perform well with respect to the evaluation measure. This is an important takeaway message learned from data pre-processing: even buggy neural models may still produce reasonable outputs. In this thesis, all constituents such as (26d) are excluded from the list of candidates for the antecedent.

- (26) a. **Original sentence:** He [VP said [SBAR [S the fourth quarter will be "challenging"]], and maintained his conservative forecast that 1990 "won't be a barn burner".

- b. **Anaphoric sentence:** He *said* this, and maintained his conservative forecast that 1990 "won't be a barn burner".
- c. **Antecedent:** the fourth quarter will be "challenging"
- d. **Candidate that differs in one word:** *said* the fourth quarter will be "challenging"

Gold unrestricted AAR data. We use anaphors from the WSJ part of the ARRAU corpus (Uryupina et al., 2016; Poesio et al., 2018) categorized as *abstract* or *plan*. We discard 15 anaphors that have more than two antecedents from two different sentences. This subset includes both nominal and pronominal abstract anaphors. It is also only available resource for studying *unrestricted* abstract anaphora resolution. Since it contains less than 585 example, we use this subset only for testing.

Gold shell noun resolution data. We use the ASN corpus (Section 2.3.3 on page 41) annotated with six anaphoric shell nouns: *decision*, *fact*, *issue*, *possibility*, *reason*, and *question*, that occur in the NYT corpus (Kolhatkar et al., 2013a,b). While Kolhatkar et al. (2013b) train six separate models for six shell nouns, we aim at a single model for resolution of all abstract anaphors. At the same time, we want to compare our models to at least one model from Kolhatkar et al. (2013b). Therefore, we set aside all *decision* instances for testing. For the development set we randomly pick 60 examples from every shell noun data (except *decision*) and use the rest for training. The authors provided us with the crowd workers' annotations of the sentence that contains the antecedent, antecedents chosen by the workers, and links to the NYT corpus. The extraction of the anaphoric sentence and the candidates had to be redone. We follow Kolhatkar (2015) in the requirement that the confidence of the antecedent is higher than 0.5. However, many instances in the provided dataset do not have such an antecedent. For such cases we took the antecedent with the highest confidence, no matter how low it is. We discard antecedents which match with the string *EMPTYEMPTY*.

Gold event resolution data. Following Jauhar et al. (2015), we extract *this*, *that*, and *it* pronouns in the CoNLL-12 shared task dataset (Pradhan et al., 2012) whose preceding mention in the coreference cluster is verbal. Unlike us, Jauhar et al. (2015) also consider cataphoric pronouns i.e. pronouns which are the first mention in the cluster and the next one is verbal. We use the train, dev, and test shared task data splits for our experiments. The CoNLL-12 dataset, besides anaphors from news articles, contains anaphors from magazine articles, broadcast news, and conversations, web data and conversational speech. We discard broadcast and telephone conversation, since we focus on the in-domain evaluation in this

thesis. We recovered the full antecedent phrase in CoNLL-12 using the attribute `parse_tree` in the `cort`⁴ library.

4.4.2 Experimental Configurations

Candidates for the antecedent are *S-like* syntactic constituents obtained from a given *window* using the Stanford constituency parser (Bowman et al., 2016). In particular, we say a constituent is S-like if it is assigned one of the following constituent tag labels: S, VP, ROOT, SBAR, or SBARQ. If the antecedent is not assigned a constituent tag label from this set, the corresponding anaphor is discarded from training and testing.

Window for selecting candidates. The window from which candidates are selected consists of d sentences that precede the anaphoric sentence, $d \in \{1, 2, 3, 4\}$, as well as the anaphoric sentence. We denote these windows by $W_i = \{\text{AnaphS}_{-i}, \dots, \text{AnaphS}_{-1}, \text{AnaphS}\}$ where AnaphS_{-i} denotes a sentence that occurs before the anaphoric sentence and there are $i - 1$ sentences between them. Here a sentence is a string that starts with a capital letter and ends with a period which indicates the end of the sentence. All constituents that are not the antecedent are referred to as *negative candidates*. We follow the prior work on shell noun resolution (Kolhatkar et al., 2013b) and consider every constituent that differs in one word and any number of punctuation symbols from the true antecedent also as the antecedent. Table 4.2 shows the number of anaphors across datasets without any requirements on the antecedent, the number of anaphors which have the S-like antecedent, and the number of anaphors which have the S-like antecedent that occurs in a given window.

Training with silver vs. gold vs. mixed data. We experiment with training on the silver data only (i.e. the data extracted using the VC-SS-Extract method), on the gold data only (the training parts of the ASN and the CoNLL-12 corpus), and on the mixture of the silver and the gold training data. We conduct mixing such that we alternate batches from the gold and silver data from the beginning of the training. Table 4.3 shows which datasets are used for training our models in each of these setups.

Dev and test sets. The development set consist of the development parts of the ASN and the CoNLL-12 corpus. For testing we use the test part of the ASN corpus and the CoNLL-12 corpus as well as the whole ARRAU corpus.

⁴<https://github.com/smartschat/cort>

# anaphors					
test	ASN 373	CoNLL-12 121	ARRAU _{all} 585	ARRAU _{nominal} 386	ARRAU _{pronom} 199
train	ASN 1563	CoNLL-12 1141	WSJ 2490	NYT 78126	
dev	ASN 300	CoNLL-12 100			
★ # anaphors with the antecedent in $\in \{S, VP, ROOT, SBAR, SBARQ\}$					
test	ASN 363	CoNLL-12 121	ARRAU _{all} 578	ARRAU _{nominal} 381	ARRAU _{pronom} 197
train	ASN 1507	CoNLL-12 1141	WSJ 2490	NYT 78126	
dev	ASN 281	CoNLL-12 100			
★ \wedge antecedent \in window= $\{AnaphS_{-1}, AnaphS\}$					
test	ASN 303	CoNLL-12 112	ARRAU _{all} 447	ARRAU _{nominal} 258	ARRAU _{pronom} 189
train	ASN 1324	CoNLL-12 1033	WSJ 2191	NYT 68625	
dev	ASN 255	CoNLL-12 91			
★ \wedge antecedent \in window= $\{AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$					
test	ASN 335	CoNLL-12 116	ARRAU _{all} 490	ARRAU _{nominal} 296	ARRAU _{pronom} 194
train	ASN 1442	CoNLL-12 1088	WSJ 2398	NYT 74938	
dev	ASN 271	CoNLL-12 98			
★ \wedge antecedent \in window= $\{AnaphS_{-3}, AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$					
test	ASN 350	CoNLL-12 117	ARRAU _{all} 510	ARRAU _{nominal} 314	ARRAU _{pronom} 196
train	ASN 1476	CoNLL-12 1099	WSJ 2455	NYT 76926	
dev	ASN 279	CoNLL-12 98			
★ \wedge antecedent \in window= $\{AnaphS_{-4}, AnaphS_{-3}, AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$					
test	ASN 363	CoNLL-12 117	ARRAU _{all} 527	ARRAU _{nominal} 331	ARRAU _{pronom} 196
train	ASN 1502	CoNLL-12 1105	WSJ 2490	NYT 78126	
dev	ASN 281	CoNLL-12 98			

Table 4.2 Data statistics: number of anaphors. $AnaphS_{-d}$ denotes the sentence that occurs $d - 1$ positions (in number of intervening sentences) before the anaphoric sentence. The requirement "antecedent \in window" indicates that the shown number is the number of anaphors whose antecedent occurs in this window. The symbol ★ indicates that we exclude anaphors whose antecedent's syntactic tag label is not in the set $\{S, VP, ROOT, SBAR, SBARQ\}$. Difference between rows (1–3) and (4–6) shows how many anaphors can not be resolved with our models since their antecedent is not an S-like syntactic constituent.

	NYT	WSJ	ASN (train)	CoNLL-12 (train)
silver	✓	✓	✗	✗
gold	✗	✗	✓	✓
mixed	✓	✓	✓	✓

Table 4.3 Datasets used in different training data types: silver vs. gold vs. mixed.

$x \backslash p(d = i)$	$p(d = 0)$	$p(d = 1)$	$p(d = 2)$	$p(d = 3)$	$p(d = 4)$
1	0.23	0.77	-	-	-
2	0.21	0.7	0.08	-	-
3	0.21	0.68	0.08	0.03	-
4	0.2	0.67	0.08	0.025	0.015

Table 4.4 Distance probabilities calculated from a balanced subset of the development parts of the ASN and CoNLL-12 corpora.

Linguistic features. In silver data, the word embedding of the verb that matches the extraction pattern (i.e. see v in Figure 4.2 on page 82) is used as the verb feature. For other datasets, we use a dependency parser⁵ to extract the verb that governs the head of the anaphor and use its 100-dimensional GloVe embedding (Pennington et al., 2014) as the verb feature. If we do not manage to extract the verb, we use an embedding of the UNK token.

For gold data examples, the distance between a candidate for the antecedent and the anaphor is measured as the distance between the sentence that contains the candidate and the anaphoric sentence, measured in number of sentences. For example, a candidate that occurs in the sentence that precedes the anaphoric sentence is assigned the distance value 1. For silver examples, we sample a distance from a distribution calculated on a sample of the development set that contains the same number of shell nouns from the subset of the ASN corpus and events from the subset of the CoNLL-12 shared task dataset. We first count occurrences of distance values $i \leq 4$ in the development set sample. Then we define the probability $p(d = i)$ of distance taking the value i , $i \leq 4$, as the inverse of its occurrence count. Table 4.4 shows the exact probabilities that are used.

To test the impact of the two features on the full model, we train variants of the MR-LSTM model: (i) without the VERB feature and (ii) without the DISTANCE feature, on both the gold and the silver data. Note that we always omit one feature at the time.

⁵<https://spacy.io/usage/linguistic-features>

4.4.3 Evaluation Metrics

Following the prior work on shell noun resolution (Kolhatkar et al., 2013b), we report success@n which measures whether the antecedent or a candidate that differs in one word and any number of punctuation symbols occurs in the first n ranked candidates,

$$\text{success@n} = \mathbb{1}_{\{\text{the antecedent (or a slightly different candidate) occurs in the first } n \text{ ranked}\}}, \quad (4.3)$$

for $n \in \{1, \dots, 4\}$. The standard precision equals to the success@1 score.

We also propose two new evaluation measures: token-F1 score and sentence-accuracy. The token-F1 score is motivated by work on sentence compression (Filippova et al., 2015). It should indicate the degree of the overlap between the predicted antecedent and the correct antecedent. We simply compute the token-recall and token-precision in terms of tokens in the antecedent and the predicted antecedent:

$$\text{token-precision} = \frac{|\text{tokens that occur in the predicted antecedent and the antecedent}|}{|\text{tokens in the predicted antecedent}|} \quad (4.4)$$

$$\text{token-recall} = \frac{|\text{tokens that occur in the predicted antecedent and the antecedent}|}{|\text{tokens in the antecedent}|} \quad (4.5)$$

$$\text{token-F1} = \frac{2 \cdot \text{token-precision} \cdot \text{token-recall}}{\text{token-precision} + \text{token-recall}} \quad (4.6)$$

We introduce the sentence-accuracy to measure how well our models detect the sentence where the antecedent occurs. The sentence-accuracy checks if the predicted antecedent occurs in the same sentence as the correct antecedent:

$$\text{sentence-accuracy} = \mathbb{1}_{\{\text{the predicted antecedent \& the antecedent occur in the same sentence}\}}. \quad (4.7)$$

We report the average success@n , the average token-F1, and the average sentence accuracy calculated across anaphors that our models can resolve (i.e. anaphors that have a S-like antecedent that occurs in a given window).

4.4.4 Baselines

We compare our models to three baselines: $\text{BL}_{\text{dist,sent}}$, $\text{BL}_{\text{dist>tag}}$, and $\text{BL}_{\text>tag}}$.

First, we calculate a distribution of distance values from a sample of the development set (i.e. development part of the ASN and CoNLL-12 corpora) which contains the same number of shell nouns and events. For a given window size $x \in \{1, 2, 3, 4\}$, we count the occurrence

of distance values $i \leq x$ in the development set sample. Then we define the probability $p(d = i)$ of distance taking the value $i, i \leq x$, as the inverse of its occurrence count. Table 4.4 on page 88 shows the calculated probabilities.

Given a window size x , the $BL_{dist,sent}$ baseline samples a distance d from a distribution calculated on the development part of the ASN and CoNLL-12 corpora (i.e. from row x in Table 4.4 on page 88). Then it picks for the antecedent the full sentence $AnaphS_{-d}$ that occurs before the anaphoric sentence and there are $d - 1$ other sentences between them.

Next, given a window size x , the $BL_{dist,tag}$ baseline also samples a distance d from the distance distribution. Then it randomly picks four S, VP, ROOT, SBAR, or SBARQ constituents from the sentence $AnaphS_{-d}$. The four chosen constituents are ranked in the order they are sampled, i.e. the first sampled is ranked first. We report the average of 10 runs of the $BL_{dist,tag}$ baseline.

Finally, given a window size x , BL_{tag} randomly chooses four S, VP, ROOT, SBAR, or SBARQ constituents from the x preceding sentences as well as the anaphoric sentence. The four chosen constituents are also ranked in the order they are sampled. We report the average of 10 runs of the BL_{tag} baseline.

There is only one prior work we could compare our models to. That is the reported result of Kolhatkar (2015) for resolution of the shell noun *decision* with the window size 4. However, since we had to process the ASN corpus on our own, the number of anaphors we managed to extract differs from the number of anaphors Kolhatkar (2015) used for her evaluation. We extracted 373 instances of *decision* (the first row in Table 4.2 on page 87), while Kolhatkar (2015) reports 390. Moreover, the model of Kolhatkar (2015) is trained only for resolution of the shell noun *decision*, on the subset of the CSN training data of 62451 *decision* instances. We do not train our models only on this subset. Nevertheless, we will give a comparison to get an impression of how well do our models work.

Since Jauhar et al. (2015) resolve cataphoric pronouns, our CoNLL-12 test sets are not comparable. Therefore we do not compare our models.

4.4.5 Hyperparameters

For selecting hyperparameters (HPs) we follow suggestions in the comprehensive HP study of Reimers and Gurevych (2017).

Input representation. We used 100d GloVe word embeddings (Pennington et al., 2014) pre-trained on Gigaword and Wikipedia and we do not fine-tune them. The vocabulary was built from the anaphoric sentence and sentences from which candidates are extracted for all instances in the training data. The out-of-the-vocabulary words were replaced with an UNK token. Embeddings for distances are initialized with values drawn from the uniform

distribution $\mathcal{U}\left(-\frac{1}{\sqrt{d+t}}, \frac{1}{\sqrt{d+n}}\right)$, where $n = 6$ is the number of possible distance values (from 0 to 4 and an additional marker for padding) and $d = 20$ the size of the distance embeddings. The distance embeddings are tuned during training. For the verb embedding we use the corresponding GloVe embedding.

Weights initialization. The size of all LSTM hidden states was set to 50. We initialized the LSTM weights with random orthogonal matrices (Henaff et al., 2016), all other weight matrices with the *He initialization* (He et al., 2015). LSTM forget biases were initialized with ones (Jozefowicz et al., 2015), all other biases with zeros. The size of the feature FF-ELU layer is set to 50, the size of the FF-ELU layer applied to individual representations to 200, and of the size of the FF-ELU layer applied to the joint representation to 400.

Optimization. We trained our models in mini-batches of size 10 using Adam (Kingma and Ba, 2015) with the learning rate of 10^{-4} and 100K iterations. We clip gradients by global norm (Pascanu et al., 2013), with a clipping value set to 1. We stop training if the model did not improve results on the development set in 2500 iterations. For the minmax optimization we use a gradient reversal layer (Ganin and Lempitsky, 2015). The discriminator’s cross-entropy loss is scaled with 0.5.

Regularization. Variational dropout (Gal and Ghahramani, 2016) with the keep probability $k_p = 0.8$ was applied to the outputs and $k_p = 0.85$ to the recurrent connections of the LSTMs. Standard dropout (Srivastava et al., 2014) was applied to the input with $k_p = 0.8$ and to the FF layer of individual representations and the FF layer of the joint representation with the keep probability $k_p = 0.75$. We used the L2 regularization with $\lambda = 1.75^{-5}$.

4.5 Performance of the MR-LSTM Model

All variants of our MR-LSTM models are evaluated every 250 iterations (2500 training examples) and saved if they improve success@1 score on the development set (i.e. the development parts of the ASN and CoNLL-12 corpora). The saved models are then used for evaluation on the test sets.

Note that the number of anaphors that we aim to resolve differs across different windows W_i , $i \in \{1, 2, 3, 4\}$, for the extraction of candidates, since we exclude anaphors that do not have an antecedent in the given window. For this reason, the performance of our models across different windows is not directly comparable. Finally, in all configurations in this section, anaphors that do not have an S-like antecedent are excluded from the evaluation. This also holds for the baselines.

table	anaphors s. t. \exists S-like antecedent $a, a \subseteq S, S \in W$
Table 4.11 (p. 99)	$W=W_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$
Table 4.15 (p. 103)	$W=W_2=\{\text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$
Table 4.19 (p. 107)	$W=W_3=\{\text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$
Table 4.23 (p. 111)	$W=W_4=\{\text{AnaphS}_{-4}, \text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$

Table 4.5 An overview of the evaluation configurations for the baselines and corresponding tables.

4.5.1 A Comparison Between MR-LSTM and Baselines

We compare our models to the baselines: $BL_{dist,sent}$, $BL_{dist,tag}$, and BL_{tag} (described on pages 89–90 in Section 4.4). Tables 4.11, 4.15, 4.19, and 4.23 on pages 99, 103, 107, and 111, show the results of the three baselines in four different configurations, and in all of the evaluation measures described in Section 4.4.3 on page 89. The configurations differ in the window $W_i = \{\text{AnaphS}_{-i}, \dots, \text{AnaphS}\}$, $i \in \{1, 2, 3, 4\}$, from which candidates for the antecedent are extracted. In Table 4.5, we give an overview of the configurations and corresponding tables.

For all datasets except the development and test parts of the CoNLL-12 corpus, there is a variant of our MR-LSTM model (including those where we omit features) that beats all the baselines in large margin. This result suggests that our models are resolving abstract anaphors by looking beyond the distance from the anaphor and the syntactic type of the antecedent. The fact that the $BL_{dist,tag}$ baseline achieves better results than our models only for anaphors in the CoNLL-12 corpus is a first suggestion that the resolution of *this*, *that*, *it* events in the CoNLL-12 corpus requires different modeling.

4.5.2 MR-LSTM Results Across Different Training Data Types

In this section, we analyze the performance of the full MR-LSTM model with respect to the different types of training data: silver, gold, and mixed. Refer to Table 4.3 on page 88 for the information which datasets correspond to which training data type. We report results of our MR-LSTM models with all components described in Section 4.2 in four tables (4.12, 4.16, 4.20, 4.24) for four candidate extraction windows $W_i = \{\text{AnaphS}_{-i}, \dots, \text{AnaphS}\}$, $i \in \{1, 2, 3, 4\}$. For the comparison, we repeat the results of the most effective baseline across windows, $BL_{dist,tag}$. In Table 4.6, we give an overview of the configurations and corresponding tables.

table	features	anaphors s. t. \exists S-like antecedent $a, a \subseteq S, S \in \mathbf{W}$	training
Table 4.12 (p. 100)	all	$\mathbf{W}=\mathbf{W}_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$	gold, silver, mixed
Table 4.16 (p. 104)	all	$\mathbf{W}=\mathbf{W}_2=\{\text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$	
Table 4.20 (p. 108)	all	$\mathbf{W}=\mathbf{W}_3=\{\text{AnaphS}_{-3}, \dots, \text{AnaphS}\}$	
Table 4.24 (p. 112)	all	$\mathbf{W}=\mathbf{W}_4=\{\text{AnaphS}_{-4}, \dots, \text{AnaphS}\}$	
Table 4.13 (p. 101)	-VERB	$\mathbf{W}=\mathbf{W}_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$	gold, silver
Table 4.17 (p. 105)	-VERB	$\mathbf{W}=\mathbf{W}_2=\{\text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$	
Table 4.21 (p. 109)	-VERB	$\mathbf{W}=\mathbf{W}_3=\{\text{AnaphS}_{-3}, \dots, \text{AnaphS}\}$	
Table 4.25 (p. 113)	-VERB	$\mathbf{W}=\mathbf{W}_4=\{\text{AnaphS}_{-4}, \dots, \text{AnaphS}\}$	
Table 4.14 (p. 102)	-DIST	$\mathbf{W}=\mathbf{W}_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$	gold, silver
Table 4.18 (p. 106)	-DIST	$\mathbf{W}=\mathbf{W}_2=\{\text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$	
Table 4.22 (p. 110)	-DIST	$\mathbf{W}=\mathbf{W}_3=\{\text{AnaphS}_{-3}, \dots, \text{AnaphS}\}$	
Table 4.26 (p. 114)	-DIST	$\mathbf{W}=\mathbf{W}_4=\{\text{AnaphS}_{-4}, \dots, \text{AnaphS}\}$	

Table 4.6 An overview of the evaluation configurations for our MR-LSTM models and corresponding tables.

		success@1				token-F1					
		W ₁	W ₂	W ₃	W ₄			W ₁	W ₂	W ₃	W ₄
dev	ASN	gold	silver	gold	gold	dev	ASN	silver	silver	silver	silver
	CoNLL-12	gold	gold	gold	gold		CoNLL-12	gold	gold	gold	gold
test	ASN	gold	silver	gold	silver	test	ASN	silver	silver	silver	silver
	CoNLL-12	gold	gold	gold	gold		CoNLL-12	gold	gold	gold	gold
	ARRAU _{all}	gold	silver	gold	silver		ARRAU _{all}	silver	silver	silver	silver
	ARRAU _{nominal}	gold	gold	gold	silver		ARRAU _{nominal}	silver	silver	silver	silver
	ARRAU _{pronominal}	silver	silver	silver	silver		ARRAU _{pronominal}	silver	silver	silver	silver
average-test		gold	gold	gold	silver	average-test		silver	silver	silver	silver

Table 4.7 Comparison between the training on the silver data to the training on the gold data.

Large-scale silver data vs. small-scale gold data. For each of Tables 4.12, 4.16, 4.20, and 4.24, we compare the results of the full MR-LSTM model trained on the large-scale silver data to the results of the full MR-LSTM model trained on the small-scale gold data. Table 4.7 summarizes which training data type (gold or silver) is better for a given pair of an evaluation dataset (row) and a candidate extraction window $\mathbf{W}_i=\{\text{AnaphS}_{-i}, \dots, \text{AnaphS}\}$ (column).

From Table 4.7, we observe that the training on the gold data gives better results for the development and test parts of the CoNLL-12 dataset in all evaluation measures. Moreover, if we focus solely on the success@1 evaluation measure, from Tables 4.12, 4.16, 4.20, and 4.24, we observe a notable benefit for the CoNLL-12 dataset coming from training exclusively on the gold data. This indicates that perhaps our silver data is not suitable for resolution of anaphors in the CoNLL-12 dataset. However, if we examine the results in terms of the

token-F1 score and the sentence-accuracy in Tables 4.12, 4.16, 4.20, and 4.24, we observe that the difference is not tremendous. These observations suggest that (i) the silver data anaphora cases are different from the CoNLL-12 anaphora cases and that (ii) our models are not able to learn from the silver data, as good as from the gold data, to point to the exact boundaries of antecedents in the CoNLL-12 corpus, but (iii) our models can learn from the silver data nearly as good as from the gold data to choose a candidate which is similar to the correct antecedent in the CoNLL-12 corpus.

Let's turn our attention to other evaluation corpora: the ASN corpus for shell noun resolution and the ARRAU corpus for unrestricted abstract anaphora resolution. In terms of success@1 (the left part of Table 4.7), we detect that the gold training works well for the development part of the ASN corpus. However, on the test part of the ASN corpus (which consists only of examples of the shell noun *decision* that does not occur in the training set) the performance depends on the window size.

From the left part of Table 4.7, we see that the gold training results in better performance in 3 out of 4 window sizes for the nominal part of the ARRAU corpus (ARRAU_{nominal}). Since the silver data contains only pronouns as anaphors, the lower performance on the nominal part of the ARRAU corpus (ARRAU_{nominal}) can be explained by saying that the nominal ARRAU cases do not have as good representatives in the silver data as in the gold data (e.g. shell nouns). Next, the gold training is always a worse choice than the silver training for the pronominal part of the ARRAU corpus (ARRAU_{pronominal}) (see the left part of Table 4.7). We have already hinted that the silver data anaphora cases are different from the CoNLL-12 anaphora cases. If this is really the case, the pronominal anaphors in the ARRAU corpus (ARRAU_{pronominal}) do not have good representatives in the gold training dataset, since the gold training dataset besides the CoNLL-12 corpus contains only the ASN corpus which consists entirely of nominal anaphors. To conclude thus far, our MR-LSTM models do not generalize well if a training dataset does not have representatives of types of anaphors that may occur at the evaluation stage. For this reason, the training on the gold data works better in some cases (CoNLL-12, the dev part of ASN, ARRAU_{nominal}) and the training on the silver data in other (ARRAU_{pronominal}).

Finally, the training on silver data results in better performance in terms of the token-F1 scores for all datasets and window sizes, except for the development part of the ASN corpus and the window W_4 as well as the development and test part of the CoNLL-12 dataset and all windows. This observation suggests that MR-LSTM models trained on the silver data pick candidates for the antecedent that do not match exactly with the correct antecedent but on average they have a higher overlap with the correct antecedent than candidates that are chosen by the models that are trained on the gold dataset.

		success@1				token-F1					
		W ₁	W ₂	W ₃	W ₄	W ₁	W ₂	W ₃	W ₄		
dev	ASN	gold	silver	gold	mixed	dev	ASN	silver	silver	silver	silver
	CoNLL-12	gold	gold	gold	gold		CoNLL-12	gold	gold	gold	gold
test	ASN	gold	mixed	gold	silver	test	ASN	silver	silver	silver	silver
	CoNLL-12	gold	gold	gold	gold		CoNLL-12	gold	gold	gold	gold
	ARRAU _{all}	mixed	mixed	gold	mixed		ARRAU _{all}	silver	silver	mixed	silver
	ARRAU _{nominal}	mixed	gold	gold	mixed		ARRAU _{nominal}	silver	silver	mixed	silver
	ARRAU _{pronominal}	silver	mixed	mixed	silver/mixed		ARRAU _{pronominal}	silver	silver	mixed	silver
average-test		gold	gold	gold	mixed	average-test		silver	silver	silver	silver

Table 4.8 The best training strategies between gold, silver, and mixed. The window $\{\text{AnaphS}_{-i}, \dots, \text{AnaphS}\}$ is denoted by W_i . The "mixed" in **bold** indicates the cases where the training on the mixture of the gold and silver datasets results in the best performance.

Benefit of mixing silver and gold data. The results presented so far suggest that the gold training dataset contains better representatives for some types of evaluation examples (CoNLL-12, the dev part of ASN, ARRAU_{nominal}), while the silver data contains better representatives for other (ARRAU_{pronominal}). A natural question emerges: what happens when we blend these two sources of training data?

Ideally the mixed training would offer us the best of two worlds. First, good representatives from the gold training dataset for the CoNLL-12 corpus, the development part of the ASN corpus, and the nominal part of the ARRAU corpus (ARRAU_{nominal}). Second, good representatives from the silver data for the pronominal part of the ARRAU corpus (ARRAU_{pronominal}). When the training on the blend of the gold and silver data would use these representatives perfectly, the models that are trained on the mixture would result in the best performance. However, Table 4.8 shows that the mixed training is the best in some cases, but not uniformly. It remains to be seen how to blend silver and gold training data types effectively.

4.5.3 Impact of Linguistic Features on MR-LSTM

We analyze the benefit of two linguistic features: the pre-trained word embedding of the verb that governs the head of the anaphoric expression and the learned embedding of the distance between a candidate for the antecedent and the anaphor (in number of sentences).

We report results of the MR-LSTM without the VERB feature trained on silver and gold training data types in four tables (4.13, 4.17, 4.21, 4.25) for four candidate extraction windows. Likewise, for the MR-LSTM model without the DISTANCE feature (Tables 4.14, 4.18, 4.22, 4.26).

gold training		success@1				token-F1					
		W ₁	W ₂	W ₃	W ₄	W ₁	W ₂	W ₃	W ₄		
dev	ASN	✓	✗	✓	✗	dev	ASN	✓	✗	✗	✗
	CoNLL-12	✓	✓	✓	✓		CoNLL-12	✓	✓	✓	✓
test	ASN	✓	✗	✗	✗	test	ASN	✗	✗	✗	✗
	CoNLL-12	✓	✓	✓	✓		CoNLL-12	✓	✓	✓	✓
	ARRAU _{all}	✗	✗	✗	✗		ARRAU _{all}	✗	✗	✗	✗
	ARRAU _{nominal}	✗	✗	✗	✗		ARRAU _{nominal}	✗	✗	✗	✗
	ARRAU _{pronominal}	✓	✗	✗	✗		ARRAU _{pronominal}	✗	✗	✗	✗
average-test		✓	✗	✗	✗	average-test		✗	✗	✗	✗

silver training		success@1				token-F1					
		W ₁	W ₂	W ₃	W ₄	W ₁	W ₂	W ₃	W ₄		
dev	ASN	✗	✗	✗	✗	dev	ASN	✗	✗	✗	✗
	CoNLL-12	-	✓	✓	✓		CoNLL-12	✓	✓	✓	✓
test	ASN	✗	✗	✗	✗	test	ASN	✓	✗	✗	✗
	CoNLL-12	✗	✓	✓	✓		CoNLL-12	✓	✓	✗	✓
	ARRAU _{all}	✗	✗	✗	✗		ARRAU _{all}	✓	✗	✗	✗
	ARRAU _{nominal}	✗	✗	✗	✗		ARRAU _{nominal}	✗	✗	✗	✗
	ARRAU _{pronominal}	✓	✗	✗	✗		ARRAU _{pronominal}	✓	✗	✗	✗
average-test		✗	✗	✗	✗	average-test		✓	✗	✗	✗

Table 4.9 Comparison between the MR-LSTM with and without the VERB feature. The symbol ✓ denotes the cases when MR-LSTM benefits from the VERB training and ✗ when it does not.

The impact of the VERB feature. From Tables 4.13, 4.17, 4.21, and 4.25, we detect when MT-LSTM trained on the gold or the silver data benefits from the VERB feature. We summarize those cases in Table 4.9.

The only dataset that benefits from the VERB feature in large margins is the CoNLL-12 dataset. In Section 2.3.2 in Chapter 2 we observed that the verb that governs the head of the anaphor is a good indicator of the type of abstract antecedent (e.g. fact or event). Since omitting the VERB feature results in a big drop in performance on the CoNLL-12, it may be concluded that (i) the verbs that suggest that the antecedent is an event are well represented in the training part of the CoNLL-12 dataset, (ii) the MR-LSTM model detects this property and makes good use of this insight.

On the other hand, the VERB feature is not beneficial for other evaluation datasets in the majority of cases, independent whether the MR-LSTM is trained on the gold or the silver dataset. How to integrate linguistic features in neural models such that they are effective is an

MR-LSTM	s@1	s@2	s@3	s@4
gold	22.77	31.63	39.08	44.61
-VERB	23.90	36.95	43.97	49.93
silver	23.76	34.61	42.91	48.65
-VERB	24.61	36.74	44.61	48.65
mixed	20.92	32.91	40.35	45.67
Kolhatkar (2015)	28	30	33	35

Table 4.10 Comparison between variants of the MR-LSTM on the test part of the ASN corpus (shell noun *decision*) and the reported result in Kolhatkar (2015).

open research question in NLP. Presently, neural models that use linguistically oriented biases do not result in the best performance for many benchmark tasks. Recently, Moosavi and Strube (2018) showed that linguistic features help to significantly improve generalization of neural coreference resolver, but only if they are employed carefully. In particular, they use a discriminative pattern mining algorithm for finding feature-value pairs that are informative for coreference. Holtzman et al. (2018) trained their language models such that they encourage them to learn linguistic features such as relevance, style, repetition, and entailment, but in a data-driven fashion using particular loss functions. That is, their models are not reliant on the explicit use of the output of natural language understanding tools. In the future, we could consider these proposals and integrate the VERB feature more carefully.

The impact of the DISTANCE feature. From Tables 4.14, 4.18, 4.22, and 4.26, we detect that omitting the DISTANCE feature from the MR-LSTM model significantly hurts its performance across all training configurations and evaluation datasets. This result confirms that the distance between a given anaphor and its antecedent is important feature for automatic resolution of abstract anaphors.

4.5.4 Comparison between MR-LSTM and Kolhatkar (2015)

We compare different variants of our MR-LSTM model in resolution of the shell noun *decision* (i.e. a subset of the test part of the ASN corpus) to the results reported in Kolhatkar (2015). The models have to pick antecedents from 4 preceding sentences and the anaphoric sentence (i.e. from window W_4). Following Kolhatkar (2015), for this experiment we consider all anaphors, i.e. we do not exclude those whose antecedent is not an S, VP, ROOT, SBAR, or SBARQ constituent of the given window or if it does not occur in the given window. To all such anaphors we assign zero success@n scores.

We give a comparison in Table 4.10. However, note that there are major differences in the training sets and even some differences in the test set. Kolhatkar (2015) trained a ranking-SVM model only to resolve the shell noun *decision* on 62451 *decision* examples from the CSN corpus. The test set differs since we had to process it ourselves and while we extract 275 examples, Kolhatkar (2015) extract 15 more. All in all, Kolhatkar (2015) outperforms our models in terms of success@1, but all variants of our model are better in terms of success@2–4 although our models are trained to solve a much more challenging task.

$\mathbf{BL}_{\text{dist,sent}}$		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	43.33	-	-	-	55.37	69.57
	CoNLL-12	0.44	-	-	-	30.75	59.12
test	ASN	28.42	-	-	-	51.56	71.75
	CoNLL-12	4.02	-	-	-	31.21	58.39
	ARRAU _{all}	36.71	-	-	-	48.23	63.49
	ARRAU _{nominal}	42.79	-	-	-	53.29	67.64
	ARRAU _{pronominal}	28.15	-	-	-	40.64	56.30
	average	28.02	-	-	-	44.99	63.51
$\mathbf{BL}_{\text{dist,tag}}$		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	23.76	41.68	53.66	61.46	48.79	69.57
	CoNLL-12	28.57	52.75	66.78	74.55	48.13	59.12
test	ASN	23.14	41.37	54.55	63.45	49.62	71.75
	CoNLL-12	24.55	45.92	59.11	68.42	44.43	58.39
	ARRAU _{all}	20.18	36.69	48.76	56.97	47.23	63.49
	ARRAU _{nominal}	20.97	37.47	50.08	59.84	51.04	67.64
	ARRAU _{pronominal}	19.37	37.26	50.33	57.78	42.47	56.30
	average	21.64	39.74	52.57	61.29	46.96	63.51
\mathbf{BL}_{tag}		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	10.71	20.63	29.09	37.00	33.70	54.51
	CoNLL-12	14.18	27.91	38.79	50.22	34.48	61.21
test	ASN	10.86	21.45	30.10	38.91	36.17	57.29
	CoNLL-12	11.43	24.02	33.75	43.57	30.91	53.48
	ARRAU _{all}	9.60	17.87	26.02	33.83	34.32	57.70
	ARRAU _{nominal}	9.50	18.02	25.97	34.75	35.70	54.69
	ARRAU _{pronominal}	7.57	16.19	24.81	33.02	32.17	60.16
	average	9.79	19.51	28.13	36.82	33.85	56.66

Table 4.11 The results of the baselines ($\mathbf{BL}_{\text{dist,sent}}$, $\mathbf{BL}_{\text{dist,tag}}$, \mathbf{BL}_{tag}) in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_1 = \{\text{AnaphS}_{-1}, \text{AnaphS}\}$. See Section 4.5.1 on page 92 for other details.

gold training (ASN, CoNLL-12)		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	56.54	66.92	73.85	79.23	64.25	83.46
	CoNLL-12	44.00	54.00	60.00	68.00	52.54	72.00
test	ASN	39.25	54.84	64.52	68.39	59.31	84.84
	CoNLL-12	36.67	48.33	55.00	60.83	53.43	73.33
	ARRAU _{all}	35.94	44.25	51.59	57.37	52.01	69.78
	ARRAU _{nominal}	38.85	46.92	53.94	60.48	55.82	75.58
	ARRAU _{pronominal}	31.64	40.70	48.60	53.98	46.75	63.57
	average-test	36.47	47.01	54.73	60.21	53.46	73.42
silver training (NYT, WSJ)		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	47.31	61.92	69.23	74.23	69.29	85.38
	CoNLL-12	14.00	30.00	42.00	51.00	44.86	72.00
test	ASN	36.34	54.52	66.77	74.84	67.44	90.32
	CoNLL-12	5.83	17.50	30.83	42.50	41.95	65.83
	ARRAU _{all}	34.29	47.81	54.03	64.48	60.53	76.67
	ARRAU _{nominal}	34.42	49.62	56.92	67.12	65.69	84.04
	ARRAU _{pronominal}	34.44	45.03	49.77	60.29	53.17	66.32
	average-test	29.07	42.89	51.67	61.84	57.76	76.64
mixed training (gold+silver)		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	48.08	65.77	76.54	81.15	68.51	85.38
	CoNLL-12	31.00	40.00	50.00	52.00	43.94	62.00
test	ASN	36.67	55.38	65.16	71.29	62.96	87.74
	CoNLL-12	19.17	24.17	30.00	36.67	35.64	48.33
	ARRAU _{all}	38.29	55.27	65.24	71.90	58.67	75.68
	ARRAU _{nominal}	42.21	62.31	71.25	78.27	65.59	83.27
	ARRAU _{pronominal}	32.87	45.03	56.73	63.04	49.14	65.20
	average-test	33.84	48.43	57.67	64.23	54.40	72.05
BL_{dist,tag}		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	23.76	41.68	53.66	61.46	48.79	69.57
	CoNLL-12	28.57	52.75	66.78	74.55	48.13	59.12
test	ASN	23.14	41.37	54.55	63.45	49.62	71.75
	CoNLL-12	24.55	45.92	59.11	68.42	44.43	58.39
	ARRAU _{all}	20.18	36.69	48.76	56.97	47.23	63.49
	ARRAU _{nominal}	20.97	37.47	50.08	59.84	51.04	67.64
	ARRAU _{pronominal}	19.37	37.26	50.33	57.78	42.47	56.30
	average-test	21.64	39.74	52.57	61.29	46.96	63.51

Table 4.12 MR-LSTM results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_1 = \{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers that are marked in **bold** indicate the best result for the corresponding evaluation dataset across different training data types: gold, silver, mixed. See Section 4.5.2 on page 92 for other details.

gold training (ASN, CoNLL) without VERB		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	50.00	62.69	69.62	74.23	61.81	80.77
	CoNLL-12	16.00	33.00	42.00	53.00	44.22	78.00
test	ASN	38.92	57.42	68.71	74.52	60.94	89.03
	CoNLL-12	11.67	22.50	32.50	40.00	44.20	76.67
	ARRAU _{all}	41.49	51.49	58.83	65.59	55.51	80.00
	ARRAU _{nominal}	49.33	60.48	67.02	72.88	61.13	86.83
	ARRAU _{pronominal}	30.70	39.12	47.60	55.56	47.72	70.53
	average	34.42	46.20	54.93	61.71	53.90	80.61
silver training (NYT, WSJ) without VERB		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	50.77	64.23	72.69	76.92	70.16	84.23
	CoNLL-12	14.00	31.00	47.00	54.00	44.30	71.00
test	ASN	37.31	54.84	67.10	75.48	66.88	90.32
	CoNLL-12	7.50	19.17	30.00	39.17	41.73	65.00
	ARRAU _{all}	36.38	49.81	59.14	64.70	60.23	76.44
	ARRAU _{nominal}	38.46	53.08	62.69	70.48	65.83	84.04
	ARRAU _{pronominal}	33.33	44.97	53.98	56.61	52.26	65.79
	average	30.60	44.37	54.58	61.29	57.38	76.32

Table 4.13 MR-LSTM **without the VERB feature** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1 = \{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in **bold** indicate the cases when the result obtained with the MR-LSTM_{VERB} model is better than the result obtained with the full MR-LSTM model. See Section 4.5.3 on page 95 for other details.

gold training (ASN, CoNLL) without DIST		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	11.15	20.77	28.85	35.77	33.05	59.23
	CoNLL-12	29.00	40.00	46.00	54.00	40.07	65.00
test	ASN	9.78	16.88	23.12	29.89	31.04	62.26
	CoNLL-12	20.00	30.00	36.67	40.83	40.18	67.50
	ARRAU _{all}	7.11	13.33	19.52	25.97	28.46	59.37
	ARRAU _{nominal}	7.79	12.79	19.52	25.29	29.83	56.92
	ARRAU _{pronominal}	6.37	14.33	19.65	27.02	26.41	62.51
	average	10.21	17.47	23.70	29.80	31.18	61.71
	average	10.21	17.47	23.70	29.80	31.18	61.71
silver training (NYT, WSJ) without DIST		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	28.85	40.77	51.54	60.00	51.99	70.77
	CoNLL-12	12.00	31.00	35.00	39.00	36.45	65.00
test	ASN	22.69	37.63	47.63	56.34	50.55	69.68
	CoNLL-12	11.67	18.33	22.50	34.17	38.57	66.67
	ARRAU _{all}	20.41	31.30	41.97	51.75	48.70	69.68
	ARRAU _{nominal}	21.25	32.79	42.40	53.37	53.78	73.65
	ARRAU _{pronominal}	19.12	29.18	41.29	49.77	41.63	64.15
	average	19.03	29.85	39.16	49.08	46.65	68.77
	average	19.03	29.85	39.16	49.08	46.65	68.77

Table 4.14 MR-LSTM **without the DISTANCE feature** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1 = \{\text{AnaphS}_{-1}, \text{AnaphS}\}$. See Section 4.5.3 on page 95 for other details.

$\mathbf{BL}_{\text{dist,sent}}$		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	36.35	-	-	-	46.68	58.93
	CoNLL-12	0.31	-	-	-	26.55	49.59
test	ASN	22.96	-	-	-	42.39	59.25
	CoNLL-12	3.19	-	-	-	26.84	50.34
	ARRAU _{all}	30.65	-	-	-	40.30	53.31
	ARRAU _{nominal}	34.36	-	-	-	42.89	54.83
	ARRAU _{pronominal}	25.21	-	-	-	37.46	52.94
	average	23.27	-	-	-	37.98	54.13
$\mathbf{BL}_{\text{dist,tag}}$		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	23.14	40.82	52.64	60.90	46.85	58.93
	CoNLL-12	31.12	54.80	69.21	77.22	47.24	49.59
test	ASN	22.03	40.60	53.58	63.03	46.35	59.25
	CoNLL-12	26.12	46.27	61.36	69.51	43.16	50.34
	ARRAU _{all}	21.80	37.67	50.65	59.55	44.66	53.31
	ARRAU _{nominal}	22.97	38.26	51.24	60.45	47.23	54.83
	ARRAU _{pronominal}	22.32	38.07	50.65	59.10	42.67	52.94
	average	23.05	40.17	53.50	62.33	44.81	54.13
\mathbf{BL}_{tag}		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	7.12	14.24	20.30	27.22	26.16	37.34
	CoNLL-12	10.00	18.67	28.27	36.22	27.83	42.96
test	ASN	7.19	14.30	21.91	28.00	27.83	38.57
	CoNLL-12	8.71	16.81	23.97	30.43	25.02	40.52
	ARRAU _{all}	7.67	14.06	19.86	26.00	27.62	40.35
	ARRAU _{nominal}	9.16	15.78	21.96	28.72	30.69	42.03
	ARRAU _{pronominal}	5.98	12.27	18.81	24.69	24.94	44.07
	average	7.74	14.64	21.30	27.57	27.22	41.11

Table 4.15 The results of the baselines ($\mathbf{BL}_{\text{dist,sent}}$, $\mathbf{BL}_{\text{dist,tag}}$, \mathbf{BL}_{tag}) in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_2 = \{\text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. A number marked in **bold** indicates that any variant of our MR-LSTM model does not beat this baseline result. See Section 4.5.1 on page 92 for other details.

gold training (ASN, CoNLL-12)		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	44.64	50.00	53.21	59.64	58.33	71.43
	CoNLL-12	26.75	39.25	44.25	52.25	44.43	62.25
test	ASN	30.00	37.65	42.35	46.76	51.33	67.65
	CoNLL-12	28.06	37.22	42.22	42.22	41.68	53.61
	ARRAU _{all}	26.33	30.61	33.06	34.90	38.97	48.78
	ARRAU _{nominal}	29.67	34.22	37.78	38.44	41.35	48.33
	ARRAU _{pronominal}	20.00	24.00	24.00	27.50	35.33	50.00
	average-test	26.81	32.74	35.88	37.97	41.73	53.67
	silver training (NYT, WSJ)		s@1	s@2	s@3	s@4	token-F1
dev	ASN	45.00	56.43	65.36	70.36	66.22	80.71
	CoNLL-12	6.25	23.25	34.75	42.75	39.57	63.25
test	ASN	30.88	47.65	59.71	67.35	61.82	81.76
	CoNLL-12	8.06	18.89	30.56	36.94	40.63	64.72
	ARRAU _{all}	28.57	40.41	48.98	56.53	54.47	68.98
	ARRAU _{nominal}	27.56	40.89	49.11	57.00	57.35	72.00
	ARRAU _{pronominal}	29.75	39.50	48.50	55.50	50.48	65.00
	average-test	24.96	37.47	47.37	54.67	52.95	70.49
	mixed training (gold+silver)		s@1	s@2	s@3	s@4	token-F1
dev	ASN	42.14	57.50	67.14	72.50	62.95	77.50
	CoNLL-12	4.25	8.50	11.50	12.50	19.04	30.00
test	ASN	31.76	46.47	52.94	57.94	56.23	75.00
	CoNLL-12	2.50	6.39	8.89	11.39	17.76	28.89
	ARRAU _{all}	29.39	45.10	54.08	59.80	50.82	65.71
	ARRAU _{nominal}	28.67	46.44	56.00	62.67	53.64	66.44
	ARRAU _{pronominal}	30.50	43.00	51.75	55.75	46.69	65.00
	average-test	24.56	37.48	44.73	49.51	45.03	60.21
	BL_{dist,tag}		s@1	s@2	s@3	s@4	token-F1
dev	ASN	23.14	40.82	52.64	60.90	46.85	58.93
	CoNLL-12	<u>31.12</u>	<u>54.80</u>	<u>69.21</u>	<u>77.22</u>	<u>47.24</u>	49.59
test	ASN	22.03	40.60	53.58	63.03	46.35	59.25
	CoNLL-12	26.12	<u>46.27</u>	<u>61.36</u>	<u>69.51</u>	43.16	50.34
	ARRAU _{all}	21.80	37.67	50.65	59.55	44.66	53.31
	ARRAU _{nominal}	22.97	38.26	51.24	60.45	47.23	54.83
	ARRAU _{pronominal}	22.32	38.07	50.65	<u>59.10</u>	42.67	52.94
	average-test	23.05	40.17	<u>53.50</u>	<u>62.33</u>	44.81	54.13

Table 4.16 MR-LSTM results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_2 = \{\text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers that are marked in **bold** indicate the best result for the corresponding evaluation dataset across different training data types: gold, silver, mixed. The $\text{BL}_{\text{dist,tag}}$ results that are better than all MR-LSTM results for the corresponding evaluation dataset are underlined. See Section 4.5.2 on page 92 for other details.

gold training (ASN, CoNLL) without VERB		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	49.64	57.86	68.57	72.14	65.88	80.71
	CoNLL-12	6.00	22.25	29.25	35.25	38.32	64.25
test	ASN	32.65	48.53	60.00	66.47	58.67	78.82
	CoNLL-12	5.83	18.33	26.67	37.22	41.15	66.39
	ARRAU _{all}	37.96	49.59	54.90	59.39	54.55	68.37
	ARRAU _{nominal}	43.67	53.67	58.67	64.33	57.99	71.67
	ARRAU _{pronominal}	28.50	43.50	49.75	52.75	49.52	64.00
	average	29.72	42.72	50.00	56.03	52.38	69.85
silver training (NYT, WSJ) without VERB		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	47.86	58.21	66.07	71.43	68.00	81.07
	CoNLL-12	4.00	23.25	34.75	47.00	37.56	62.25
test	ASN	34.71	51.18	61.76	69.12	62.70	81.76
	CoNLL-12	6.39	18.06	29.72	38.61	40.12	63.89
	ARRAU _{all}	33.27	46.33	52.24	57.76	55.50	68.78
	ARRAU _{nominal}	32.56	49.00	55.33	60.33	58.13	71.67
	ARRAU _{pronominal}	34.50	42.50	47.50	53.50	52.03	65.00
	average	28.28	41.41	49.31	55.86	53.70	70.22

Table 4.17 MR-LSTM **without the VERB feature** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_2 = \{AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$. Numbers marked in **bold** indicate the cases when the result obtained with the MR-LSTM_{VERB} model is better than the result obtained with the full MR-LSTM model. See Section 4.5.3 on page 95 for other details.

gold training (ASN, CoNLL) without DIST		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	7.50	11.07	18.57	22.14	23.94	32.50
	CoNLL-12	15.50	21.50	25.50	29.75	27.67	31.00
test	ASN	5.00	9.41	12.35	16.76	23.40	33.82
	CoNLL-12	10.56	13.89	18.61	22.78	24.89	33.06
	ARRAU _{all}	7.35	12.65	15.10	23.27	23.98	32.65
	ARRAU _{nominal}	6.67	11.33	13.00	22.00	24.37	30.00
	ARRAU _{pronominal}	8.75	14.75	18.25	24.75	23.13	35.75
	average	7.66	12.41	15.46	21.91	23.95	33.06
silver training (NYT, WSJ) without DIST		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	19.29	28.57	38.57	43.93	40.62	52.50
	CoNLL-12	4.00	9.00	15.00	21.25	27.20	39.75
test	ASN	14.12	21.76	30.00	37.65	36.08	45.88
	CoNLL-12	7.22	12.78	16.11	20.28	31.43	46.39
	ARRAU _{all}	12.04	22.24	30.00	36.53	34.98	48.37
	ARRAU _{nominal}	12.78	23.11	31.78	38.00	39.57	51.56
	ARRAU _{pronominal}	11.00	20.50	26.50	34.25	28.21	43.25
	average	11.43	20.08	26.88	33.34	34.05	47.09

Table 4.18 MR-LSTM **without the DISTANCE feature** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_2 = \{\text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. See Section 4.5.3 on page 95 for other details.

$\mathbf{BL}_{\text{dist,sent}}$		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	34.55	-	-	-	44.28	55.73
	CoNLL-12	0.31	-	-	-	25.87	48.16
test	ASN	21.51	-	-	-	39.41	54.91
	CoNLL-12	3.33	-	-	-	26.32	48.97
	ARRAU _{all}	28.24	-	-	-	37.57	49.90
	ARRAU _{nominal}	30.89	-	-	-	38.98	49.81
	ARRAU _{pronominal}	23.98	-	-	-	35.43	50.41
	average	21.59	-	-	-	35.54	50.80
$\mathbf{BL}_{\text{dist,tag}}$		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	22.47	40.57	53.69	61.80	44.96	55.73
	CoNLL-12	31.12	54.29	68.86	77.92	47.36	48.16
test	ASN	22.29	39.93	53.74	61.90	45.35	54.91
	CoNLL-12	23.33	44.42	58.42	68.39	41.05	48.97
	ARRAU _{all}	22.25	38.39	50.65	58.20	43.39	49.90
	ARRAU _{nominal}	23.34	38.02	51.14	60.96	45.52	49.81
	ARRAU _{pronominal}	20.51	38.02	50.80	58.59	39.87	50.41
	average	22.34	39.76	52.95	61.61	43.04	50.80
\mathbf{BL}_{tag}		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	5.63	10.65	15.90	20.86	23.55	30.57
	CoNLL-12	7.55	15.51	22.55	28.16	21.67	32.86
test	ASN	5.86	11.86	17.20	21.46	23.65	30.80
	CoNLL-12	6.92	13.25	19.40	24.44	20.70	30.94
	ARRAU _{all}	6.25	11.65	16.84	21.57	24.57	34.65
	ARRAU _{nominal}	7.04	12.74	17.77	22.45	26.63	33.47
	ARRAU _{pronominal}	4.95	10.15	14.23	18.11	21.38	34.90
	average	6.20	11.93	17.09	21.61	23.39	32.95

Table 4.19 The results of the baselines ($\mathbf{BL}_{\text{dist,sent}}$, $\mathbf{BL}_{\text{dist,tag}}$, \mathbf{BL}_{tag}) in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_3 = \{\text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. A number marked in **bold** indicates that any variant of our MR-LSTM model does not beat this baseline result. See Section 4.5.1 on page 92 for other details.

gold training (ASN, CoNLL-12)		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	48.69	55.48	60.12	64.84	61.30	75.60
	CoNLL-12	19.25	37.75	40.75	42.00	43.81	75.75
test	ASN	30.57	42.86	49.71	53.43	54.11	71.71
	CoNLL-12	15.71	28.21	36.90	43.93	42.87	69.76
	ARRAU _{all}	31.76	39.41	44.71	48.43	46.64	60.00
	ARRAU _{nominal}	34.38	41.88	45.31	48.44	47.70	57.50
	ARRAU _{pronominal}	27.50	35.67	44.33	48.83	45.01	63.83
	average-test	27.99	37.60	44.19	48.61	47.27	64.56
silver training (NYT, WSJ)		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	40.44	53.37	61.59	66.59	63.44	78.06
	CoNLL-12	7.00	23.25	32.75	39.75	39.64	63.25
test	ASN	27.43	46.86	56.57	62.86	58.62	78.00
	CoNLL-12	7.86	20.36	28.21	32.38	40.52	64.29
	ARRAU _{all}	27.65	40.59	48.82	56.27	52.14	66.27
	ARRAU _{nominal}	26.09	39.69	47.81	55.00	54.09	67.66
	ARRAU _{pronominal}	30.17	42.17	50.17	57.67	49.50	64.67
	average-test	23.84	37.93	46.32	52.84	50.97	68.18
mixed training (gold+silver)		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	40.79	54.80	65.56	72.02	61.83	78.41
	CoNLL-12	8.25	17.50	25.75	30.75	37.67	60.25
test	ASN	28.00	41.43	54.00	60.86	56.26	76.57
	CoNLL-12	9.52	13.69	19.05	23.57	36.38	54.40
	ARRAU _{all}	30.98	40.39	52.35	60.39	52.76	68.24
	ARRAU _{nominal}	29.22	40.00	53.59	61.41	54.65	66.88
	ARRAU _{pronominal}	34.33	41.83	51.33	59.33	49.91	70.17
	average-test	26.41	35.47	46.07	53.11	49.99	67.25
BL_{dist.tag}		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	22.47	40.57	53.69	61.80	44.96	55.73
	CoNLL-12	<u>31.12</u>	<u>54.29</u>	<u>68.86</u>	<u>77.92</u>	<u>47.36</u>	48.16
test	ASN	22.29	39.93	53.74	61.90	45.35	54.91
	CoNLL-12	<u>23.33</u>	<u>44.42</u>	<u>58.42</u>	<u>68.39</u>	41.05	48.97
	ARRAU _{all}	22.25	38.39	50.65	58.20	43.39	49.90
	ARRAU _{nominal}	23.34	38.02	51.14	60.96	45.52	49.81
	ARRAU _{pronominal}	20.51	38.02	50.80	58.59	39.87	50.41
	average-test	22.34	39.76	<u>52.95</u>	<u>61.61</u>	43.04	50.80

Table 4.20 MR-LSTM results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_3 = \{\text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers that are marked in **bold** indicate the best result for the corresponding evaluation dataset across different training data types: gold, silver, mixed. Baseline results that are better than all MR-LSTM results for the corresponding evaluation dataset are underlined. See Section 4.5.2 on page 92 for other details.

gold training (ASN, CoNLL) without VERB		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	47.58	58.65	64.80	68.37	63.63	76.59
	CoNLL-12	8.25	26.75	39.75	44.75	40.65	64.25
test	ASN	34.00	50.86	59.14	65.71	58.78	76.86
	CoNLL-12	12.02	20.36	32.74	40.24	42.10	63.45
	ARRAU _{all}	37.84	46.27	52.55	57.25	54.31	65.49
	ARRAU _{nominal}	40.78	49.53	55.78	60.78	56.86	67.66
	ARRAU _{pronominal}	33.83	41.67	47.67	52.00	50.56	62.67
	average	31.70	41.74	49.58	55.20	52.52	67.22
silver training (NYT, WSJ) without VERB		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	45.83	55.12	64.44	68.73	65.50	77.66
	CoNLL-12	6.00	24.25	35.50	45.00	39.39	62.25
test	ASN	31.14	47.43	57.71	65.71	59.83	78.00
	CoNLL-12	7.02	19.52	30.36	36.90	41.48	63.45
	ARRAU _{all}	34.12	44.31	51.96	56.86	54.76	65.88
	ARRAU _{nominal}	34.06	45.00	53.28	58.28	57.18	67.66
	ARRAU _{pronominal}	34.67	43.17	50.17	54.67	51.38	63.67
	average	28.20	39.89	48.70	54.49	52.93	67.73

Table 4.21 MR-LSTM **without the VERB feature** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_3 = \{\text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in **bold** indicate the cases when the result obtained with the MR-LSTM_{VERB} model is better than the result obtained with the full MR-LSTM model. See Section 4.5.3 on page 95 for other details.

gold training (ASN, CoNLL) without DIST		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	7.50	12.14	16.11	20.40	23.53	33.37
	CoNLL-12	10.25	15.25	22.25	26.25	22.70	25.25
test	ASN	4.29	8.29	11.14	14.57	20.93	28.86
	CoNLL-12	9.52	15.71	22.74	24.40	20.84	27.86
	ARRAU _{all}	4.12	6.86	10.20	16.27	21.38	30.98
	ARRAU _{nominal}	4.69	8.75	12.19	18.75	21.93	30.47
	ARRAU _{pronominal}	3.00	3.50	6.50	11.83	20.72	32.33
	average	5.12	8.62	12.55	17.17	21.16	30.10
silver training (NYT, WSJ) without DIST		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	12.14	19.33	23.29	29.01	31.78	33.65
	CoNLL-12	6.00	9.00	12.00	12.00	22.72	23.75
test	ASN	9.14	16.86	22.29	28.29	27.92	28.86
	CoNLL-12	8.33	10.83	12.50	15.36	23.44	27.38
	ARRAU _{all}	8.82	16.67	21.18	25.88	27.33	31.57
	ARRAU _{nominal}	8.13	15.47	21.41	27.34	29.06	30.78
	ARRAU _{pronominal}	9.50	18.50	20.50	23.00	24.25	32.33
	average	8.78	15.67	19.57	23.97	26.40	30.18

Table 4.22 MR-LSTM **without the DISTANCE feature** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_3 = \{AnaphS_{-3}, AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$. See Section 4.5.3 on page 95 for other details.

$\mathbf{BL}_{\text{dist,sent}}$		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	32.53	-	-	-	42.60	53.99
	CoNLL-12	0.31	-	-	-	25.60	47.45
test	ASN	19.97	-	-	-	37.15	51.57
	CoNLL-12	3.33	-	-	-	26.00	48.29
	ARRAU _{all}	26.60	-	-	-	34.89	46.26
	ARRAU _{nominal}	29.34	-	-	-	36.20	46.25
	ARRAU _{pronominal}	23.72	-	-	-	35.09	49.80
	average	20.59	-	-	-	33.87	48.43
$\mathbf{BL}_{\text{dist,tag}}$		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	23.88	41.16	53.83	61.16	45.13	53.99
	CoNLL-12	31.02	54.50	68.93	78.90	47.41	47.45
test	ASN	23.58	40.60	53.71	62.17	44.93	51.57
	CoNLL-12	23.33	43.21	59.43	69.37	40.98	48.29
	ARRAU _{all}	22.26	37.78	50.72	58.61	42.46	46.26
	ARRAU _{nominal}	23.38	39.31	51.95	60.00	44.41	46.25
	ARRAU _{pronominal}	19.85	37.49	50.16	57.46	39.76	49.80
	average	22.48	39.68	53.19	61.52	42.51	48.43
\mathbf{BL}_{tag}		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	4.80	9.79	13.18	17.25	20.88	25.98
	CoNLL-12	5.92	12.35	18.88	23.78	18.87	25.61
test	ASN	5.01	10.14	14.88	19.26	22.04	26.91
	CoNLL-12	5.73	10.60	15.38	19.23	18.57	27.86
	ARRAU _{all}	5.12	9.70	14.16	18.44	22.44	30.04
	ARRAU _{nominal}	6.07	11.36	15.56	19.91	23.97	29.43
	ARRAU _{pronominal}	4.64	8.83	12.65	15.66	19.92	28.57
	average	5.31	10.13	14.53	18.50	21.39	28.56

Table 4.23 The results of the baselines ($\mathbf{BL}_{\text{dist,sent}}$, $\mathbf{BL}_{\text{dist,tag}}$, \mathbf{BL}_{tag}) in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_4 = \{\text{AnaphS}_{-4}, \text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. A number marked in **bold** indicates that any variant of our MR-LSTM model does not beat this baseline result. See Section 4.5.1 on page 92 for other details.

gold training (ASN, CoNLL-12)		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	42.41	54.14	60.00	69.31	60.50	75.52
	CoNLL-12	26.50	37.75	41.75	45.00	44.96	66.50
test	ASN	28.92	40.18	49.64	56.67	51.94	71.53
	CoNLL-12	18.21	35.24	45.24	51.07	45.12	68.93
	ARRAU _{all}	24.99	34.96	45.42	51.08	45.37	61.83
	ARRAU _{nominal}	24.71	34.12	47.06	52.06	46.45	60.00
	ARRAU _{pronominal}	24.00	35.00	45.00	51.50	44.66	66.83
	average-test	24.17	35.90	46.47	52.48	46.71	65.83
	silver training (NYT, WSJ)		s@1	s@2	s@3	s@4	token-F1
dev	ASN	41.72	53.10	61.03	68.97	64.68	78.28
	CoNLL-12	6.25	22.25	33.75	39.75	39.62	63.25
test	ASN	30.18	43.96	54.50	61.80	58.19	75.05
	CoNLL-12	7.02	21.19	29.05	34.05	40.94	64.29
	ARRAU _{all}	28.09	39.78	45.61	52.96	52.18	63.99
	ARRAU _{nominal}	24.71	37.94	45.88	53.24	54.11	65.00
	ARRAU _{pronominal}	32.67	41.17	47.17	54.17	50.08	64.17
	average-test	24.53	36.81	44.44	51.24	51.10	66.50
	mixed training (gold+silver)		s@1	s@2	s@3	s@4	token-F1
dev	ASN	42.76	54.48	63.45	71.03	63.06	75.52
	CoNLL-12	6.00	13.25	19.50	25.50	38.90	65.25
test	ASN	26.58	41.80	51.26	58.02	53.75	68.83
	CoNLL-12	11.19	15.36	21.90	31.07	42.55	67.62
	ARRAU _{all}	28.54	38.81	49.00	56.17	49.70	62.59
	ARRAU _{nominal}	27.94	38.24	50.29	56.76	51.30	59.71
	ARRAU _{pronominal}	32.67	42.50	49.33	57.33	49.12	69.67
	average-test	25.38	35.34	44.36	51.87	49.29	65.68
	BL_{dist.tag}		s@1	s@2	s@3	s@4	token-F1
dev	ASN	23.88	41.16	53.83	61.16	45.13	53.99
	CoNLL-12	<u>31.02</u>	<u>54.50</u>	<u>68.93</u>	<u>78.90</u>	<u>47.41</u>	47.45
test	ASN	23.58	40.60	53.71	62.17	44.93	51.57
	CoNLL-12	<u>23.33</u>	<u>43.21</u>	<u>59.43</u>	<u>69.37</u>	<u>40.98</u>	48.29
	ARRAU _{all}	22.26	37.78	50.72	58.61	42.46	46.26
	ARRAU _{nominal}	23.38	39.31	51.95	60.00	44.41	46.25
	ARRAU _{pronominal}	19.85	37.49	50.16	57.46	39.76	49.80
	average-test	22.48	39.68	53.19	61.52	42.51	48.43

Table 4.24 MR-LSTM results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_4 = \{\text{AnaphS}_{-4}, \text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers that are marked in **bold** indicate the best result for the corresponding evaluation dataset across different training data types: gold, silver, mixed. Baseline results that are better than all MR-LSTM results for the corresponding evaluation dataset are underlined. See Section 4.5.2 on page 92 for other details.

gold training (ASN, CoNLL) without VERB		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	49.66	58.28	63.79	71.72	65.18	77.93
	CoNLL-12	5.25	22.25	34.25	40.25	38.06	64.00
test	ASN	30.36	46.94	55.86	63.42	56.81	74.23
	CoNLL-12	6.19	19.88	30.24	37.74	39.89	65.95
	ARRAU _{all}	38.92	48.25	53.23	56.63	53.59	63.80
	ARRAU _{nominal}	42.94	51.18	54.71	58.82	56.47	65.29
	ARRAU _{pronominal}	34.00	44.83	52.50	54.50	50.46	63.17
	average	30.48	42.22	49.31	54.22	51.44	66.49
silver training set (NYT, WSJ) without VERB		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	45.86	56.55	64.48	69.66	66.19	77.93
	CoNLL-12	6.00	22.25	34.50	44.75	39.43	62.25
test	ASN	31.26	46.67	56.67	61.80	58.49	75.05
	CoNLL-12	5.36	19.52	29.52	36.07	40.25	63.45
	ARRAU _{all}	30.81	42.96	48.63	54.29	53.22	63.61
	ARRAU _{nominal}	30.59	43.53	49.12	55.29	55.29	65.00
	ARRAU _{pronominal}	34.17	44.17	49.67	54.17	51.83	63.17
	average	26.44	39.37	46.72	52.32	51.82	66.06

Table 4.25 MR-LSTM **without the VERB feature** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_4 = \{\text{Anaph}_{-4}, \text{AnaphS}_{-3}, \text{Anaph}_{-2}, \text{Anaph}_{-1}, \text{AnaphS}\}$. Numbers marked in **bold** indicate the cases when the result obtained with the MR-LSTM_{VERB} model is better than the result obtained with the full MR-LSTM model. See Section 4.5.3 on page 95 for other details.

gold training set (ASN, CoNLL) without DIST		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	3.45	9.31	13.45	19.66	22.07	28.28
	CoNLL-12	14.25	19.50	22.50	23.50	27.14	31.50
test	ASN	5.68	8.92	11.89	15.41	20.00	25.68
	CoNLL-12	5.71	7.38	9.40	14.40	16.11	24.40
	ARRAU _{all}	3.96	7.55	10.57	13.58	19.05	24.69
	ARRAU _{nominal}	5.29	8.53	12.06	15.00	20.43	21.76
	ARRAU _{pronominal}	1.50	5.50	7.50	10.50	17.35	29.00
	average	4.43	7.58	10.28	13.78	18.59	25.11
silver training (NYT, WSJ) without DIST		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	6.90	13.10	18.97	23.10	25.79	27.59
	CoNLL-12	8.00	9.00	13.50	15.50	23.32	31.75
test	ASN	3.51	8.38	12.43	15.68	20.72	23.24
	CoNLL-12	3.33	6.67	8.33	9.17	14.80	25.83
	ARRAU _{all}	5.47	11.70	16.60	20.75	24.36	28.95
	ARRAU _{nominal}	6.76	13.53	17.65	21.18	27.81	30.88
	ARRAU _{pronominal}	3.00	8.00	14.00	19.00	19.74	28.83
	average	4.42	9.65	13.80	17.15	21.49	27.55

Table 4.26 MR-LSTM **without the DISTANCE feature** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_4 = \{AnaphS_{-4}, AnaphS_{-3}, AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$. See Section 4.5.3 on page 95 for other details.

4.6 MR-LSTM with Adversarial Training

In the previous section, we showed that training with the large silver data does not perform better than training with the small gold data for all evaluation datasets. We speculated that this is due to the fact that the silver data represents some types of anaphors better than others. For example, anaphors in the silver data are always pronouns, therefore it is expected that the silver data represents the pronominal part of the ARRAU dataset better than the nominal part. We speculated that blending the gold and silver training data could alleviate this problem, but we showed that mixing this two types of training data does not always work. In Section 4.1, we noted that we could potentially ensure that our models do not fit specificities of the extracted (silver) data if we train our models with mixed data and adversarial training. This training strategy could potentially enforce learning representations that are relevant for all anaphors encountered during the training, without favoring one type over the other.

The adversarial training works as follows. We use two dataset type discriminators: one after the LSTM layer and one after the feed-forward (FF-ELU) layer that produces representations of the linguistic features. The dataset discriminator (Figure 4.3, green triangles before "silver or gold" output) predicts the class of data $c_{\text{data}} \in \{\text{gold}, \text{silver}\}$ that the current batch of data belongs to based on the given representation. This discriminator should perform poorly if the LSTM and the feature FF-ELU layer are indeed invariant to the class of data (gold or silver). Since this is what we want, we need to update the LSTM's parameters such that the LSTM discriminator's cross-entropy loss $L_{\text{discrm}}^{\text{LSTM}}$ in (4.8) is maximized with respect to the LSTM's parameters θ_{LSTM} . The same holds for the discriminator after the FF-ELU layer, but for simplicity we focus on the discriminator after the LSTM layer. To maximize the LSTM discriminator's loss $L_{\text{discrm}}^{\text{LSTM}}$ with respect to the LSTM's parameters θ_{LSTM} , we make a step in the direction opposite of the direction of the negative gradient. Therefore, there is a plus sign before the gradient $\nabla L_{\text{discrm}}^{\text{LSTM}}(\theta_{\text{LSTM}})$ in (4.9). The symbol α is used for the learning rate and the symbol λ for the scaling factor of the discriminator's loss. The scaling factor is introduced because we do not want that the discriminator's loss disrupts the dynamics of the max-margin loss $L_{\text{max-margin}}$.

At the same time we want the discriminators to challenge the LSTM layer and the feature FF-ELU layer. Therefore, we update their parameters to minimize their cross-entropy losses. This is what we typically do when we train a machine learning system with a gradient descent. Note that in (4.10) there is a negative sign before the gradient since we move in the direction of the negative gradient. The whole process is illustrated in Figure 4.4.

$$L_{\text{dscrm}}^{\text{LSTM}} = -\left(c_{\text{data}} \cdot \log(\text{softmax}(\theta_{\text{dscrm}})) + (1 - c_{\text{data}}) \cdot \log(1 - \text{softmax}(\theta_{\text{dscrm}}))\right) \quad (4.8)$$

$$\theta_{\text{LSTM}} \leftarrow \theta_{\text{LSTM}} + \alpha \lambda \nabla L_{\text{dscrm}}^{\text{LSTM}}(\theta_{\text{LSTM}}) - \alpha \nabla L_{\text{max-margin}}(\theta_{\text{LSTM}}) \quad (4.9)$$

$$\theta_{\text{dscrm}} \leftarrow \theta_{\text{dscrm}} - \alpha \nabla L_{\text{dscrm}}^{\text{LSTM}}(\theta_{\text{dscrm}}) \quad (4.10)$$

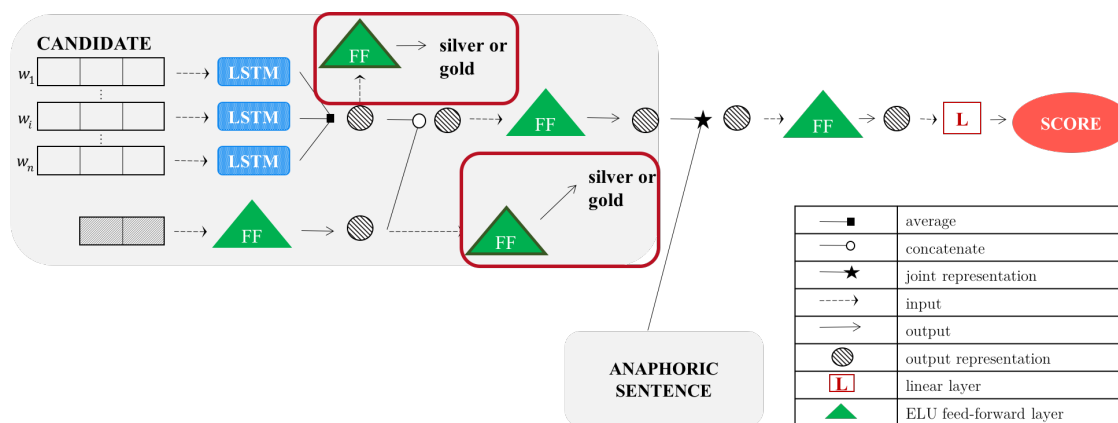


Fig. 4.3 Adversarial MR-LSTM (forward pass).

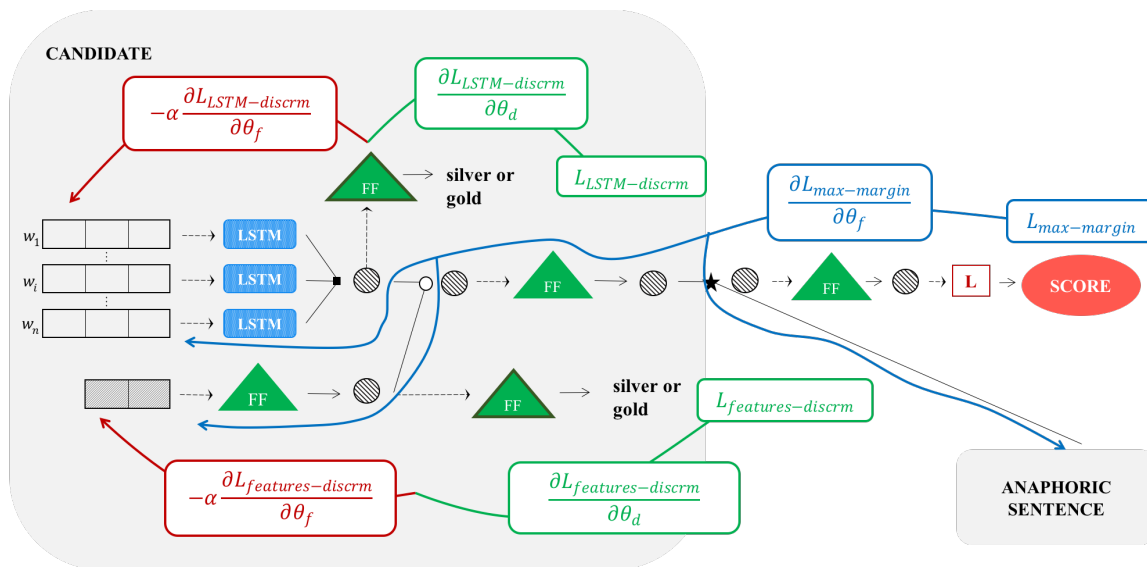


Fig. 4.4 Adversarial MR-LSTM (backward pass).

4.6.1 Results

We report results with adversarial training in Tables 4.29, 4.30, 4.31, and 4.32, on pages 118 and 119. From these tables, we detect when adversarial training is beneficial and summarize such cases in Table 4.27. The adversarial training is helpful in the majority of cases for the windows W_1 and W_2 , but not for W_3 and W_4 . The larger candidate extraction windows results in more candidates which potentially makes a challenging minmax optimization unstable. We can not trace the dynamics of the discriminator’s gradient anymore to check if some instabilities occurred. However, we could train the model again and examine this.

In Table 4.28, we report which training strategy (gold, silver, mixed, or mixed with adversarial training) is the best. The adversarial training improved the mixed training only in three configurations. Therefore, making the mixed training effective remains a challenge.

		success@1						token-F1			
		W_1	W_2	W_3	W_4			W_1	W_2	W_3	W_4
dev	ASN	✓	✓	✗	✗	dev	ASN	✗	✓	✗	✗
	CoNLL-12	✓	✓	✗	✗		CoNLL-12	✓	✓	✗	✗
test	ASN	✓	✓	✓	✓	test	ASN	✓	✓	✗	✓
	CoNLL-12	✓	✓	✗	✗		CoNLL-12	✓	✓	✗	✗
	ARRAU _{all}	✓	✓	✗	✗		ARRAU _{all}	✓	✗	✗	✓
	ARRAU _{nominal}	✗	✗	✗	✗		ARRAU _{nominal}	✗	✗	✗	✓
	ARRAU _{pronominal}	✓	✓	✗	✗		ARRAU _{pronominal}	✓	✓	✗	✗
	average-test	✓	✓	✗	✓		average-test	✓	✓	✗	✓

Table 4.27 Comparison between the MR-LSTM with and without adversarial training. The symbol ✓ denotes the cases when adversarial training is beneficial and ✗ when it is not.

		success@1						token-F1			
		W_1	W_2	W_3	W_4			W_1	W_2	W_3	W_4
dev	ASN	gold	silver	gold	mixed	dev	ASN	silver	silver	silver	silver
	CoNLL-12	gold	gold	gold	gold		CoNLL-12	gold	gold	gold	gold
test	ASN	gold	mix-adv	gold	silver	test	ASN	silver	silver	silver	silver
	CoNLL-12	gold	gold	gold	gold		CoNLL-12	gold	gold	gold	gold
	ARRAU _{all}	mixed	mix-adv	gold	mixed		ARRAU _{all}	silver	silver	mixed	silver
	ARRAU _{nominal}	mixed	gold	gold	mixed		ARRAU _{nominal}	silver	silver	mixed	silver
	ARRAU _{pronominal}	silver	mixed	mixed	silver/mixed		ARRAU _{pronominal}	silver	silver	mixed	silver
	average-test	gold	gold	gold	mix-adv		average-test	silver	silver	silver	silver

Table 4.28 The best training strategies between gold, silver, mixed, and mixed with adversarial training. The window $\{AnaphS_{-i}, \dots, AnaphS\}$ is denoted by W_i . The "mix-adv" in **bold** indicates the cases where adversarial training results in the best performance between all training strategies.

mixed + adversarial training		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	50.38	65.00	74.62	80.77	67.98	84.62
	CoNLL-12	36.00	48.00	50.00	56.00	49.44	69.00
test	ASN	37.96	54.73	63.76	72.90	64.13	88.06
	CoNLL-12	22.50	34.17	39.17	44.17	40.48	58.33
	ARRAU _{all}	38.51	55.02	64.22	68.89	59.27	80.35
	ARRAU _{nominal}	40.29	60.87	70.48	74.04	64.26	84.81
	ARRAU _{pronominal}	36.02	46.67	55.15	61.46	52.37	74.15
	average-test	35.06	50.29	58.56	64.29	56.10	77.14

Table 4.29 MR-LSTM with **adversarial training** results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_1 = \{AnaphS_{-1}, AnaphS\}$. Numbers marked in **bold** indicate the cases when the adversarial training resulted in a better performance than gold, silver, or mixed training (Table 4.12 on page 100).

mixed + adversarial training		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	45.36	60.71	68.57	73.21	65.53	78.93
	CoNLL-12	9.50	12.75	16.75	18.75	20.96	29.00
test	ASN	34.41	47.65	54.12	60.00	58.41	77.65
	CoNLL-12	5.56	14.44	18.61	19.44	20.29	31.11
	ARRAU _{all}	31.02	47.55	54.69	60.61	50.66	68.16
	ARRAU _{nominal}	28.67	48.44	56.00	62.67	52.34	67.44
	ARRAU _{pronominal}	34.00	45.50	52.75	56.75	47.54	68.50
	average-test	26.73	40.72	47.23	51.89	45.85	62.57

Table 4.30 MR-LSTM with **adversarial training** results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_2 = \{AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$. Numbers marked in **bold** indicate the cases when the adversarial training resulted in a better performance than gold, silver, or mixed training (Table 4.16 on page 104).

mixed + adversarial training		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	40.44	55.48	63.77	71.63	60.81	75.87
	CoNLL-12	4.00	14.25	19.50	21.50	35.28	61.25
test	ASN	28.86	41.71	53.71	59.43	55.05	73.71
	CoNLL-12	6.19	14.52	19.88	23.21	33.69	56.43
	ARRAU _{all}	28.04	41.76	51.76	60.98	51.39	67.45
	ARRAU _{nominal}	26.09	41.56	53.59	61.09	53.27	65.63
	ARRAU _{pronominal}	31.83	42.83	49.83	61.33	48.54	70.17
	average-test	24.20	36.48	45.76	53.21	48.39	66.68

Table 4.31 MR-LSTM with **adversarial training** results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_3 = \{\text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in **bold** indicate the cases when the adversarial training resulted in a better performance than gold, silver, or mixed training (Table 4.20 on page 108).

mixed + adversarial training		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	39.66	57.24	64.83	72.07	60.57	75.52
	CoNLL-12	6.00	19.25	20.25	25.25	38.86	64.25
test	ASN	30.09	42.61	52.34	60.18	56.86	73.42
	CoNLL-12	9.52	13.21	16.90	23.57	39.68	61.43
	ARRAU _{all}	28.17	41.64	50.70	58.06	49.82	62.21
	ARRAU _{nominal}	27.94	40.29	50.59	57.06	51.94	61.18
	ARRAU _{pronominal}	32.00	46.83	53.33	61.83	48.43	66.17
	average-test	25.54	36.92	44.77	52.14	49.35	64.88

Table 4.32 MR-LSTM with **adversarial training** results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_4 = \{\text{AnaphS}_{-4}, \text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in **bold** indicate the cases when the adversarial training resulted in a better performance than gold, silver, or mixed training (Table 4.24 on page 112).

4.7 MR-LSTM with Multi-Task Learning

All variants of our MR-LSTM model calculate the score for any type of abstract anaphor using the same scorer given by Equation (4.1) on page 81. However, different decisions may be necessary for scoring candidates of different types of abstract anaphora i.e. different types of relations between the anaphor and its antecedent (see Section 4.1 on page 76). Motivated by this observation, we wonder can we improve the mixed training by replacing a single scorer for all types of anaphors with three scores for three different types of anaphors that occur in the mixed training data: events, shell nouns, and pronouns (see Figure 4.5). All other components of our model stay the same. In other words, we use the Fully-Shared (FS-MTL) model described in Chapter 2 (Section 2.4.2 on page 54). Figure 4.6 illustrates the propagation of the gradients if the model has to resolve a shell noun. Notice that the other two scorers are not affected, but all the other parameters are.

4.7.1 Results

We report results with Multi-Task Learning (MTL) in Tables 4.12, 4.16, 4.20, and 4.24, on pages 100, 104, 108, and 112, respectively. From these tables, we detect when MR-LSTM trained on the mixture of the gold and the silver data benefits from MTL and summarize those cases in Table 4.33. We notice that MTL is beneficial in the majority of cases for the windows W_1 and W_2 , but, as it was the case with adversarial training, for the window W_3 and W_4 there are mostly no improvements from MTL.

Since we repeatedly observe that the CoNLL-12 dataset differs from other datasets, it is reasonable to expect that different decisions are necessary to score candidates for the antecedent of the CoNLL-12 anaphors. We notice that the anaphors that benefit the most from MTL indeed come from the CoNLL-12 dataset.

Finally, we examine which training strategy is the best between all possibilities: gold, silver, mixed, mixed with adversarial training, and mixed with MTL. Table 4.34 provides an overview. From this table we observe once more that training on the gold data works well if we have the same type of anaphors at the evaluation time. However, given that the ARRAU corpus covers variety of anaphor types, our models benefit from the silver data. This is evident from the observation that in the majority of cases some version of the mixed training is the best training strategy for the ARRAU corpus. Figure 4.7 additionally illustrates a comparison between different training strategies for the resolution of anaphors in the test part of the ASN corpus, the test part of the CoNLL corpus, the nominal part of the ARRAU corpus ($ARRAU_{\text{nominal}}$), and the pronominal part of the ARRAU corpus ($ARRAU_{\text{pronom}}$), but only in terms of success@1 .

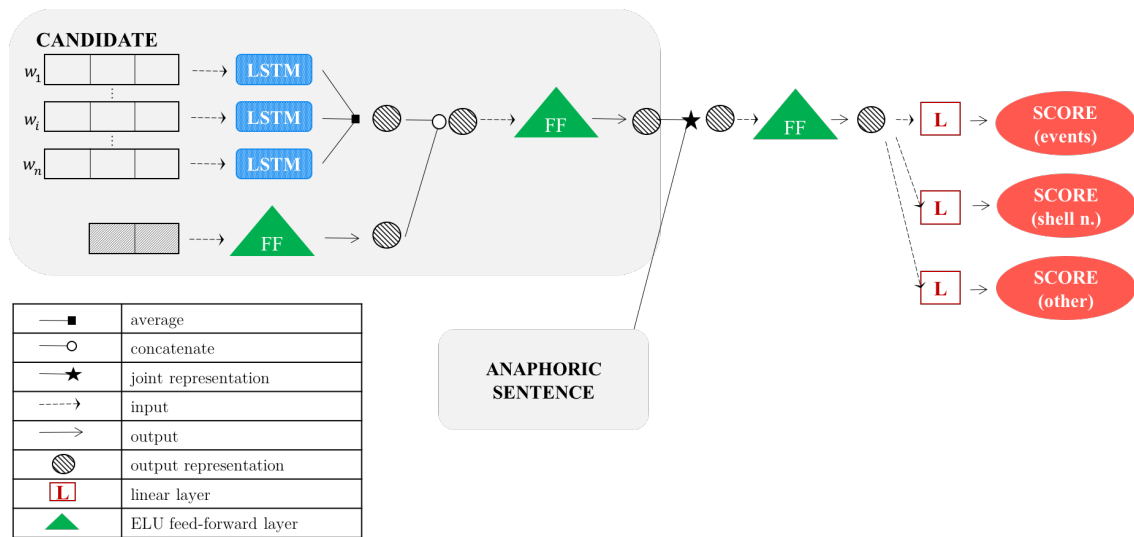


Fig. 4.5 Multi-task learning MR-LSTM.

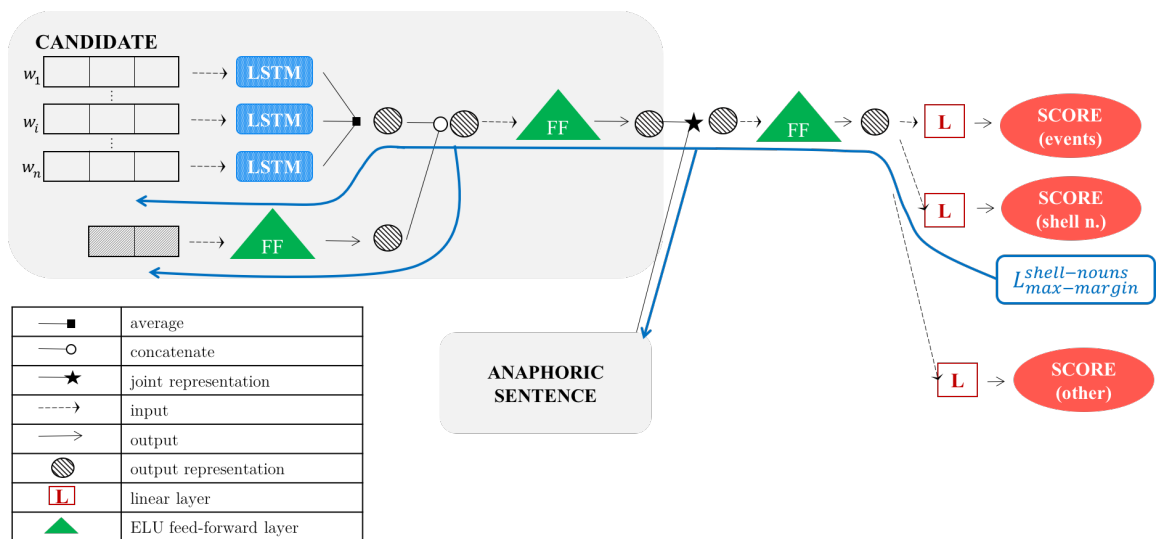


Fig. 4.6 Multi-task learning MR-LSTM (backward pass).

		success@1				token-F1			
		W ₁	W ₂	W ₃	W ₄	W ₁	W ₂	W ₃	W ₄
dev	ASN	✓	✓	✓	✗	✗	✓	✓	✗
	CoNLL-12	✓	✓	✗	✗	✓	✓	✗	✗
	ASN	✗	✗	✗	✗	✓	✓	✗	✗
	CoNLL-12	✓	✓	✗	✗	✓	✓	✓	✗
shell noun (ASN) scorer					shell noun (ASN) scorer				
test	ARRAU _{all}	✓	✗	✗	✗	✓	✗	✗	✗
	ARRAU _{nominal}	✓	✓	✗	✗	✓	✗	✗	✗
	ARRAU _{pronominal}	✓	✗	✗	✗	✓	✗	✗	✗
	average-test	✗	✓	✗	✗	✓	✓	✗	✗
	events (CoNLL-12) scorer					events (CoNLL-12) scorer			
	ARRAU _{all}	✗	✗	✗	✗	✗	✓	✗	✗
	ARRAU _{nominal}	✗	✗	✗	✗	✗	✓	✗	✓
	ARRAU _{pronominal}	✗	✗	✗	✗	✗	✓	✗	✗
	average-test	✗	✓	✗	✗	✗	✓	✗	✗
	silver scorer					silver scorer			
	ARRAU _{all}	✗	✗	✗	✓	✗	✗	✗	✓
	ARRAU _{nominal}	✗	✗	✗	✓	✗	✗	✗	✓
	ARRAU _{pronominal}	✗	✗	✗	✗	✗	✗	✗	✗
	average-test	✗	✓	✗	✗	✗	✓	✗	✗

Table 4.33 Comparison between the MR-LSTM with and without multi-task learning. The symbol ✓ denotes the cases when MR-LSTM benefits from multi-task learning and ✗ when it does not.

		success@1				token-F1			
		W ₁	W ₂	W ₃	W ₄	W ₁	W ₂	W ₃	W ₄
dev	ASN	gold	mix-mtl	gold	mixed	silver	silver	silver	silver
	CoNLL-12	gold	gold	gold	gold	gold	gold	gold	gold
test	ASN	gold	mix-adv	gold	silver	silver	silver	silver	silver
	CoNLL-12	gold	gold	gold	gold	gold	gold	gold	gold
	ARRAU _{all}	mix-mtl	mix-adv	gold	mix-mtl	mix-mtl	silver	mixed	silver
	ARRAU _{nominal}	mix-mtl	mix-mtl	gold	mix-mtl	mix-mtl	silver	mixed	silver
	ARRAU _{pronominal}	mix-mtl	mix-mtl	mixed	silver/mixed	mix-mtl	silver	mixed	silver
average-test	mix-mtl	gold	gold	mix-adv	silver	silver	silver	silver	

Table 4.34 The best training strategies between all possibilities: gold, silver, mixed, mixed with adversarial training, and mixed with multi-task learning. The window {AnaphS_{-i}, ..., AnaphS} is denoted by W_i. The "mix-mtl" in **bold** indicates the cases where multi-task learning results in the best performance between all training strategies.

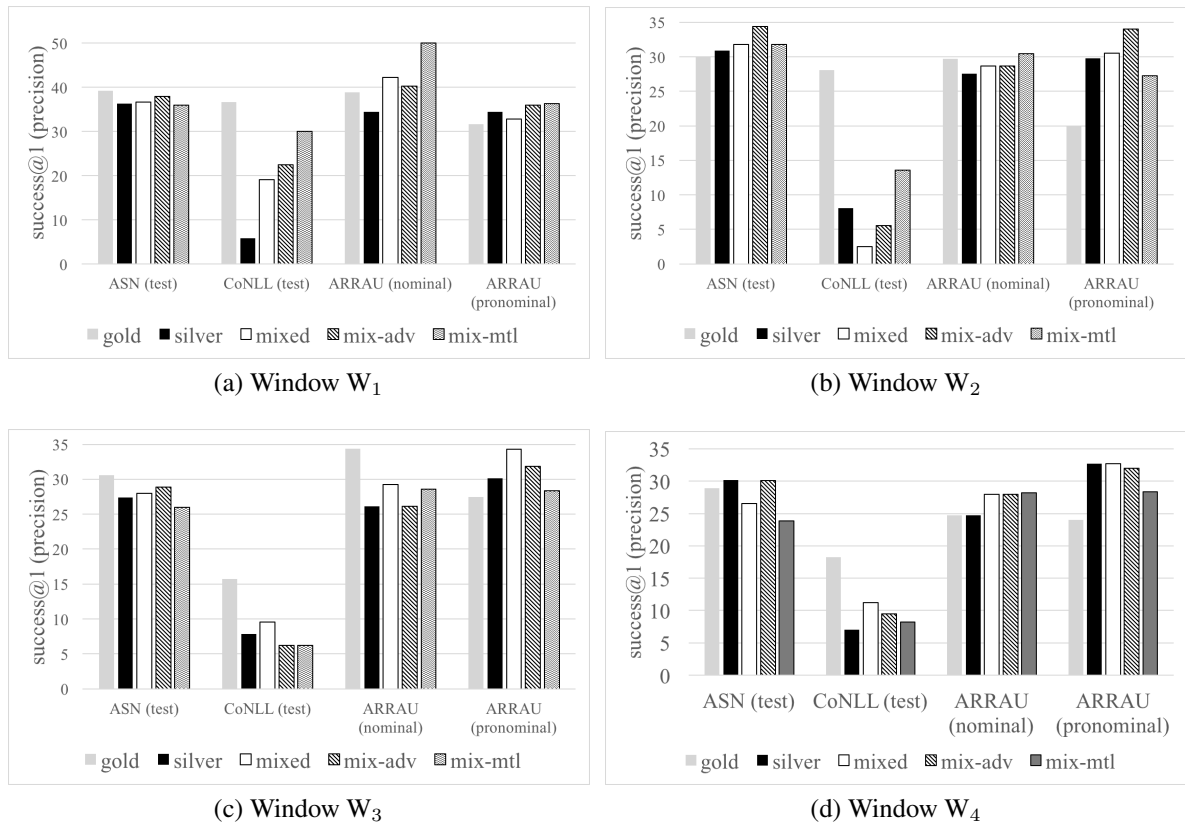


Fig. 4.7 Comparison of success@n scores across different training configurations and windows $W_i = \{\text{AnaphS}_{-i}, \dots, \text{AnaphS}\}$, $i \in \{1, 2, 3, 4\}$. We report the best MTL result for the ARRAU corpus between the results obtained with three scorers.

mixed + multi-task learning		s@1	s@2	s@3	s@4	token-F1	SentAcc
dev	ASN	48.46	66.92	73.08	79.62	66.81	81.92
	CoNLL-12	37.00	49.00	51.00	60.00	51.54	78.00
ASN		36.02	51.94	60.65	65.81	63.89	85.16
CoNLL-12		30.00	47.50	51.67	55.83	47.31	67.50
shell noun (ASN) scorer							
ARRAU _{all}		44.19	59.94	68.16	74.13	62.67	80.35
ARRAU _{nominal}		50.00	63.85	71.92	77.02	67.49	83.27
ARRAU _{pronominal}		36.37	54.50	62.92	70.41	55.99	76.26
average-test		39.32	55.54	63.06	68.64	59.47	78.51
test	event (CoNLL-12) scorer						
	ARRAU _{all}	15.87	28.32	35.65	38.32	42.33	60.60
ARRAU _{nominal}		16.35	29.42	37.21	40.67	41.88	55.10
ARRAU _{pronominal}		15.32	26.43	33.27	34.85	43.13	69.18
average-test		22.71	36.72	43.69	47.10	47.71	67.51
silver scorer							
ARRAU _{all}		26.38	39.05	50.70	60.48	56.43	78.67
ARRAU _{nominal}		23.75	36.44	48.75	60.67	60.97	83.75
ARRAU _{pronominal}		29.77	42.40	52.98	59.88	50.05	71.52
average-test		29.18	43.46	52.95	60.53	55.73	77.32

Table 4.35 MR-LSTM with **multi-task learning** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1 = \{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in **bold** indicate the cases when the multi-task learning resulted in a better performance than gold, silver, or mixed training (Table 4.12 on page 100), or adversarial training (Table 4.29 on page 118).

mixed + multi-task learning		s@1	s@2	s@3	s@4	token-F1	SentAcc	
dev	ASN	47.50	57.86	64.64	68.57	65.08	79.64	
	CoNLL-12	8.25	23.50	30.50	40.75	41.37	65.25	
	ASN	31.76	44.71	52.35	58.82	57.62	74.41	
	CoNLL-12	13.61	24.44	32.78	41.11	43.13	68.06	
shell noun (ASN) scorer								
test	ARRAU _{all}	29.39	40.20	45.31	51.43	49.12	64.08	
	ARRAU _{nominal}	30.44	41.78	47.44	53.44	50.51	63.11	
	ARRAU _{pronominal}	27.25	37.25	41.25	48.50	46.57	65.50	
	average-test	26.49	37.68	43.83	50.66	49.39	67.03	
	event (CoNLL-12) scorer							
	ARRAU _{all}	27.55	40.82	47.76	54.49	51.92	67.96	
	ARRAU _{nominal}	28.67	41.33	49.67	56.67	54.76	69.33	
	ARRAU _{pronominal}	26.25	39.75	44.25	51.75	47.84	66.50	
	average-test	25.57	38.21	45.36	52.57	51.06	69.25	
	silver scorer							
ARRAU _{all}	26.33	37.14	46.53	54.29	48.77	66.33		
ARRAU _{nominal}	25.33	37.22	46.78	55.44	50.50	65.78		
ARRAU _{pronominal}	27.25	36.25	46.00	52.75	46.31	67.50		
average-test	24.86	35.95	44.89	52.48	49.27	68.41		

Table 4.36 MR-LSTM with **multi-task learning** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1 = \{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in **bold** indicate the cases when the multi-task learning resulted in a better performance than gold, silver, or mixed training (Table 4.16 on page 104), or adversarial training (Table 4.30 on page 118).

mixed + multi-task learning		s@1	s@2	s@3	s@4	token-F1	SentAcc	
dev	ASN	40.87	53.37	59.44	64.44	59.16	70.63	
	CoNLL-12	6.00	16.00	19.25	26.50	36.44	65.25	
	ASN	26.00	39.43	48.29	52.86	51.42	62.86	
	CoNLL-12	6.19	17.38	20.71	28.57	39.58	65.60	
shell nouns (ASN) scorer								
test	ARRAU _{all}	26.08	37.84	45.69	50.59	49.02	62.35	
	ARRAU _{nominal}	26.41	36.41	44.06	49.69	48.86	58.44	
	ARRAU _{pronominal}	25.33	39.33	47.83	52.00	49.19	68.67	
	average-test	22.00	34.08	41.32	46.74	47.62	63.58	
	event (CoNLL) scorer							
	ARRAU _{all}	26.86	38.63	46.86	52.55	50.56	64.12	
	ARRAU _{nominal}	25.78	37.97	46.88	53.44	50.73	61.56	
	ARRAU _{pronominal}	28.33	38.83	46.33	50.67	50.05	68.17	
	average-test	22.63	34.45	41.81	47.62	48.47	64.46	
	silver scorer							
ARRAU _{all}	27.65	39.80	46.67	52.94	51.00	66.47		
ARRAU _{nominal}	28.59	41.72	48.13	54.06	52.88	66.88		
ARRAU _{pronominal}	25.83	36.17	44.17	51.00	47.68	65.67		
average-test	22.85	34.90	41.59	47.89	48.51	65.49		

Table 4.37 MR-LSTM with **multi-task learning** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1 = \{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in **bold** indicate the cases when the multi-task learning resulted in a better performance than gold, silver, or mixed training (Table 4.20 on page 108), or adversarial training (Table 4.31 on page 119).

mixed + multi-task learning		s@1	s@2	s@3	s@4	token-F1	SentAcc	
dev	ASN	35.52	51.03	55.86	62.07	56.89	66.55	
	CoNLL-12	3.00	14.00	21.25	27.25	37.80	66.50	
	ASN	23.87	35.86	47.21	53.96	48.48	60.72	
	CoNLL-12	8.21	16.90	25.24	31.07	42.46	69.29	
shell nouns (ASN) scorer								
test	ARRAU _{all}	23.37	33.64	42.13	48.06	46.88	57.12	
	ARRAU _{nominal}	24.12	34.41	43.24	48.53	47.39	52.35	
	ARRAU _{pronominal}	25.33	35.50	43.00	49.83	48.27	67.33	
	average-test	20.98	31.26	40.16	46.29	46.70	61.36	
	event (CoNLL) scorer							
	ARRAU _{all}	23.94	33.83	43.45	50.32	49.24	61.64	
	ARRAU _{nominal}	25.88	36.76	45.29	52.06	52.20	61.18	
	ARRAU _{pronominal}	25.88	36.76	45.29	52.06	52.20	61.18	
	average-test	25.88	36.76	45.29	52.06	52.20	61.18	
	silver scorer							
ARRAU _{all}	28.92	39.11	45.34	52.40	51.04	63.99		
ARRAU _{nominal}	28.24	37.94	45.29	51.76	52.92	65.00		
ARRAU _{pronominal}	28.24	37.94	45.29	51.76	52.92	65.00		
average-test	23.52	33.73	41.18	48.04	48.58	64.63		

Table 4.38 MR-LSTM with **multi-task learning** results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1 = \{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in **bold** indicate the cases when the multi-task learning resulted in a better performance than gold, silver, or mixed training (Table 4.24 on page 112), or adversarial training (Table 4.32 on page 119).

4.8 Summary

In this chapter we have proposed a Mention-Ranking LSTM-siamese (MR-LSTM) neural model for the challenging task of unrestricted abstract anaphora resolution, i.e. finding a (typically) non-nominal antecedent of pronouns and noun phrases that refer to abstract objects like facts, events, actions, or situations in the preceding discourse.

We designed the MR-LSTM model on the basis of the assumption that we can learn what is the correct antecedent for a given abstract anaphor by learning characteristics of the relation that holds between the sentence with the abstract anaphor and its antecedent. Since there is a number of properties associated with abstract anaphora we opted for a neural model which captures these attributes in a data-driven fashion. To circumvent the problems that arise from the scarcity of labeled data, we automatically extracted antecedent-anaphoric sentence pairs from parsed corpora. In particular, we searched for constructions with embedded sentences, applied a simple transformation that replaces the embedded sentence with an abstract anaphor, and used the "cut-off" embedded sentence as the antecedent. To be able to find antecedents from at least few preceding sentences we sampled a distance between an extracted antecedent and the anaphor from a distribution calculated on the dataset of naturally occurring anaphors.

Our MR-LSTM model performs better than baselines in the majority of cases. The only exception is the CoNLL-12 corpus for which we repeatedly show that it covers different phenomena from what we encounter in other datasets. Our experiments suggest that the gold training data contains better representatives for some types of evaluation examples (CoNLL-12, the dev part of ASN, and the nominal part of the ARRAU corpus), while the silver data contains better representatives for other (the pronominal part of the ARRAU corpus). Mixing two sources of data (gold and silver) is the best strategy in some cases, but not uniformly across all evaluation configuration. The final conclusion is that training on the gold data works well if we have exactly the same type of anaphors at the evaluation time. However, since the ARRAU corpus covers variety of anaphor types our models benefit from the extracted data.

We attempted to improve the training on a mixture of the gold and silver data with adversarial training and multi-task learning. We suspected that adversarial training could filter artifacts of the silver data and that multi-task learning could enable our models to take different decisions for scoring candidates of different types of abstract anaphors. The results from these two additional training strategies differ across training and evaluation configurations. Therefore, making the mixed training effective remains a challenge. In the future work, we suggest to assess the benefit of adversarial training and multi-task model with a MR-LSTM model without the verb feature which negatively effected the performance

and with a better alternative to the CoNLL-12 dataset which proved to be unsuitable for studying abstract anaphora resolution.

Finally, this work is—to our knowledge—the first abstract anaphora resolution system that handles the unrestricted phenomenon in a realistic setting.

Chapter 5

Outlook and Conclusions

In this chapter, we summarize the contributions of this thesis and the insights gained. We then discuss the limitations and various potential future directions.

5.1 Contributions and Takeaways

The focus of this thesis is designing machine learning models for two tasks that could potentially assist models for sentiment inference: opinion role labeling and abstract anaphora resolution. At the beginning of the thesis, we raised a series of research questions centered on the problem of limited labeled data for labeling of opinion roles and on the automatic resolution of abstract anaphors. We revisit these questions, summarize how they have been addressed, and highlight our core contributions and insights.

Opinion Role Labeling (ORL). We investigated whether Multi-Task Learning (MTL) can overcome divergences in the annotation schemes of the opinion and semantic role labeling by adopting one of the recent successful SRL neural models and experimenting with different MTL frameworks. We empirically confirm the raised question. We show that MTL models achieve significant improvements with all evaluation measures, for both holders and targets, on both development and test sets, when evaluated with repeated 4-fold CV. Since nowadays it is common to report only one run of a 10-fold CV, we propose to instead repeat a 4-fold CV twice. The 4-fold CV provides equally large development and test sets unlike the 10-fold CV used by the prior work. We recommend evaluation with comparable development and test set sizes for future work because it enables more reliable evaluation.

We conducted a thorough HP tuning to compare different MTL frameworks. We show that that the simplest MTL model performs the best and that more sophisticated MTL models are mutually comparable.

To understand better what works and what is next for ORL, we carried out a thorough analysis. We summarize some of the important results as follows. First, we observed that our models struggle to capture long-range dependencies, the same problem current SRL models struggle with. Since we adapted the SRL model for ORL, the research contributions in SRL can now immediately be explored for ORL as well. This is not the case for the neural opinion analysis model of Katiyar and Cardie (2016). Second, we observed that our models have difficulty handling roles that occur with the corresponding opinion in complicated syntactic constructions. For this reason, we think a good line of future work should investigate whether training the ORL model jointly with a neural dependency parser helps the ORL model to cope with complex sentence constructions. The next challenge for our ORL model pose roles that require special inference skills to be detected. To handle such cases, it would be worth trying to train a model to recognize textual entailment together with the ORL model. Finally, we observe from the analysis that the biggest contribution of the MTL comes from having far fewer missing roles.

We analyzed what else besides enhancing ORL models might be worth improving for the task. We show that the evaluation scripts need to be extended such that predicting any entity from the coreference cluster is considered to be correct instead of only the entity closest to the corresponding opinion expression. Next, the evaluation scripts needs to be extended such that they handle discontinuous roles predicted by a system. Finally, we confirm the insight from Katiyar and Cardie (2016) that computational models predict reasonable roles that are not annotated in the MPQA corpus. Therefore, the missing roles have to be curated.

Abstract Anaphora Resolution (AAR). We asked ourselves how can we apply computational methods to resolve abstract anaphors automatically. To the best of our knowledge, we propose the first abstract anaphora resolution model that handles the unrestricted phenomenon in a realistic setting.

We cast AAR as a task of learning the characteristics of the relation that holds between the sentence with the anaphor (i.e anaphoric sentence) and the antecedent. We propose a neural model since it can learn the mentioned characteristics in a data-driven fashion and therefore does not force us to make certain assumptions that might be limiting. In particular, we present a Mention-Ranking LSTM-siamese neural network (MR-LSTM) for learning the mentioned characteristics.

To be able to train a neural model we have to circumvent the problems that arise from the limited labeled data. For this reason, we propose a method for extracting antecedent-anaphoric sentence pairs from parsed corpora by searching for a common construction which consists of a verb with an embedded sentence (complement or adverbial).

We evaluate our models in a realistic task setup in which models need to rank embedded sentences and verb phrases from the sentence with the anaphor as well as few preceding sentences. Therefore, we need to provide our models with a distance between a candidate for the antecedent and the anaphor. Since the extracted antecedent pairs do not have a distance from the anaphor, we propose a simple solution. We sample the distances from a distribution calculated from naturally occurring anaphors in the development set. Furthermore, motivated by insights from the literature in linguistics, we propose an additional feature: the verb that governs the head of the anaphor. We hypothesized that this verb may serve as a good indicator of anaphor’s preferences for the syntactic and semantic type of antecedent. However, the verb feature had a negative impact on the performance of the model.

We provide a thorough evaluation of our MR-LSTM models. We analyze the performance of the full MR-LSTM model with respect to the different sources of training data: (i) large-scale extracted (silver) data, (ii) small-scale gold data, (iii) a mixture of gold and silver data. We analyze the performance across different windows for extraction of candidates for the antecedent. We also provide results of our MR-LSTM models without the verb and distance feature. To gain an accurate understanding of capabilities of our models we provide results in supplementary evaluation measures.

We summarize some of the important insights as follows. Our MR-LSTM model performs better than baselines in the majority of cases. The only exception is the CoNLL-12 dataset for which we repeatedly show that it covers different phenomena from what we encounter in other datasets. Our experiments suggest that the gold training data contains better representatives for some types of evaluation examples (CoNLL-12, the dev part of ASN, and the nominal part of the ARRAU corpus), while the silver data contains better representatives for other (the pronominal part of the ARRAU corpus). Mixing two sources of data (gold and silver) is the best strategy in some cases, but not uniformly across all evaluation configurations. The final conclusion is that training on the gold data works well if we have exactly the same type of anaphors at the evaluation time. However, since the ARRAU corpus covers a variety of anaphor types our models benefit from the extracted silver data.

We attempted to improve the training on a mixture of the gold and silver data with adversarial training and multi-task learning. The results from these two additional training strategies differ across training and evaluation configurations. Therefore, making the mixed training effective remains a challenge.

5.2 Discussion

Although this thesis breaks new ground for opinion role labeling and abstract anaphora resolution from a computational perspective, there are still open problems in this field. Here we discuss some of the limitations of this work.

Opinion Role Labeling (ORL). We evaluated the benefit of MTL for the ORL model which is given a gold (oracle) opinion expression. It is still an open question whether improvements from MTL persist in the ORL without the oracle opinion expression. Therefore, we have to extend the model presented in Chapter 3 with a component that first predicts the opinion expression. We have replicated results of the deep bidirectional-LSTM model of Irsoy and Cardie (2014b) on MPQA 1.0, but there is a significant drop in performance when we evaluate their model on MPQA 2.0. We get the binary F1 score 71.17 for opinion extraction on MPQA 1.0 (reported is 71.72), whereas on MPQA 2.0 the same model achieves only 62.71. For this reason, a pipeline approach would most likely have a big error propagation. Therefore, the model presented in Chapter 3 has to be trained jointly with the component that extracts the opinion expression. However, since the input to the ORL model are embeddings of the predicate and its context, some propagation of error will probably still persist. Therefore we need to design an input to the ORL model that signals the predicate in the sentence, but less explicitly as in the current model. We could use some suggestions that are already proposed for SRL (He et al., 2017; Peng et al., 2018).

Abstract Anaphora Resolution (AAR). We note that our extraction method covers a much wider range of anaphoric types compared to the method proposed by Kolhatkar et al. (2013b). However, it also has two limitations: the anaphors are always pronominal expressions and antecedents are uniformly of type S. It is likely that the former effects the performance on the nominal part of the ARRAU corpus.

We have measured the distance between a candidate for the antecedent and the anaphor in number of sentences. However, there are different ways we could define the distance. For example, in number of tokens or the number of edges between nodes in some discourse structure. Moreover, we use the simplest solution to sampling distance for extracted (silver) antecedents. However, a distance is dependent on the type of anaphor and we should think about more sophisticated methods for distance sampling.

We show that the verb feature has negative impact on our models. Since this feature seem to not work as expected in practice, it may be concluded that a good line of work must be exploring new ways of integrating this feature in our models.

We experimented with blending the gold and silver data, but we tried only one of many possible ways to blend these two data types. We alternate batches from the gold and silver data from the beginning of the training of our model. That is, every even iteration we sample a batch of gold data and every odd iteration we sample a batch of silver data. Recently, Shnarch et al. (2018) proposed a methodology to blend high quality but scarce labeled data with noisy but abundant weak labeled data during the training of neural networks. They experiment with different blending factors as well as with pre-training the architecture with weak labeled data. We could follow some of their suggestions. Moreover, we assessed the impact of the mixed training, adversarial training, and multi-task learning only with the full model with the verb feature and the CoNLL-12 dataset. Since the feature and the dataset proved to be ineffective, they might also affect proper utilization of additional modeling proposals such as the mixed training, adversarial training, and multi-task learning. We could re-assess the value of these training strategies using the best version of our MR-LSTM model. However, there is no alternative to the CoNLL-12 dataset yet.

Finally, although we give a through evaluation of our MR-LSTM models we are still focused solely on the in-domain evaluation. Once we have a robust in-domain model, we must test how does it generalize to other domains.

5.3 Direction for Future Work

In this section, we discuss a few promising directions of future research.

Pre-trained language models. The most recent advances in NLP show that the pre-trained language models can be used to achieve state-of-the-art results on a wide range of NLP tasks (Peters et al., 2018a; Howard and Ruder, 2018; Devlin et al., 2018). We expect they could benefit both opinion role labeling and abstract anaphora resolution.

We have already noted that the ORL model could be improved if we train it together with a neural dependency parser and a model that recognizes textual entailment. Peters et al. (2018b) show that deep bi-directional Language Models (LMs) learn a hierarchical relations of both words and spans. The lower LM layers specialize in local syntactic relationships, allowing the higher layers to model longer range relationships such as coreference. Their results show that LMs act as a general purpose feature extractor for natural language. Therefore, it might be sufficient to use the representation produced from such a language model to handle the cases that the ORL model struggles with.

Our MR-LSTM models for abstract anaphora resolution have to learn the individual representations of the anaphoric sentence and a candidate for the antecedent from scratch.

However, since these new pre-trained language models are available, it would be worthwhile to try their representations and supervise model for abstract anaphora resolution only for layers that come after individual representations are produced.

Cross-lingual opinion analysis. The neural approaches are suitable for cross-lingual studies since in theory the only necessary change to the model is to replace monolingual word embeddings with bi-lingual word embeddings (Ruder et al., 2017). Moreover, the most recent pre-trained language model BERT¹ (Devlin et al., 2018) is available for 104 languages. Therefore, we could pre-training the entire model with hierarchical representations in any of the 104 languages and from there we need only some labeled data.

Utilizing extracted AAR data. Nie et al. (2017) use discourse markers like *but* and *because* as a natural signal to learn the meaning of sentences they connect. We could follow their approach and use the verb that embeds the sentence (or the conjunction for sentential adjuncts) to learn the meaning of the embedding sentence and the embedded sentence.

Sentiment inference in discourse. The recent advances in machine learning enable us to design neural versions of sentiment inference models outlined in Chapter 2 (Section 2.1). Neural alternatives could be combined with our neural models using a joint training objective.

Battaglia et al. (2018)² present a general framework for entity- and relation-based reasoning (which they term graph networks) for unifying and extending existing methods which operate on graphs, and describe key design principles for building powerful architectures using graph networks as building blocks. We could investigate their framework to design a neural alternative of the graph-based sentiment inference model (Deng and Wiebe, 2014).

Minervini et al. (2017) proposed a new training algorithm for knowledge base completion. They first define a set of constraints in the form of function-free first-order logic clauses. From these clauses they then derive an inconsistency loss that measures the extent to which constraints are violated. The learning architecture is composed of two models, an adversary and a discriminator, having two competing goals. The adversary tries to find a set of adversarial input representations for which, according to the discriminator (a link prediction model), the constraints do not hold. Such a set is found by maximizing the inconsistency loss. The discriminator, on the other hand, uses the inconsistency loss on the adversarial input representations for regularizing its training process. This training algorithm may be used to achieve the same goals as the ILP training objective in Deng et al. (2014).

¹<https://github.com/google-research/bert>

²https://github.com/deepmind/graph_nets

Finally, even probabilistic soft logic has been successfully integrated in a neural architecture (Aditya et al., 2018). Therefore, it is likely that the probabilistic soft logic approach of Deng and Wiebe (2015a) may be adapted to a neural model.

Appendix A

MPQA Pre-processing

The MPQA corpus is challenging not only because it captures a variety of ORL cases as we have illustrated in Chapter 3 (Section 3.1), but also because it is hard to process it in a way that it can be presented to a neural sequence labeling model. Code or sufficient description how the corpus was processed is not available from the prior work (Yang and Cardie, 2013; Katiyar and Cardie, 2016).

The first difficulty is that we are designing a model that labels at the token-level, but annotation spans are given in bytes. Thus, we used Stanford CoreNLP (Manning et al., 2014) which tokenizes text and gives the byte span of every token.¹ However, due to the absence of punctuation for transcripts of spoken conversations the sentence splitter treats a whole document as one sentence. Therefore, for sentences longer than 150 tokens, we take 15 tokens preceding the opinion expression, the expression itself, and 15 tokens after as proxy for a sentence that we present to the model.

Opinion expressions of interest are annotated in MPQA as *direct subjectives* (DSEs). We discard implicit DSEs which frequently point to the attitude which covers the whole sentence and reflects the attitude of the author of the document as in Example (27). These DSEs are not useful for the task we are looking into. Although such DSEs should be marked with the `implicit` attribute, sometimes they are not. Some of such cases we capture by demanding that a DSE is longer than one byte and that the author is not the only holder. There are few DSEs for which byte spans did not match with any sentence and we discard those as well.

(27) But there can not be any real [**talk**]_O of success until the broad strategy against terrorism begins to bear fruit.

For every document, we collected from the corresponding annotation file: identifiers and byte spans of all holders marked with `GATE_agent` (\mathcal{H}), attitudes marked with `GATE_`-

¹We used python wrapper: https://github.com/brendano/stanford_corenlp_pywrapper

	# DSEs (including ignored)	# implicit DSEs (ignored)	# inferred DSEs (ignored)	# filtered DSEs	# somewhat uncertain DSEs (filtered)
TRAIN (avg)	3723.5	481	101.25	3141.25	133
TEST (avg)	1229.5	159	33.75	1036.75	44
DEV	1263	168	40	1055	43
	# very uncertain DSEs (filtered)	# DSEs w/o Hs (filtered)	# DSEs w/o roles (filtered)	# DSEs w/o Ts (filtered)	# insubstantial DSEs (filtered)
TRAIN (avg)	40.5	171.25	66.75	413.75	528.75
TEST (avg)	13.5	56.75	22.25	136.25	174.25
DEV	15	56	22	146	180
	# Hs of filtered DSEs	# Ts of filtered DSEs	# somewhat uncertain Hs	# somewhat uncertain Ts	# overlapping entites
TRAIN (avg)	2903.25	19528.5	17.25	27.75	961
TEST (avg)	957.75	6424.5	5.75	9.25	318
DEV	977	6073	5	6	305
	sentiment neg	sentiment pos	arguing pos	other attitude	intention pos
TRAIN (avg)	946	817.5	438.25	381	238.75
TEST (avg)	314	270.5	143.75	126	79.25
DEV	299	300	131	126	66
	arguing neg	agree pos	speculation	agree neg	intention neg
TRAIN (avg)	110.5	99.25	64.5	68.25	19.5
TEST (avg)	35.5	32.75	20.5	22.75	6.5
DEV	48	40	25	31	5

Table A.1 Statistics of the ORL (MPQA) data for 4-fold CV.

attitude, and targets marked with `GATE_target`. Holders and targets can be marked multiple times with the same id, but with different byte spans. If the attribute `nested-source` of a DSE or the `target-link` attribute of its attitude point to identifiers of such holders and targets, we pick the byte spans which are closest to the DSE. In many cases the `nested-source` attribute of a DSE pointed to a holder which is not marked in the annotation file ($\notin \mathcal{H}$). We tried to fix the `nested-source` attribute by doing the following transformations: (1) adding 'w' to the beginning (e.g. `nhs` \mapsto `w, nhs`), (2) removing 'w' from the beginning (e.g. `w, ip` \mapsto `ip`), (3) removing duplicates (e.g. `w, mug, mug` \mapsto `w, mug`). Although these transformations helped a lot, they are a few holders and targets we could not trace.

In some cases, as in Example (28), an opinion expression and its opinion roles overlap. In average, we discard 74.7 such holders and 16.2 targets, because we train the output CRF to predict only one label by token. Notice that the prior work (Katiyar and Cardie, 2016) had to do the same.

(28) Mugabe said [Zimbabwe]_T needed their continued support against what he called [hostile [international]_H attention]_O.

	# DSEs (including ignored)	# implicit DSEs (ignored)	# inferred DSEs (ignored)	# filtered DSEs	# somewhat uncertain DSEs (filtered)
TRAIN (avg)	4173.3	537.3	119.7	3516.3	137.7
TEST (avg)	457.8	43.9	29.9	349.3	15.2
DEV	1579	211	42	1326	67
	# very uncertain DSEs (filtered)	# DSEs w/o Hs (filtered)	# DSEs w/o roles (filtered)	# DSEs w/o Ts (filtered)	# insubstantial DSEs (filtered)
TRAIN (avg)	47.7	187.2	77.4	459.9	567.9
TEST (avg)	7.3	19.3	11.8	82.6	150.5
DEV	16	76	25	185	252
	# Hs of filtered DSEs	# Ts of filtered DSEs	# somewhat uncertain Hs	# somewhat uncertain Ts	# overlapping entites
TRAIN (avg)	3251.7	21664.8	17.1	27.9	1064.7
TEST (avg)	957.4	1700	19.4	37.8	84.9
DEV	1225	7978	9	12	401
	sentiment neg	sentiment pos	arguing pos	other attitude	intention pos
TRAIN (avg)	1008.9	949.5	471.6	440.1	266.4
TEST (avg)	107.8	89.4	50.7	40.2	25.6
DEV	438	333	189	144	88
	arguing neg	agree pos	speculation	agree neg	intention neg
TRAIN (avg)	133.2	115.2	80.1	74.7	16.2
TEST (avg)	14.1	11.1	8.6	6.5	1.428571429
DEV	46	44	21	39	13

Table A.2 Statistics of the ORL (MPQA) data for 10-fold CV.

We discard inferred attitudes, as labeling of their targets is considered to be another task (sentiment inference).

Further, a DSE can have multiple attitudes and each attitude can point to different targets. Again, because the model can predict only one label by token, we have to pick one attitude and non-overlapping targets. We chose attitudes according to the following priorities: sentiment, intention, agreement, arguing, other-attitude, speculation.

We kept DSEs with the `insubstantial` attribute which are either not significant (29) or not not real within the discourse (30). Our models should demonstrate the ability of properly labeling roles of insubstantial DSEs. However, note that when FGOA is used for opinion-oriented summarization or QA, opinion roles of insubstantial opinions should not be labeled. A full FGOA system should additionally predict whether an opinion is substantial within the discourse, before labeling its opinion roles.

(29) [...] it completely supports the [U.S.]_H [**stance**]_O [...].

(30) [...] Antonio Martino, meanwhile, said [...] that his country would not support an attack on Iraq without "proven proof" that [Baghdad]_H is [**supporting**]_O [al Qaeda]_T.

Finally, DSE, holder and target annotations allow an attribute that indicates whether an annotator was uncertain with possible values: somewhat- and very-uncertain. We did not

discard those believing that they would have been discarded by the corpus creators if they are really incorrect.

For reproducibility we report detailed data statistics in Tables A.1 and A.2: average number (calculated over folds) of all extracted DSEs, implicit DSEs, inferred DSEs, DSEs used in experiments (not implicit or inferred), somewhat uncertain DSEs used in experiments, very uncertain DSEs used in experiments, insubstantial DSEs used in experiments, the average number (calculated over folds) of DSEs used in experiments without a holder, without a target, without the `attitude-link` attribute, without both roles, the average number (calculated over folds) of holders, somewhat uncertain holders, very uncertain holders, targets, somewhat uncertain targets and very uncertain targets, the average number (calculated over folds) of different attitude types used in the experiments.

Examples how to easily use our MPQA pre-processing scripts can be found at <https://github.com/amarasovic/naacl-mpqa-srl4orl/blob/master/mpqa2-pytools.ipynb>.

Appendix B

Comparison between MR-LSTM and the Prior Work in Shell Noun Resolution

The variant of our MR-LSTM model proposed in Marasović et al. (2017) differs from the MR-LSTM model described in this thesis. We will refer to the MR-LSTM model in Marasović et al. (2017) as MR-LSTM 1.0. The input to the LSTM layers in the MR-LSTM 1.0 model are not only the word embeddings. Each word is represented with a vector constructed by concatenating the pre-trained word embedding, the embedding of the context of the anaphor (i.e. the average of embeddings of the anaphoric phrase, the previous, and the next word), the pre-trained word embedding of the head of the anaphoric phrase, and, finally, an embedding of the constituent tag label of the candidate, or the *S* constituent tag if the word is in the anaphoric sentence. The embedding of the context of the anaphor and the embedding of the head of the anaphoric phrase obtained mixed results in evaluation on the ARRAU corpus. Therefore we did not utilize them in this thesis. The constituent tag label embedding should not be beneficial since in the thesis we restrict the set of candidates to constituents in {S, VP, ROOT, SBAR, SBARQ}.

Table B.1 provides the published results of MR-LSTM 1.0 on the ASN corpus using default HPs (Marasović et al., 2017). Column 2 states which model produced the results: KZH13 refers to the best reported results in (Kolhatkar et al., 2013b) and TAG_{BL} is the baseline that randomly picks a candidate with a constituent tag label in {S, VP, ROOT, SBAR, SBARQ}. The candidates are all syntactic constituents of only the sentence that contains the antecedent. Thus, this is a restricted task setup compared to the evaluation setup presented in this thesis.

In terms of *success*@1 score (i.e. precision), MR-LSTM 1.0 outperforms both KZH13's results and TAG_{BL} without even necessitating HP tuning. For the outlier *reason* we tuned HPs for different variants of the architecture: the full architecture, without embedding of

the context of the anaphor (ctx), of the anaphor (aa), of both constituent tag embedding and shortcut (tag,cut), dropping only the shortcut (cut), using only word embeddings as input (ctx,aa,tag,cut), without the first (ffl1) and second (ffl2) layer. From Table B.2 we see that MR-LSTM with HPs obtains results well beyond KZH13.

		s @ 1	s @ 2	s @ 3	s @ 4
fact (train: 43809, test: 472)	MR-LSTM	83.47	85.38	86.44	87.08
	KZH13	70.00	86.00	92.00	95.00
	TAG _{BL}	46.99	-	-	-
reason (train: 4529, test: 442)	MR-LSTM 1.0	71.27	77.38	80.09	80.54
	+ tuning	87.78	91.63	93.44	93.89
	KZH13	72.00	86.90	90.00	94.00
issue (train: 2664, test: 303)	TAG _{BL}	42.40	-	-	-
	MR-LSTM 1.0	88.12	91.09	93.07	93.40
	KZH13	47.00	61.00	72.00	81.00
decision (train: 42289, test: 389)	TAG _{BL}	44.92	-	-	-
	MR-LSTM 1.0	76.09	85.86	91.00	93.06
	KZH13	35.00	53.00	67.00	76.00
question (train: 9327, test: 440)	TAG _{BL}	45.55	-	-	-
	MR-LSTM 1.0	89.77	94.09	95.00	95.68
	KZH13	70.00	83.00	88.00	91.00
possibility (train: 11874, test: 277)	TAG _{BL}	42.02	-	-	-
	MR-LSTM 1.0	93.14	94.58	95.31	95.67
	KZH13	56.00	76.00	87.00	92.00
	TAG _{BL}	48.66	-	-	-

Table B.1 Shell noun resolution results.

						reason			
ctx	aa	tag	cut	ffl1	ffl2	s@1	s@2	s@3	s@4
✓	✓	✓	✓	✓	✓	87.78	91.63	93.44	93.89
✗	✓	✓	✓	✓	✓	85.97	87.56	89.14	89.82
✓	✗	✓	✓	✓	✓	86.65	88.91	91.18	91.40
✓	✓	✗	✗	✓	✓	68.10	80.32	85.29	89.37
✓	✓	✓	✗	✓	✓	85.52	88.24	89.59	90.05
✗	✗	✗	✗	✓	✓	66.97	80.54	85.75	88.24
✓	✓	✓	✓	✗	✓	87.56	91.63	92.76	94.12
✓	✓	✓	✓	✓	✗	85.97	88.69	89.14	90.05

Table B.2 Architecture ablation for *reason*.

Appendix C

Quality Evaluation of the VC-SS-Extract Method

Two experienced annotators independently assessed the quality of 10 randomly chosen instances per possible value of the head of the SBAR clause. Each instance was marked as either being sound (✓) or unusable (✗). Cases with marginal acceptability due to unnatural sounding anaphora expression or position in the sentence could be marked with a special sign (⚡).

This annotation scheme was applied in two phases. In the first phase, we evaluated the data which is extracted using heads of the SBAR clause proposed in Marasović et al. (2017). In the second phase, we evaluated data we extracted after refining the VC-SS-Extract patterns. Tables C.1 and C.2 give an overview of the replacement types in the two phases. We list the embedding sentence type, the head of SBAR, and point to the tables that display the annotation results. Table C.2 further displays which extraction types were used for our refined data extraction, based on the annotation results (column 4).

We refine the VC-SS-Extract pattern as follows:

- We eliminate the cases with the wh-adverb (WHADV), wh-noun (WHNP), or wh-propositional (WHPP) subordinate SBAR clause.
- We require that VPs have exactly two children if *that* is the head of the SBAR clause.
- We eliminate constructions such as: *Under the direction of its new chairman, Francisco Luzon, Spain's seventh largest bank is undergoing a tough restructuring* $[[that]_{WHNP} [analysts [say [[may\ be\ the\ first\ step\ toward\ the\ bank's\ privatization]_S]_{SBAR2}]_{VP}]_{S1}]_{SBAR1}$.

embedding sentence type	head of SBAR	table number
argument	zero	Table C.3
argument	\emptyset	Table C.4
argument	that	Table C.5
adjunct	as	Table C.7
adjunct	if	Table C.6
...		

Table C.1 Tables with manual evaluations for (up to) 10 randomly drawn examples for different embedding types (arguments and adjuncts) in Trial 1.

embedding sentence type	head of SBAR	table number	used
argument	zero	Table C.8	YES
argument	none	Table C.9	NO
argument interrog.	whether	Table C.10	YES
adjunct causal	because	Table C.11	YES
adjunct causal	since-prp	Table C.12	YES
adjunct causal	as-prp	Table C.13	YES
adjunct conditional	if-adv	Table C.14	YES
adjunct temporal	after	Table C.15	YES
adjunct temporal	since-tmp	Table C.16	YES
adjunct temporal	until	Table C.17	NO
adjunct temporal	while	Table C.18	NO
adjunct temporal	as-tmp	Table C.19	NO

Table C.2 Tables with manual evaluations for (up to) 10 randomly drawn examples for different embedding types (arguments and adjuncts) in Trial 2.

- We allow conjunctions *since*, *as*, and *if* to be the head of the SBAR clause only if the special PRP, TMP, and ADV parse attributes are available.
- To avoid extractions from sentences such as *But the test may prove to be more sensitive in determining whether a tumor has spread or returned following treatment, Dr. Wilson said.* we filter instances extracted using the VC-SS-extract pattern with the SBAR head *whether* and the coordination indicated by *or*. With *whether* we cannot settle the case (one or the other holds true). Such examples are mostly unnatural and cannot serve as antecedent.

SBAR head = zero						
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1	#2
that	first city , which recently purchased three small texas banking concerns , said it would use the proceeds to pursue additional expansion opportunities in the southwest and elsewhere .	first city , which recently purchased three small texas banking concerns , said <u>that</u> .		<ol style="list-style-type: none"> 1. it would use the proceeds to pursue additional expansion opportunities in the southwest and elsewhere 2. would use the proceeds to pursue additional expansion opportunities in the southwest and elsewhere 3. said it would use the proceeds to pursue additional expansion opportunities in the southwest and elsewhere 	✓	✓
this	although british air is waiting to see what the buy-out group comes up with , mr. stevens said a revised transaction with less debt leverage is likely to be more attractive to banks .	although british air is waiting to see what the buy-out group comes up with , mr. stevens said <u>this</u> .		<ol style="list-style-type: none"> 1. a revised transaction with less debt leverage is likely to be more attractive to banks 2. said a revised transaction with less debt leverage is likely to be more attractive to banks 	✓	✓
that	she said her employer ca n't afford the rate increases , and she fears she wo n't find another job with a benefit plan covering her ailment .	she said her employer ca n't afford the rate increases , and she fears <u>that</u> .		<ol style="list-style-type: none"> 1. she wo n't find another job with a benefit plan covering her ailment 2. fears she wo n't find another job with a benefit plan covering her ailment 3. wo n't find another job with a benefit plan covering her ailment 	✓✱	✓✱
that	justin 's attorney , charles e. baxley , said justin would ask an appeals court to set aside the order temporarily , pending an expedited appeal .	justin 's attorney , charles e. baxley , said <u>that</u> .		<ol style="list-style-type: none"> 1. justin would ask an appeals court to set aside the order temporarily , pending an expedited appeal 2. said justin would ask an appeals court to set aside the order temporarily , pending an expedited appeal 3. would ask an appeals court to set aside the order temporarily , pending an expedited appeal 	✓	✓
that	james kochan , chief fixed-income strategist at merrill lynch , is touting shorter-term securities , which he says should benefit more quickly than longer-term bonds as interest rates fall .	james kochan , chief fixed-income strategist at merrill lynch , is touting shorter-term securities , which he says <u>that</u> .		<ol style="list-style-type: none"> 1. should benefit more quickly than longer-term bonds as interest rates fall 2. says should benefit more quickly than longer-term bonds as interest rates fall 	✗	✗ ¹

SBAR head = zero						
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1	#2
this	he said the fourth quarter will be "challenging," and maintained his conservative forecast that 1990 "won't be a barn burner."	he said <u>this</u> , " and maintained his conservative forecast that 1990 " won't be a barn burner . "		<ol style="list-style-type: none"> 1. the fourth quarter will be "challenging 2. said the fourth quarter will be "challenging 	✓	✓
that	but " their present financial condition means i 'd have a hard time convincing the vice president in charge of purchasing . "	but " their present financial condition means <u>that</u> . "		<ol style="list-style-type: none"> 1. i 'd have a hard time convincing the vice president in charge of purchasing 2. means i 'd have a hard time convincing the vice president in charge of purchasing 3. 'd have a hard time convincing the vice president in charge of purchasing 	✗	✗
this	moody 's investors service inc. said it reduced its rating on \$ 281 million of senior and subordinated debt of this thrift holding company , to c from ca , saying it believes bondholders will recover only " negligible principal . "	moody 's investors service inc. said <u>this</u> saying it believes bondholders will recover only " negligible principal . "		<ol style="list-style-type: none"> 1. it reduced its rating on \$ 281 million of senior and subordinated debt of this thrift holding company , to c from ca , 2. reduced its rating on \$ 281 million of senior and subordinated debt of this thrift holding company , to c from ca , 	✓	✓
that	one new investment style called " asset allocation " shifts portfolio weightings between stocks , bonds and cash when computer models say one is more attractive .	one new investment style called " asset allocation " shifts portfolio weightings between stocks , bonds and cash when computer models say <u>that</u> .		<ol style="list-style-type: none"> 1. one is more attractive 2. say one is more attractive 3. is more attractive 	✓✚	✗
this	ms. bryant , the head of the state securities group , said drexel has done a better job of settling with the states than e.f. hutton did after its guilty plea to a massive check-kiting scheme several years ago .	ms. bryant , the head of the state securities group , said <u>this</u> .		<ol style="list-style-type: none"> 1. drexel has done a better job of settling with the states than e.f. hutton did after its guilty plea to a massive check-kiting scheme several years ago 2. said drexel has done a better job of settling with the states than e.f. hutton did after its guilty plea to a massive check-kiting scheme several years ago 3. has done a better job of settling with the states than e.f. hutton did after its guilty plea to a massive check-kiting scheme several years ago 	✓	✓

Table C.3 The first trial of the manual evaluation of the data extracted with the VC-SS-Extract method and 0 as the head of the SBAR clause.

		SBAR head = NONE			
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1 #2
that	on a video screen , riders can see 30 different " rides , " including urban , mountain and desert scenes , and check how many calories are burned a minute .	on a video screen , riders can see 30 different " rides , " including urban , mountain and desert scenes , and check how many calories <u>that</u> .		1. are burned a minute	X X
this	late yesterday afternoon , ddb needham executives were scrambling to figure out what to do about a new business presentation that had been scheduled for today , a spokesman said .	late yesterday afternoon , ddb needham executives were scrambling to figure out what <u>this</u> is true , a spokesman said .		1. to do about a new business presentation that had been scheduled for today 2. do about a new business presentation that had been scheduled for today, (2) what to do about a new business presentation that had been scheduled for today	X X
this	" the essence of being a trial lawyer is understanding how people of diverse backgrounds react to you and your presentation , " says barry ostrager of simpson thacher & bartlett , who recently won a huge case on behalf of insurers against shell oil co .	" the essence of being a trial lawyer is understanding how <u>this</u> , " says barry ostrager of simpson thacher & bartlett , who recently won a huge case on behalf of insurers against shell oil co .		1. people of diverse backgrounds react to you and your presentation 2. how people of diverse backgrounds react to you and your presentation	X X
this	it 's a trade secret how many were plastic , and most writers still do n't know what they 're using .	it 's a trade secret how many were plastic , and most writers still do n't know what <u>this</u> is true .		1. they 're using 2. what they 're using 3. they 're using	X X
this	alusuisse , of new york , declined to say how much it expects to get for the unit ; the company has hired first boston corp. to help identify bidders .	alusuisse , of new york , declined to say how much <u>this</u> is true ; the company has hired first boston corp. to help identify bidders .		1. it expects to get for the unit 2. expects to get for the unit	X X

this	" it 's too early to tell " what happens after that , he says .	" it 's too early to tell " what <u>this</u> is true , he says .	1. happens after that 2. what happens after that	X X
that	it is the right-wing guerrillas who are aligned with the drug traffickers , not the left wing .	it is the right-wing guerrillas who <u>that</u> , not the left wing .	1. are aligned with the drug traffickers 2. who are aligned with the drug traffickers	X X
this	from his new , million-dollar-a-year perch on wall street as a managing director of wertheim schroder & co. , mr. coelho reports that many of his former colleagues have contacted him to find out how they , too , can pursue investment banking careers .	from his new , million-dollar-a-year perch on wall street as a managing director of wertheim schroder & co. , mr. coelho reports that many of his former colleagues have contacted him to find out how <u>this</u> .	1. they , too , can pursue investment banking careers 2. how they , too , can pursue investment banking careers	X X
this	" what i thought i was saying is that the market is troubled but still viable and , appropriately enough , quite quality-conscious , which is not at all bad , " he says .	" what i thought i was saying is that the market is troubled but still viable and , appropriately enough , quite quality-conscious , which <u>this</u> is true , " he says .	1. is not at all bad 2. which is not at all bad	X X
this	this will " require us to define – and redefine – what is ' necessary ' or ' appropriate ' care .	this will " require us to define – and redefine – what <u>this</u> is true .	1. is ' necessary ' or ' appropriate ' care 2. what is ' necessary ' or ' appropriate ' care	X X

Table C.4 The first trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *NONE* as the head of the SBAR clause.

SBAR head = that					
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1 #2
that	for the broader market , the greatest significance of the vitro-anchor deal may be that it was put together late friday night – after the market rout – and involves a \$ 155 million temporary “ bridge ” loan from donaldson , lufkin & jenrette securities and a \$ 139 million loan from security pacific national bank .	for the broader market , the greatest significance of the vitro-anchor deal may be that .		<ol style="list-style-type: none"> 1. it was put together late friday night – after the market rout – and involves a \$ 155 million temporary “ bridge ” loan from donaldson , lufkin & jenrette securities and a \$ 139 million loan from security pacific national bank, 2. was put together late friday night – after the market rout – and involves a \$ 155 million temporary “ bridge ” loan from donaldson , lufkin & jenrette securities and a \$ 139 million loan from security pacific national bank, 	✓ ✓
this	in another takeover battle , a spokesman for mccaw cellular communications said yesterday that mccaw has been advised by three commercial banks that they remain “ highly confident ” they can arrange \$ 4.5 billion of bank loans for mccaw ’s tender offer for about 45 % of lin broadcasting , “ notwithstanding recent events . ”	in another takeover battle , a spokesman for mccaw cellular communications said yesterday that mccaw has been advised by three commercial banks this . ”		<ol style="list-style-type: none"> 1. they remain “ highly confident ” they can arrange \$ 4.5 billion of bank loans for mccaw ’s tender offer for about 45 % of lin broadcasting , “ notwithstanding recent events, 2. remain “ highly confident ” they can arrange \$ 4.5 billion of bank loans for mccaw ’s tender offer for about 45 % of lin broadcasting , “ notwithstanding recent events, 	✗ ✗
that	in another takeover battle , a spokesman for mccaw cellular communications said yesterday that mccaw has been advised by three commercial banks that they remain “ highly confident ” they can arrange \$ 4.5 billion of bank loans for mccaw ’s tender offer for about 45 % of lin broadcasting , “ notwithstanding recent events . ”	in another takeover battle , a spokesman for mccaw cellular communications said yesterday that . ” (0) mccaw has been advised by three commercial banks that they remain “ highly confident ” they can arrange \$ 4.5 billion of bank loans for mccaw ’s tender offer for about 45 % of lin broadcasting , “ notwithstanding recent events,		<ol style="list-style-type: none"> 1. has been advised by three commercial banks that they remain “ highly confident ” they can arrange \$ 4.5 billion of bank loans for mccaw ’s tender offer for about 45 % of lin broadcasting , “ notwithstanding recent events, 2. been advised by three commercial banks that they remain “ highly confident ” they can arrange \$ 4.5 billion of bank loans for mccaw ’s tender offer for about 45 % of lin broadcasting , “ notwithstanding recent events, 	✓ ✓✚

that	but investors should keep in mind , before paying too much , that the average annual return for stock holdings , long-term , is 9 % to 10 % a year ; a return of 15 % is considered praiseworthy .	but investors should keep in mind , before paying too much , that ; a return of 15 % is considered praiseworthy .	1. the average annual return for stock holdings , long-term , is 9 % to 10 % a year,	✓	✓✚
this	the ruling stems from a 1984 suit filed by shareholders of apple computer inc. , claiming that company officials misled investors about the expected success of the lisa computer , introduced in 1983 .	the ruling stems from a 1984 suit filed by shareholders of apple computer inc. , claiming this .	1. company officials misled investors about the expected success of the lisa computer , introduced in 1983,	✓	✓
that	lawyers are worried about the ruling 's implication in other shareholder suits but pointed out that the court stressed that the ruling should be regarded as very specific to the apple case .	lawyers are worried about the ruling 's implication in other shareholder suits but pointed out that .	1. the court stressed that the ruling should be regarded as very specific to the apple case, 2. stressed that the ruling should be regarded as very specific to the apple case,	✗	✗
that	“ the court was careful to say that the adverse information appeared in the very same articles and received the same attention as the company 's statements , ” said patrick grannon , a los angeles lawyer at the firm of greenfield & chimicles , which was n't involved in the case .	“ the court was careful to say that , ” said patrick grannon , a los angeles lawyer at the firm of greenfield & chimicles , which was n't involved in the case .	1. the adverse information appeared in the very same articles and received the same attention as the company 's statements,	✓	✓
this	mount lebanon high school , near pittsburgh , sought \$ 21 million in compensatory damages from grace , arguing that the asbestos , which can cause respiratory diseases and lung cancer , posed a risk to students .	mount lebanon high school , near pittsburgh , sought \$ 21 million in compensatory damages from grace , arguing this .	1. the asbestos , which can cause respiratory diseases and lung cancer , posed a risk to students,	✓	✓
this	rorer group inc. will report that third-quarter profit rose more than 15 % from a year earlier , though the gain is wholly due to asset sales , robert cawthorn , chairman , president and chief executive officer , said .	rorer group inc. will report this , robert cawthorn , chairman , president and chief executive officer , said .	1. third-quarter profit rose more than 15 % from a year earlier , though the gain is wholly due to asset sales,	✓	✓
that	an investor who may have placed a stop-loss order at \$ 90 under a stock that was trading at \$ 100 a share on the friday before the crash was stunned to discover that the order was filled at \$ 75 when the stock opened at that price on monday .	an investor who may have placed a stop-loss order at \$ 90 under a stock that was trading at \$ 100 a share on the friday before the crash was stunned to discover that .	1. the order was filled at \$ 75 when the stock opened at that price on monday, 2. was filled at \$ 75 when the stock opened at that price on monday,	✓	✓

Table C.5 The first trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *that* as the head of the SBAR clause.

SBAR head = if					
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1 #2
that	the little radio fizzes as other boats want to see we have found any fish – spotting location is everything in this sport .	the little radio fizzes as other boats want to see <u>that</u> – spotting location is everything in this sport .		1. we have found any fish 2. have found any fish 3. if we have found any fish	X X
this	" it 's hard to know people are responding truthfully .	" it 's hard to know if <u>this</u> .		1. people are responding truthfully 2. if people are responding truthfully 3. are responding truthfully	X X
that	he was further questioned to determine he was " a real working man or an exploiter . "	he was further questioned to determine <u>that</u> .		1. he was " a real working man or an exploiter 2. if he was " a real working man or an exploiter 3. was " a real working man or an exploiter	X ✓
that	asked the soviets , like chinese officials , wo n't one day face a similar conflict between the desire to liberalize economically and yet retain political control , mr. lee said , " i would think that the soviets face a deeper dilemma because they have been more in blinkers than the chinese – i mean keeping their people cut off from the outside world . "	asked <u>that</u> , mr. lee said , " i would think that the soviets face a deeper dilemma because they have been more in blinkers than the chinese – i mean keeping their people cut off from the outside world . "		1. the soviets , like chinese officials , wo n't one day face a similar conflict between the desire to liberalize economically and yet retain political control 2. if the soviets , like chinese officials , wo n't one day face a similar conflict between the desire to liberalize economically and yet retain political control	X X
this	voters generally agree when they are given a chance to decide they want to sink their own tax dollars into a new mega-stadium .	voters generally agree when they are given a chance to decide <u>this</u> .		1. they want to sink their own tax dollars into a new mega-stadium 2. want to sink their own tax dollars into a new mega-stadium 3. if they want to sink their own tax dollars into a new mega-stadium	X X

this	asked the administration agreed , he curtly replied : " the administration will have to speak for itself . "	asked <u>this</u> , he curtly replied : " the administration will have to speak for itself . "	<ol style="list-style-type: none"> 1. the administration agreed 2. the administration 3. if the administration agreed 	X	X
that	in his review of " saturday night with connie chung , " tom shales , the tv critic of the washington post and generally an admirer of cbs , wrote that while the show is " impressive , ... one has to wonder this is the proper direction for a network news division to take . "	in his review of " saturday night with connie chung , " tom shales , the tv critic of the washington post and generally an admirer of cbs , wrote that while the show is " impressive , ... one has to wonder <u>that</u> . "	<ol style="list-style-type: none"> 1. this is the proper direction for a network news division to take 2. if this is the proper direction for a network news division to take 3. is the proper direction for a network news division to take 	X	X
that	nbi , a maker of word-processing systems , said it ca n't predict any of the preferred stock will be converted .	nbi , a maker of word-processing systems , said it ca n't predict <u>that</u> .	<ol style="list-style-type: none"> 1. any of the preferred stock will be converted 2. if any of the preferred stock will be converted 	✓	✓
this	bertin nadeau , newly appointed chairman and interim chief executive of provigo , would n't say mr. lortie was asked to leave .	bertin nadeau , newly appointed chairman and interim chief executive of provigo , would n't say <u>this</u> .	<ol style="list-style-type: none"> 1. mr. lortie was asked to leave 2. if mr. lortie was asked to leave 	✓	✓
that	since then , the government has left observers wondering it ever meant to join .	since then , the government has left observers wondering <u>that</u> .	<ol style="list-style-type: none"> 1. it ever meant to join 2. if it ever meant to join 	X	X

Table C.6 The first trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *if* as the head of the SBAR clause.

SBAR head = as						
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1	#2
this	positive currency rates and strong sales growth led to a substantial rise in consolidated profit in the period , although the company did n't provide figures , as is customary with swiss companies .	positive currency rates and strong sales growth led to a substantial rise in consolidated profit in the period , although the company did n't provide figures , because of <u>this</u> .		<ol style="list-style-type: none"> 1. is customary with swiss companies 2. as is customary with swiss companies 	X	X
therefore	the representative responded that noriega had numerous assets in place in nicaragua and could accomplish many essential things , just noriega had helped -lcb- the u.s. . -rcb- the previous year in blowing up a sandinista arsenal . "	the representative responded that noriega had numerous assets in place in nicaragua and could accomplish many essential things , just <u>therefore</u> . "		<ol style="list-style-type: none"> 1. noriega had helped -lcb- the u.s. . -rcb- the previous year in blowing up a sandinista arsenal 2. had helped -lcb- the u.s. . -rcb- the previous year in blowing up a sandinista arsenal 	X	X
this	gray friday reflects a panic mainly by the takeover arbitragers , rather than the small investor , their highly margined investments in the " deal " stocks are jeopardized * by the unexpected drying up of the lubricant for deal financing .	gray friday reflects a panic mainly by the takeover arbitragers , rather than the small investor , because of <u>this</u> .		<ol style="list-style-type: none"> 1. their highly margined investments in the " deal " stocks are jeopardized by the unexpected drying up of the lubricant for deal financing 2. as their highly margined investments in the " deal " stocks are jeopardized by the unexpected drying up of the lubricant for deal financing 	✓	✓
this	mr. barnicle was hardly kinder to the renderings of colleagues michael madden -lrb- " appears to be a pervert " -rrb- , will mcdonough -lrb- " looks he drove for abe lincoln " -rrb- or bella english , whose " little girl now screams hysterically every time she sees a newspaper . "	mr. barnicle was hardly kinder to the renderings of colleagues michael madden -lrb- " appears to be a pervert " -rrb- , will mcdonough -lrb- " looks because of <u>this</u> " -rrb- or bella english , whose " little girl now screams hysterically every time she sees a newspaper . "		<ol style="list-style-type: none"> 1. he drove for abe lincoln 2. drove for abe lincoln 	X✱	X
that	adds robert juliano , the head lobbyist for a variety of interests that want to protect the tax deduction for travel and entertainment expenses : " it appears the whole thing is wide open again . "	adds robert juliano , the head lobbyist for a variety of interests that want to protect the tax deduction for travel and entertainment expenses : " it appears because of <u>that</u> . "		<ol style="list-style-type: none"> 1. the whole thing is wide open again 	X✱	X

this	" i feel people should be allowed * to remember players they were . "	" i feel people should be allowed to remember players because of <u>this</u> . "	1. they were 2. as they were	X X
this	though he nominally supports both programs , mr. bush has n't been a passionate champion of either cause , mr. reagan was .	though he nominally supports both programs , mr. bush has n't been a passionate champion of either cause , due to <u>this</u> .	1. mr. reagan was 2. as mr. reagan was	X X
therefore	try they might , the communists could neither replace nor break him .	try <u>therefore</u> , the communists could neither replace nor break him .	1. they might 2. as they might	X X
that	ual's stock skidded an additional \$ 24.875 , to \$ 198 , british airways indicated it may balk at any hastily revised version of the aborted \$ 6.79 billion buy-out of united air 's parent .	ual 's stock skidded an additional \$ 24.875 , to \$ 198 , due to <u>that</u> .	1. british airways indicated it may balk at any hastily revised version of the aborted \$ 6.79 billion buy-out of united air 's parent 2. as british airways indicated it may balk at any hastily revised version of the aborted \$ 6.79 billion buy-out of united air 's parent	✓ ✓
that	huge machines that look as they came from the star wars desert-battle scene lumber among the dunes .	huge machines that look due to <u>that</u> lumber among the dunes .	1. they came from the star wars desert-battle scene 2. came from the star wars desert-battle scene	X X

Table C.7 The first trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *as* as the head of the SBAR clause.

SBAR head = 0					
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1 #2
it	In a filing with the Securities and Exchange Commission , DPC Acquisition said it expects it will need about \$ 215 million to buy the shares and pay related fees and expenses .	in a filing with the securities and exchange commission , dpc acquisition said it expects it .		<ol style="list-style-type: none"> 1. it will need about \$ 215 million to buy the shares and pay related fees and expenses, 2. will need about \$ 215 million to buy the shares and pay related fees and expenses, 3. it expects it will need about \$ 215 million to buy the shares and pay related fees and expenses, 	✓ ✓
that	In a filing with the Securities and Exchange Commission , DPC Acquisition said it expects it will need about \$ 215 million to buy the shares and pay related fees and expenses .	in a filing with the securities and exchange commission , dpc acquisition said that .		<ol style="list-style-type: none"> 1. it expects it will need about \$ 215 million to buy the shares and pay related fees and expenses, 2. it will need about \$ 215 million to buy the shares and pay related fees and expenses, 3. expects it will need about \$ 215 million to buy the shares and pay related fees and expenses, 	✓ ✓
it	Richard Luehrs , president of the Newport Harbor Area Chamber of Commerce , calls boiler rooms a “ negative we wish we could get rid of . ”	richard luehrs , president of the newport harbor area chamber of commerce , calls boiler rooms a “ negative we wish it . ”		<ol style="list-style-type: none"> 1. we could get rid of, 2. could get rid of, 3. we wish we could get rid of, 	✗ ✗
that	Despite federal approval , General Dynamics says it decided it wo n’t go ahead with the matching program .	despite federal approval , general dynamics says it decided that .		<ol style="list-style-type: none"> 1. it wo n’t go ahead with the matching program, 2. it decided it wo n’t go ahead with the matching program, 3. wo n’t go ahead with the matching program, 	✗ ✓

that	Some of those debtholders have filed a suit , saying they believed they were buying government-insured certificates of deposit .	some of those debtholders have filed a suit , saying they believed it .	<ol style="list-style-type: none"> 1. they were buying government-insured certificates of deposit, 2. they believed they were buying government-insured certificates of deposit, 3. were buying government-insured certificates of deposit, 	✓ ✓✚
it	Some of those debtholders have filed a suit , saying they believed they were buying government-insured certificates of deposit	some of those debtholders have filed a suit , saying it .	<ol style="list-style-type: none"> 1. they believed they were buying government-insured certificates of deposit, 2. they were buying government-insured certificates of deposit, 3. believed they were buying government-insured certificates of deposit, 	✓ ✓✚
it	Compaq , which said it discovered the bugs , still plans to announce new 486 products on Nov. 6 .	compaq , which said it , still plans to announce new 486 products on nov. 6 .	<ol style="list-style-type: none"> 1. it discovered the bugs, 2. discovered the bugs, 3. the bugs, 4. about the bugs, 	✓ ✓✚
it	Separately , the New York Times reported that the Israeli government had provided its correspondent in Jerusalem with different documents that Israel said prove the PLO has been conducting terrorism from the occupied Arab territories .	separately , the new york times reported that the israeli government had provided its correspondent in jerusalem with different documents that israel said it .	<ol style="list-style-type: none"> 1. prove the plo has been conducting terrorism from the occupied arab territories, 2. the plo has been conducting terrorism from the occupied arab territories, 	✗ ✗
it	If you bought , you wish you had n't ? , and if you sold , you wish you had n't ? . "	if you bought , you wish it , and if you sold , you wish you had n't . "	<ol style="list-style-type: none"> 1. you had n't, 2. had n't, 	✗ ✗
this	If you bought , you wish you had n't ? , and if you sold , you wish you had n't ? . "	if you bought , you wish you had n't , and if you sold , you wish this . "	<ol style="list-style-type: none"> 1. you had n't, 2. had n't, 	✗ ✗

Table C.8 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and θ as the head of the SBAR clause.

SBAR head = NONE						
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1	#2
this	That selling of futures contracts by elevators is what helps keep downward pressure on crop prices during the harvest .	that selling of futures contracts by elevators is what this .		<ol style="list-style-type: none"> 1. helps keep downward pressure on crop prices during the harvest, 2. keep downward pressure on crop prices during the harvest, 	X	X
that	While rival ABC News outstripped the competition in live coverage of the event by sheer luck – the network was broadcasting the World Series from Candlestick Park when the quake struck – NBC News was unable to get its signal out of San Francisco for the first hour after the quake .	while rival abc news outstripped the competition in live coverage of the event by sheer luck – the network was broadcasting the world series from candlestick park when that – nbc news was unable to get its signal out of san francisco for the first hour after the quake .		<ol style="list-style-type: none"> 1. the quake struck, 2. the quake, 	X	X
it	“ That ’s really where the leverage hurt , ” says Thomas Herzfeld , a Miami-based investment manager who specializes in closed-end funds .	“ that ’s really where it , ” says thomas herzfeld , a miami-based investment manager who specializes in closed-end funds .		<ol style="list-style-type: none"> 1. the leverage hurt, 2. the leverage, 	X	X
that	High cash positions help buffer a fund when the market falls .	high cash positions help buffer a fund when that .		<ol style="list-style-type: none"> 1. the market falls, 2. the market, 	X	X
this	To understand what Mr. Engelken means , one must go back to a sunny October afternoon in 1951 at New York ’s Polo Grounds stadium , where , it can be argued , the most dramatic moment in baseball history was played out .	to understand what this , one must go back to a sunny october afternoon in 1951 at new york ’s polo grounds stadium , where , it can be argued , the most dramatic moment in baseball history was played out .		<ol style="list-style-type: none"> 1. mr. engelken means, 2. mr. engelken, 3. to mr. engelken, 	X	X

that	Families that do not need the loan can make money simply by putting the loan in the bank and paying it back when the student graduates .	families that do not need the loan can make money simply by putting the loan in the bank and paying it back when that .	1. the student graduates, 2. the student,	X X
that	Prime will still manage Ramada 's domestic franchise system when the sale closes .	prime will still manage ramada 's domestic franchise system when that .	1. the sale closes, 2. the sale,	X X
it	Several phone calls and a visit to his broker 's office later , the dentist found out that the \$ 9,000 drop represented the current value of the premium he paid when he bought the CD , and that the amount was n't insured .	several phone calls and a visit to his broker 's office later , the dentist found out that the \$ 9,000 drop represented the current value of the premium he paid when it , and that the amount was n't insured . when this happened	1. he bought the cd, 2. bought the cd, 3. the cd,	X X
it	Mr. Lang surprised Time soon after joining forces when he said he would negotiate rates individually with advertisers , a practice common in broadcasting but considered taboo by magazine publishers .	mr. lang surprised time soon after joining forces when it .	1. he said he would negotiate rates individually with advertisers , a practice common in broadcasting but considered taboo by magazine publishers, 2. he would negotiate rates individually with advertisers , a practice common in broadcasting but considered taboo by magazine publishers, 3. said he would negotiate rates individually with advertisers , a practice common in broadcasting but considered taboo by magazine publishers,	X X
that	Pravda gave no estimate for overall unemployment but said an " Association of the Unemployed " has cropped up that says the number of jobless is 23 million Soviets , or 17 % of the work force .	pravda gave no estimate for overall unemployment but said an " association of the unemployed " has cropped up that that .	1. says the number of jobless is 23 million soviets , or 17 % of the work force, 2. the number of jobless is 23 million soviets , or 17 % of the work force,	X X

Table C.9 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *NONE* as the head of the SBAR clause.

		SBAR head = whether				
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1	#2
this	various ministries decided the products businessmen could produce and how much ? ; and government-owned banks controlled the financing of projects and monitored whether companies came through on promised plans .	various ministries decided the products businessmen could produce and how much ; and government-owned banks controlled the financing of projects and monitored this .		<ol style="list-style-type: none"> 1. companies came through on promised plans, 2. came through on promised plans, 	✓	✓
that	white house spokesmen last week said mr. bush is considering simply declaring that the constitution gives him the power , exercising a line-item veto and inviting a court challenge to decide whether he has the right .	white house spokesmen last week said mr. bush is considering simply declaring that the constitution gives him the power , exercising a line-item veto and inviting a court challenge to decide that .		<ol style="list-style-type: none"> 1. he has the right, 2. has the right, 	✓	✓
that	senate minority leader robert dole r. , kan. , for one , accepts this argument and earlier this year publicly urged mr. bush “ to use the line-item veto and allow the courts to decide whether or not it is constitutional . ”	senate minority leader robert dole r. , kan. , for one , accepts this argument and earlier this year publicly urged mr. bush “ to use the line-item veto and allow the courts to decide that . ”		<ol style="list-style-type: none"> 1. it is constitutional, 2. is constitutional, 	✓	✓
this	the appeals court , however , said the judge did n’t adequately consider whether the delay would actually hurt the chances of a fair trial .	the appeals court , however , said the judge did n’t adequately consider this .		<ol style="list-style-type: none"> 1. the delay would actually hurt the chances of a fair trial, 2. would actually hurt the chances of a fair trial, 	✓	✓
it	passport applicants now must give social security numbers , enabling the irs to see whether americans living abroad are filing required u.s. returns .	passport applicants now must give social security numbers , enabling the irs to see it .		<ol style="list-style-type: none"> 1. americans living abroad are filing required u.s. returns, 	✗	✓✚

this	the committee is to recommend at the end of the month whether doe should support cold fusion research .	the committee is to recommend at the end of the month this .	<ol style="list-style-type: none"> 1. doe should support cold fusion research, 2. should support cold fusion research, 	✓✚ ✓✚
it	at one point , he asked a worker whether he thought east germans were fleeing the country because of restrictive travel policies .	at one point , he asked a worker it .	<ol style="list-style-type: none"> 1. he thought east germans were fleeing the country because of restrictive travel policies, 2. thought east germans were fleeing the country because of restrictive travel policies, 	✓✚ ✓✚
it	“ it ’s hard to know whether it was intended to be funny , ” says the east berlin shopkeeper , “ but everyone i know laughed about it .	“ it ’s hard to know it , ” says the east berlin shopkeeper , “ but everyone i know laughed about it .	<ol style="list-style-type: none"> 1. it was intended to be funny, 2. was intended to be funny, 	✓ ✓✚
that	four workers at gte corp. ’s headquarters have been diagnosed as having hepatitis , and city health officials are investigating whether a cafeteria worker may have exposed hundreds of other gte employees to the viral infection , company and city officials said .	four workers at gte corp. ’s headquarters have been diagnosed as having hepatitis , and city health officials are investigating that , company and city officials said .	<ol style="list-style-type: none"> 1. a cafeteria worker may have exposed hundreds of other gte employees to the viral infection, 	✗ ✓✚
that	the census bureau counts all cash income in determining whether families are below the line , but it does n’t consider other government benefits , such as medicare .	the census bureau counts all cash income in determining that , but it does n’t consider other government benefits , such as medicare .	<ol style="list-style-type: none"> 1. families are below the line, 2. are below the line, 	✓ ✓

Table C.10 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *whether* as the head of the SBAR clause.

SBAR head = because					
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1 #2
it	these securities get top credit ratings because the issuers have put aside u.s. bonds that will be sold to pay off holders when the municipals are retired .	these securities get top credit ratings because of it .		<ol style="list-style-type: none"> 1. the issuers have put aside u.s. bonds that will be sold to pay off holders when the municipals are retired, 2. have put aside u.s. bonds that will be sold to pay off holders when the municipals are retired, 	✓ ✓✚
it	a few years ago , state farm , the nation 's largest home insurer , stopped buying reinsurance because no one carrier could provide all the coverage that it needed and the company found it cheaper to self-reinsure .	a few years ago , state farm , the nation 's largest home insurer , stopped buying reinsurance because of it and the company found it cheaper to self-reinsure .		<ol style="list-style-type: none"> 1. no one carrier could provide all the coverage that it needed, 	✓ ✗
that	john bruner , associate director of communications for cincinnati public schools , said channel one was rejected because students watching the program did n't fare particularly better on a 28-question current events quiz than a control school without the program and school absences were almost unchanged during the period when the program was being aired .	john bruner , associate director of communications for cincinnati public schools , said channel one was rejected because of that .		<ol style="list-style-type: none"> 1. students watching the program did n't fare particularly better on a 28-question current events quiz than a control school without the program and school absences were almost unchanged during the period when the program was being aired, 	✓ ✓
that	" cheerios and honey nut cheerios have eaten away sales normally going to kellogg 's corn-based lines simply because they are made of oats , " says merrill lynch food analyst william maguire .	" cheerios and honey nut cheerios have eaten away sales normally going to kellogg 's corn-based lines simply because of that , " says merrill lynch food analyst william maguire .		<ol style="list-style-type: none"> 1. they are made of oats, 2. are made of oats, 	✓ ✓
that	general mills , meanwhile , finds itself constrained from boosting sales further because its plants are operating at capacity .	general mills , meanwhile , finds itself constrained from boosting sales further due to that .		<ol style="list-style-type: none"> 1. its plants are operating at capacity, 	✓ ✓✚

this	we know firsthand the discrimination addressed by the act : to be told there 's no place for your child in school ; to spend lonely hours at home because there is no transportation for someone in a wheelchair ; to be denied employment because you are disabled .	we know firsthand the discrimination addressed by the act : to be told there 's no place for your child in school ; to spend lonely hours at home due to this ; to be denied employment because you are disabled .	1. there is no transportation for someone in a wheelchair, 2. is no transportation for someone in a wheelchair,	✓	✗
this	we know firsthand the discrimination addressed by the act : to be told there 's no place for your child in school ; to spend lonely hours at home because there is no transportation for someone in a wheelchair ; to be denied employment because you are disabled .	we know firsthand the discrimination addressed by the act : to be told there 's no place for your child in school ; to spend lonely hours at home because there is no transportation for someone in a wheelchair ; to be denied employment due to this .	1. you are disabled, 2. are disabled, 3. are themselves disabled,	✓✘	✓✘
that	he also said traders should keep an eye on the stock market , because “ if the stock market rallies , that could spell trouble for the precious metals . ”	he also said traders should keep an eye on the stock market , due to that . ”	1. if the stock market rallies , that could spell trouble for the precious metals,	✓	✓
it	but japanese agencies are cautious about expanding abroad because client relationships are different .	but japanese agencies are cautious about expanding abroad due to it .	1. client relationships are different,	✓✘	✓
it	the slowdown raises questions about the economy 's strength because spending fueled much of the third-quarter gnp growth .	the slowdown raises questions about the economy 's strength because of it .	1. spending fueled much of the third-quarter gnp growth, 2. fueled much of the third-quarter gnp growth,	✓✘	✓✘

Table C.11 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *because* as the head of the SBAR clause.

SBAR head = since-pp						
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1	#2
it	carried even further , some investors assumed that since leveraged buy-outs are the only thing propping up stock prices , the market would collapse if no more lbos could be done .	carried even further , some investors assumed that since leveraged buy-outs are the only thing propping up stock prices , the market would collapse because of it .		<ol style="list-style-type: none"> 1. no more lbos could be done, 2. if no more lbos could be done, 	✓	✓
it	the board could eventually come under some pressure to sell the company because its members can be ousted by a majority shareholder vote , particularly since one-third of ual stock is held by takeover stock speculators who favor a sale .	the board could eventually come under some pressure to sell the company because its members can be ousted by a majority shareholder vote , particularly because of it .		<ol style="list-style-type: none"> 1. one-third of ual stock is held by takeover stock speculators who favor a sale, 	✓	✓✚
this	traditionally , boiler rooms operate on the cheap , since few , if any , customers ever visit their offices .	traditionally , boiler rooms operate on the cheap , because of this .		<ol style="list-style-type: none"> 1. few , if any , customers ever visit their offices, 	✓	✓
this	' investigators stress that building owners are victims , too , since boiler rooms often leave without paying rent . '	investigators stress that building owners are victims , too , because of this .		<ol style="list-style-type: none"> 1. boiler rooms often leave without paying rent, 	✓	✓
that	a licensed government intellectual , francis fukuyama , recently announced in the national interest that history is , so to speak , at an end since the course of human progress has now culminated in the glorious full stop of american civilization .	a licensed government intellectual , francis fukuyama , recently announced in the national interest that history is , so to speak , at an end because of that .		<ol style="list-style-type: none"> 1. the course of human progress has now culminated in the glorious full stop of american civilization, 	✓	✓

that	“ there are no commercials to make up for since we ’re going to eventually broadcast the world series , ” said a network spokesman . “	there are no commercials to make up for because of that , ” said a network spokesman .	1. we ’re going to eventually broadcast the world series, 2. ’re going to eventually broadcast the world series,	X X
this	linear technology , milpitas , calif. , called the settlement “ positive , ” since products covered by the disputed patents account for about 20 % of its annual sales .	linear technology , milpitas , calif. , called the settlement “ positive , ” because of this .	1. products covered by the disputed patents account for about 20 % of its annual sales,	✓ ✓
it	the 19-member cabinet is led by prime minister jan syse , who acknowledged a “ difficult situation ” since the coalition controls only 62 seats in oslo ’s 165-member legislature .	the 19-member cabinet is led by prime minister jan syse , who acknowledged a “ difficult situation ” due to it .	1. the coalition controls only 62 seats in oslo ’s 165-member legislature,	✓ ✓✚
it	it ’s almost impossible to track the number of companies trashing junk mail , since the decision is usually made in the mail room – not the board room .	it ’s almost impossible to track the number of companies trashing junk mail , due to it .	1. the decision is usually made in the mail room – not the board room, 2. is usually made in the mail room – not the board room,	✓ ✓✚
this	apart from those two actions , mr. sikes and the three other commissioners said they expect to re-examine how at & t is regulated *ppa* since competition has increased .	apart from those two actions , mr. sikes and the three other commissioners said they expect to re-examine how at & t is regulated *ppa* due to this .	1. competition has increased, 2. has increased,	✓ ✓✚

Table C.12 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *since-prp* as the head of the SBAR clause.

		SBAR head = as-prp				
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1	#2
that	general motors corp. and ford motor co. are now going head to head in the markets for shares of jaguar plc , as gm got early clearance from the federal trade commission to boost its stake in the british luxury car maker .	general motors corp. and ford motor co. are now going head to head in the markets for shares of jaguar plc , because of that .		1. gm got early clearance from the federal trade commission to boost its stake in the british luxury car maker, 2. got early clearance from the federal trade commission to boost its stake in the british luxury car maker,	✓	✓
that	toy makers complain that electricity in guangdong has been provided only three days a week in recent months , down from five days a week , as the province 's rapid industrialization has outstripped its generating capacity .	toy makers complain that electricity in guangdong has been provided only three days a week in recent months , down from five days a week , due to that .		1. the province 's rapid industrialization has outstripped its generating capacity,	✓	✓
that	" we did n't trade much today , as our policy now is to wait and see , " said a fund manager at taisho life insurance co .	" we did n't trade much today , due to that , " said a fund manager at taisho life insurance co .		1. our policy now is to wait and see,	✓	✓
this	if you asked me to select a stock with the highest expected return , i would select a stock with the greatest amount of undiversifiable risk , as i am sure your pros do ? .	if you asked me to select a stock with the highest expected return , i would select a stock with the greatest amount of undiversifiable risk , because of this .		1. i am sure your pros do, 2. am sure your pros do,	✗	✗
that	analysts said skf 's results for the first nine months lived up to market expectations as brokerage firms had predicted a pretax profit of 1.74 billion to 1.86 billion kronor .	analysts said skf 's results for the first nine months lived up to market expectations due to that .		1. brokerage firms had predicted a pretax profit of 1.74 billion to 1.86 billion kronor,	✓	✓

this	but the proposals also display political savvy , couching some of the most controversial ideas in cautious language so as not to alienate powerful conservatives in the government who stand to lose out if they are implemented .	but the proposals also display political savvy , couching some of the most controversial ideas in cautious language so as not to alienate powerful conservatives in the government who stand to lose out because of this .	<ol style="list-style-type: none"> 1. they are implemented, 2. are implemented, 3. if they are implemented, 	✓ ✓
it	one airline official said about three times as many free-travel coupons are being turned in as in previous years – not surprisingly , as the airlines last year allowed many travelers to build up mileage at triple the normal rate .	one airline official said about three times as many free-travel coupons are being turned in as in previous years – not surprisingly , because of it .	<ol style="list-style-type: none"> 1. the airlines last year allowed many travelers to build up mileage at triple the normal rate, 	✓ ✗
this	the first analyst said that the japanese , as well as the chinese , bought copper earlier in the week in london , but that this purchasing has since slackened as the supply situation , at least over the long term , appears to have improved .	the first analyst said that the japanese , as well as the chinese , bought copper earlier in the week in london , but that this purchasing has since slackened due to this .	<ol style="list-style-type: none"> 1. the supply situation , at least over the long term , appears to have improved, 	✓ ✓
that	they also are concerned about the persistent strength of the dollar against the yen , as a weaker yen leads to higher import prices in japan and adds to domestic inflationary pressures .	they also are concerned about the persistent strength of the dollar against the yen , because of that .	<ol style="list-style-type: none"> 1. a weaker yen leads to higher import prices in japan and adds to domestic inflationary pressures, 	✓ ✓
this	but traders said the market lacks a base on which to set long-term buying strategy , as the future direction of u.s. interest rates remains unclear .	but traders said the market lacks a base on which to set long-term buying strategy , because of this .	<ol style="list-style-type: none"> 1. the future direction of u.s. interest rates remains unclear, 	✓ ✓

Table C.13 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *as-prp* as the head of the SBAR clause.

SBAR head = if						
suggested anaphor head	original sentence	suggested anaphoric sentence	all possible antecedents	#1	#2	
this	buyers can look forward to double-digit annual returns if they are right .	buyers can look forward to double-digit annual returns because of this .	1. they are right, 2. are right,	✓	✓	
this	but they will have disappointing returns or even losses if interest rates rise instead .	but they will have disappointing returns or even losses due to this .	1. interest rates rise instead,	✓	✓	
that	the short term bond fund , with an average maturity of 2 12 years , would deliver a total return for one year of about 10.6 % if rates drop one percentage point and a one-year return of about 6.6 % if rates rise by the same amount .	the short term bond fund , with an average maturity of 2 12 years , would deliver a total return for one year of about 10.6 % due to that and a one-year return of about 6.6 % if rates rise by the same amount .	1. rates drop one percentage point, 2. drop one percentage point,	✓	✓	
it	the short term bond fund , with an average maturity of 2 12 years , would deliver a total return for one year of about 10.6 % if rates drop one percentage point and a one-year return of about 6.6 % if rates rise by the same amount .	the short term bond fund , with an average maturity of 2 12 years , would deliver a total return for one year of about 10.6 % because of it and a one-year return of about 6.6 % rates rise by the same amount .	1. if rates drop one percentage point, 2. rates drop one percentage point,	✗	✓✱	
that	“ you get equity-like returns ” from bonds if you guess right on rates , says james e. wilson , a columbia , s.c. , planner .	“ you get equity-like returns ” from bonds due to that , says james e. wilson , a columbia , s.c. , planner .	1. you guess right on rates, 2. guess right on rates,	✓	✓	

this	james snedeker , senior vice president of gill & roeser inc. , a new york-based reinsurance broker , says insurers who took big losses this fall and had purchased little reinsurance in recent years will be asked to pay some pretty hefty rates if they want to buy reinsurance for 1990 .	james snedeker , senior vice president of gill & roeser inc. , a new york-based reinsurance broker , says insurers who took big losses this fall and had purchased little reinsurance in recent years will be asked to pay some pretty hefty rates because of this .	<ol style="list-style-type: none"> 1. they want to buy reinsurance for 1990, 2. want to buy reinsurance for 1990, 	✓ ✓
this	saturday , he amended his remarks to say that he would continue to abide by the cease-fire if the u.s. ends its financial support for the contras .	saturday , he amended his remarks to say that he would continue to abide by the cease-fire due to this .	<ol style="list-style-type: none"> 1. the u.s. ends its financial support for the contras, 2. ends its financial support for the contras, 	✓ ✓
that	the treasury also said noncompetitive tenders will be considered timely if postmarked no later than sunday , oct. 29 , and received no later than tomorrow .	the treasury also said noncompetitive tenders will be considered timely because of that .	<ol style="list-style-type: none"> 1. postmarked no later than sunday , oct. 29 , and received no later than tomorrow, 	✓ ✗
that	in addition , a second part b premium to cover the cost of the new program of insurance against catastrophic illness will rise to \$ 4.90 a month from \$ 4 , if congress does n't change the program .	in addition , a second part b premium to cover the cost of the new program of insurance against catastrophic illness will rise to \$ 4.90 a month from \$ 4 , because of that .	<ol style="list-style-type: none"> 1. congress does n't change the program, 2. does n't change the program, 	✓ ✓
this	but traders said the market 's tone could pick up this week if new york city 's \$ 787 million bond offering goes well .	but traders said the market 's tone could pick up this week due to this .	<ol style="list-style-type: none"> 1. new york city 's \$ 787 million bond offering goes well, 	✓ ✓

Table C.14 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *if* as the head of the SBAR clause.

SBAR head = after						
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1	#2
this	the announcement said the acquisition should be completed by december after a definitive agreement is completed and regulatory approval is received .	the announcement said the acquisition should be completed by december after this .		1. a definitive agreement is completed and regulatory approval is received,	✓	✓
this	dr. toseland , a toxicologist , said he was preparing an article for a british forensic medical journal raising the possibility that the deaths may have occurred after human insulin blunted critical warning signs indicating hypoglycemia , or low blood sugar , which can kill diabetics .	dr. toseland , a toxicologist , said he was preparing an article for a british forensic medical journal raising the possibility that the deaths may have occurred after this .		1. human insulin blunted critical warning signs indicating hypoglycemia , or low blood sugar , which can kill diabetics,	✓	✓
this	much of the excess spending will be pushed into fiscal 1991 , and in some cases is temporarily parked in slow-spending accounts in anticipation of being transferred to faster-spending areas after the budget scorekeeping is completed .	much of the excess spending will be pushed into fiscal 1991 , and in some cases is temporarily parked in slow-spending accounts in anticipation of being transferred to faster-spending areas after this .		1. the budget scorekeeping is completed,	✓	✓✚
it	the office also said mrs. marcos and her husband were n't brought to the u.s. against their will after mr. marcos was ousted as president .	the office also said mrs. marcos and her husband were n't brought to the u.s. against their will after it .		1. mr. marcos was ousted as president,	✓	✓
that	for the next two years , the bank board , which at the time was the agency responsible for regulating thrifts , failed to act – even after federal auditors warned in may 1987 that mr. keating had caused lincoln to become insolvent .	for the next two years , the bank board , which at the time was the agency responsible for regulating thrifts , failed to act – even after that .		1. federal auditors warned in may 1987 that mr. keating had caused lincoln to become insolvent,	✓	✓✚

this	in the midst of his 1988 re-election campaign , sen. riegler , chairman of the senate banking committee , returned \$ 76,000 in contributions after a detroit newspaper said that mr. keating had gathered the money for him about two weeks before the meeting with regulators .	in the midst of his 1988 re-election campaign , sen. riegler , chairman of the senate banking committee , returned \$ 76,000 in contributions after this .	1. a detroit newspaper said that mr. keating had gathered the money for him about two weeks before the meeting with regulators,	✓ ✓✚
this	sen. deconcini , after months of fending off intense press criticism , returned \$ 48,000 only last month , shortly after the government formally accused mr. keating of defrauding lincoln .	sen. deconcini , after months of fending off intense press criticism , returned \$ 48,000 only last month , shortly after this .	1. the government formally accused mr. keating of defrauding lincoln,	✓ ✓✚
it	it was the latest in a series of setbacks for the junk bond market , where prices began weakening last month after campeau hit a cash crunch .	it was the latest in a series of setbacks for the junk bond market , where prices began weakening last month after it .	1. campeau hit a cash crunch, 2. hit a cash crunch,	✓ ✓✚
that	hoyle dropped its initial \$ 13.35 billion -lr- \$ 20.71 billion -rrb- takeover bid after it received the extension , but said it would launch a new bid if and when the proposed sale of farmers to axa receives regulatory approval .	hoyle dropped its initial \$ 13.35 billion -lr- \$ 20.71 billion -rrb- takeover bid after that , but said it would launch a new bid if and when the proposed sale of farmers to axa receives regulatory approval .	1. it received the extension, 2. received the extension, 3. the extension,	✓ ✗
it	on the other hand , symbol technologies dropped 14 to 18 after shearson lehman hutton lowered its short-term investment rating on the stock and its 1989 earnings estimate , and commodore international fell 78 to 8 after the company said it expects to post a loss for the september quarter .	on the other hand , symbol technologies dropped 14 to 18 after it , and commodore international fell 78 to 8 after the company said it expects to post a loss for the september quarter .	1. shearson lehman hutton lowered its short-term investment rating on the stock and its 1989 earnings estimate,	✓ ✓

Table C.15 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *after* as the head of the SBAR clause.

		SBAR head = since-tmp				
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1	#2
then	the indian stock markets have been on a five-year high , with dips and corrections , since prime minister rajiv gandhi started liberalizing industry .	the indian stock markets have been on a five-year high , with dips and corrections , since then .		1. prime minister rajiv gandhi started liberalizing industry,	✓	✓
then	mr. tonkin , who has been feuding with the big three since he took office earlier this year , said that with half of the nation 's dealers losing money or breaking even , it was time for " emergency action . "	mr. tonkin , who has been feuding with the big three since then , said that with half of the nation 's dealers losing money or breaking even , it was time for " emergency action . "		1. he took office earlier this year, 2. took office earlier this year,	✓	✓
then	mr. tonkin , who has been feuding with the big three since he took office earlier this year , said that with half of the nation 's dealers losing money or breaking even , it was time for " emergency action . "	mr. tonkin , who has been feuding with the big three since he took office earlier this year , said since then . "		1. with half of the nation 's dealers losing money or breaking even , it was time for " emergency action , 2. that with half of the nation 's dealers losing money or breaking even , it was time for " emergency action ,	✗	✗ ²
then	shearson 's application is the first since the taiwan securities and exchange commission announced june 21 that it would allow foreign brokerage firms to do business in that country .	shearson 's application is the first since the taiwan securities and exchange commission announced june 21 since then .		1. it would allow foreign brokerage firms to do business in that country, 2. would allow foreign brokerage firms to do business in that country, 3. that it would allow foreign brokerage firms to do business in that country,	✗	✗ ³
then	" i have n't been able to get a decent night 's sleep since this has been going on , " he says .	" i have n't been able to get a decent night 's sleep since then , " he says .		1. this has been going on, 2. has been going on,	✓	✓

then	the existence of the guidelines has become known since president bush disclosed them privately to seven republican senators at a white house meeting last monday .	the existence of the guidelines has become known since then .	1. president bush disclosed them privately to seven republican senators at a white house meeting last monday,	✓	✓
then	import competition for u.s. furs has risen sharply since furriers started aggressively marketing “ working-girl mink ” and similar lower-priced imported furs in recent years .	import competition for u.s. furs has risen sharply since then .	1. furriers started aggressively marketing “ working-girl mink ” and similar lower-priced imported furs in recent years, 2. started aggressively marketing “ working-girl mink ” and similar lower-priced imported furs in recent years,	✓	✓
then	the asset privatization trust , the agency chiefly responsible for selling government-held properties , has recorded sales of more than \$ 500 million since it began functioning in december 1986 .	the asset privatization trust , the agency chiefly responsible for selling government-held properties , has recorded sales of more than \$ 500 million since then .	1. it began functioning in december 1986, 2. began functioning in december 1986,	✓	✓
then	sir john has spurned ford ’s advances since the u.s. auto giant launched a surprise bid for as much as 15 % of jaguar last month .	sir john has spurned ford ’s advances since then .	1. the u.s. auto giant launched a surprise bid for as much as 15 % of jaguar last month,	✓	✓
then	he said jaguar started negotiating with gm and several other car makers over a year ago , but the rest “ dropped by the wayside ever since the share price went above # 4 -lrb- \$ 6.30 -rrb- a share . ”	he said jaguar started negotiating with gm and several other car makers over a year ago , but the rest “ dropped by the wayside ever since then . ”	1. the share price went above # 4 -lrb- \$ 6.30 -rrb- a share,	✓	✓

Table C.16 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *since-imp* as the head of the SBAR clause.

		SBAR head = until			
suggested anaphor head	original sentence	suggested sentence	anaphoric sentence	all possible antecedents	#1 #2
this	these insurance company contracts feature some of the same tax benefits and restrictions as non-deductible individual retirement accounts : investment gains are compounded without tax consequences until money is withdrawn , but a 10 % penalty tax is imposed on withdrawals made before age 59 12 .	these insurance company contracts feature some of the same tax benefits and restrictions as non-deductible individual retirement accounts : investment gains are compounded without tax consequences until this , but a 10 % penalty tax is imposed on withdrawals made before age 59 12 .		1. money is withdrawn, 2. is withdrawn,	X X
it	however , workers ca n't break ground until legal maneuvers to block the complex are resolved , moves which caused the signing to remain questionable up to the last moment .	however , workers ca n't break ground until it .		1. legal maneuvers to block the complex are resolved , moves which caused the signing to remain questionable up to the last moment,	X X
it	a put option gives its holder the right -lrb- but not the obligation -rrb- to sell a stock -lrb- or stock index -rrb- for a specified price -lrb- the strike price -rrb- until the option expires .	a put option gives its holder the right -lrb- but not the obligation -rrb- to sell a stock -lrb- or stock index -rrb- for a specified price -lrb- the strike price -rrb- until it .		1. the option expires, 2. the option,	X X
this	last year , gm had a different program in place that continued rewarding dealers until all the 1989 models had been sold .	last year , gm had a different program in place that continued rewarding dealers until this .		1. all the 1989 models had been sold,	✓ X
that	they just keep digging me in deeper until i reach the point where i give up and go away . ”	they just keep digging me in deeper until that . ”		1. i reach the point where i give up and go away, 2. reach the point where i give up and go away, 3. the point where i give up and go away,	X X

this	what better place to turn than sen. edward kennedy 's labor committee , that great stove of government expansionism , where many a stagnant pot of porridge is kept on the back burner until it can be brought forward and presented as nouvelle cuisine ?	what better place to turn than sen. edward kennedy 's labor committee , that great stove of government expansionism , where many a stagnant pot of porridge is kept on the back burner until this ?	1. it can be brought forward and presented as nouvelle cuisine, 2. can be brought forward and presented as nouvelle cuisine,	X X
that	but the company said the amount can't be determined until it knows how many managers opt to retire .	but the company said the amount can't be determined until that .	1. it knows how many managers opt to retire, 2. knows how many managers opt to retire,	X X
this	mutual funds arrived in the u.s. during the roaring twenties -lrb- they had been in britain for a century -rrb- , but they did n't boom until the money market fund was created in the 1970s .	mutual funds arrived in the u.s. during the roaring twenties -lrb- they had been in britain for a century -rrb- , but they did n't boom until this .	1. the money market fund was created in the 1970s,	✓ ✓✚
it	shimson gottesfeld of los alamos national laboratory said researchers there detected a burst of neutrons from an early cold fusion experiment last april but decided not to announce it until they could confirm it .	shimson gottesfeld of los alamos national laboratory said researchers there detected a burst of neutrons from an early cold fusion experiment last april but decided not to announce it until it .	1. they could confirm it, 2. could confirm it,	X X
that	likewise , mutual funds remained relatively flat until i made what was , for me , a serious investment .	likewise , mutual funds remained relatively flat until that .	1. i made what was , for me , a serious investment, 2. made what was , for me , a serious investment,	✓ ✓✚

Table C.17 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *until* as the head of the SBAR clause.

SBAR head = while						
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1	#2
that	hewlett-packard , a palo alto , calif. , computer company , said it acquired the stock “ to develop and maintain a strategic partnership in which each company remains independent while working together to market and sell their products . ”	hewlett-packard , a palo alto , calif. , computer company , said it acquired the stock “ to develop and maintain a strategic partnership in which each company remains independent during that . ”		1. working together to market and sell their products,	X	X
this	but the staffs at some of those locations will be slashed while at others the work force will be increased .	but the staffs at some of those locations will be slashed during this .		1. at others the work force will be increased,	X	X
this	kellogg 's current share is believed to be slightly under 40 % while general mills ' share is about 27 % .	kellogg 's current share is believed to be slightly under 40 % during this .		1. general mills ' share is about 27 % ,	X	X
this	imports rose 11 % to 18.443 trillion lire in september from a year earlier , while exports rose 17 % to 16.436 trillion lire .	imports rose 11 % to 18.443 trillion lire in september from a year earlier , during this .		1. exports rose 17 % to 16.436 trillion lire, 2. rose 17 % to 16.436 trillion lire,	X	X
that	in the nine months , imports rose 20 % to 155.039 trillion lire , while exports grew 18 % to 140.106 trillion lire .	in the nine months , imports rose 20 % to 155.039 trillion lire , during that .		1. exports grew 18 % to 140.106 trillion lire, 2. grew 18 % to 140.106 trillion lire,	X	X

that	import values are calculated on a cost , insurance and freight -lrb- c.i.f . -rrb- basis , while exports are accounted for on a free-on-board -lrb- f.o.b . -rrb- basis .	import values are calculated on a cost , insurance and freight -lrb- c.i.f . -rrb- basis , during that .	<ol style="list-style-type: none"> 1. exports are accounted for on a free-on-board -lrb- f.o.b . -rrb- basis, 2. are accounted for on a free-on-board -lrb- f.o.b . -rrb- basis, 	X X
this	a new in-house magazine , kidder world – which will focus on the firm ’s synergy strategy , says mr. carpenter – confides that on weekends mr. newquist “ often gets value-added ideas while flying his single-engine cessna centurion on the way to nantucket . ”	a new in-house magazine , kidder world – which will focus on the firm ’s synergy strategy , says mr. carpenter – confides that on weekends mr. newquist “ often gets value-added ideas during this . ”	<ol style="list-style-type: none"> 1. flying his single-engine cessna centurion on the way to nantucket, 	X X
that	congress sent to president bush an \$ 8.5 billion military construction bill that cuts spending for new installations by 16 % while revamping the pentagon budget to move more than \$ 450 million from foreign bases to home-state projects .	congress sent to president bush an \$ 8.5 billion military construction bill that cuts spending for new installations by 16 % during that .	<ol style="list-style-type: none"> 1. revamping the pentagon budget to move more than \$ 450 million from foreign bases to home-state projects, 	X X
that	total pentagon requests for installations in west germany , japan , south korea , the united kingdom and the philippines , for example , are cut by almost two-thirds , while lawmakers added to the military budget for construction in all but a dozen states at home .	total pentagon requests for installations in west germany , japan , south korea , the united kingdom and the philippines , for example , are cut by almost two-thirds , during that .	<ol style="list-style-type: none"> 1. lawmakers added to the military budget for construction in all but a dozen states at home, 2. added to the military budget for construction in all but a dozen states at home, 	X X
this	but now the companies are getting into trouble because they undertook a record expansion program while they were raising prices sharply .	but now the companies are getting into trouble because they undertook a record expansion program during this .	<ol style="list-style-type: none"> 1. they were raising prices sharply , 2. were raising prices sharply, 	✓ X

Table C.18 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *while* as the head of the SBAR clause.

SBAR head = <i>as-tmp</i>						
suggested anaphor head	original sentence	suggested sentence	anaphoric	all possible antecedents	#1	#2
this	mr. sim figures it will be easier to turn barry wright around since he 's now in the driver 's seat .	mr. sim figures it will be easier to turn barry wright around during this .		1. he 's now in the driver 's seat, 2. 's now in the driver 's seat, 3. since he 's now in the driver 's seat,	X	X
this	concerning your sept. 21 page-one article on prince charles and the leeches : it 's a few hundred years since england has been a kingdom .	concerning your sept. 21 page-one article on prince charles and the leeches : it 's a few hundred years during this .		1. england has been a kingdom, 2. since england has been a kingdom, 3. has been a kingdom,	X	X
that	mr. sim figures it will be easier to turn barry wright around since he 's now in the driver 's seat .	mr. sim figures it will be easier to turn barry wright around during that .		1. he 's now in the driver 's seat, 2. 's now in the driver 's seat, 3. since he 's now in the driver 's seat,	X	X
that	concerning your sept. 21 page-one article on prince charles and the leeches : it 's a few hundred years since england has been a kingdom .	concerning your sept. 21 page-one article on prince charles and the leeches : it 's a few hundred years during that .		1. england has been a kingdom, 2. since england has been a kingdom, 3. has been a kingdom,	X	X

Table C.19 Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and *as* as the head of the SBAR clause.

Appendix D

Data Management

Resources for Chapter 3. The heiDATA repository available at <https://doi.org/10.11588/data/LWN9XE> contains the code for reproducing experiments presented in Chapter 3 and the corresponding NAACL-HLT paper (Marasović and Frank, 2018). In particular,

- The MPQA 2.0 corpus is available at http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/. In the repository we provide the ipython notebook `mpqa2-pytools.ipynb` to demonstrate how to prepare the corpus for our models. The cross-validation data splits can be found in the `datasplit` folder.
- The SRL data is provided by the CoNLL-2005 Shared Task available at <http://www.lsi.upc.edu/~srlconll/>. However, the original words are from the Penn Treebank dataset which is not publicly available, but can be purchased at <https://catalog.ldc.upenn.edu/LDC99T42>.
- To train models run the `main.py` script. For example, `python main.py -adv_coef 0.0 -model fs -exp_setup_id new -n_layers_orl 0 -begin_fold 0 -end_fold 4`. The results will be written in the `outputs` folder.

Resources for Chapter 4. The heiDATA repository available at <https://doi.org/10.11588/data/UDMPY5> contains the code for reproducing experiments presented in Chapter 4 and the EMNLP paper (Marasović et al., 2017). In particular,

- Scripts and detailed instructions how to extract the silver training data from a parsed corpus as it is described in Chapter 4 (Section 4.3) are provided in the folder `Silver Data`. We extracted the silver data from the manually aparsed WSJ corpus that can be purchased at <https://catalog.ldc.upenn.edu/LDC2000T43> and the automatically

parsed NYT corpus from the annotated Gigaword that can be purchased at <https://catalog ldc.upenn.edu/LDC2003T05>.

- The ARRAU corpus (Uryupina et al., 2016; Poesio et al., 2018) can be purchased at <https://catalog ldc.upenn.edu/LDC2013T22>. The instructions for extracting the subset of abstract anaphors from the full ARRAU corpus are provided in `Gold Data/arrau_csn/instructions_arrau_construction.txt`.
- The ASN corpus (Kolhatkar et al., 2013b) is provided in personal communication with the authors. The authors provided us with the workers' annotations of the sentence with the antecedent, the antecedents chosen by the workers and links to the NYT corpus which can be purchased at <https://catalog ldc.upenn.edu/LDC2008T19>.
- The scripts to re-produce the CSN corpus can be found in the repository `Gold Data/arrau_csn/csn`.
- The CoNLL-12 shared task dataset (Pradhan et al., 2012) is available at <http://conll.cemantix.org/2012/data.html>. We used the `cort` library available at <https://github.com/smartschat/cort> to extract *this*, *that*, and *it* pronouns in the CoNLL-12 shared task dataset whose preceding mention in the coreference cluster is verbal.
- The script `Gold Data/process_aar_data.py` should be used to prepare the ASN corpus, the CoNLL-12 corpus, and the ARRAU corpus, for our models.
- The `EMNLP 2017` folder contains scripts for training and evaluating models described in Marasović et al. (2017).
- The `Thesis` folder contains scripts for training and evaluating models described in Chapter 4 (Section 4.5).

List of Figures

2.1	The GOODFOR/BADFOR graph of Example (5a): <i>The bill would lower health care costs, which would be a tremendous positive change across the entire health-care system.</i>	14
2.2	Example how resolution of anaphors results in more explicit sentiments relations and hence better sentiment propagation in the graph-based sentiment inference model.	17
2.3	MPQA annotation scheme. Figure from Wilson (2008).	27
2.4	The creation of the ANS and CSN corpora (Kolhatkar et al., 2013b).	41
2.5	Feed-Forward Neural Network (FFNN).	49
2.6	Convolution.	52
2.7	The Convolutional Neural Network for sentence classification. Figure replicated from Zhang and Wallace (2017).	53
2.8	Hard-parameter sharing architectures.	54
2.9	Adversarial training of the SP-MTL model.	55
3.1	Comparison of MTL models using violin plots.	67
4.1	MR-LSTM: Mention-Ranking LSTM-Siamese Neural Network.	79
4.2	A general pattern for creating anaphoric sentence–antecedent pairs.	82
4.3	Adversarial MR-LSTM (forward pass).	116
4.4	Adversarial MR-LSTM (backward pass).	116
4.5	Multi-task learning MR-LSTM.	121
4.6	Multi-task learning MR-LSTM (backward pass).	121
4.7	Comparison of success@n scores across different training configurations and windows $W_i = \{\text{AnaphS}_{-i}, \dots, \text{AnaphS}\}$, $i \in \{1, 2, 3, 4\}$. We report the best MTL result for the ARRAU corpus between the results obtained with three scorers.	123

List of Tables



2.1	Sentiment implicature rulesets of Deng and Wiebe (2014). Sentiment in the first column is explicitly stated in text and implies sentiment in the second column. POSPAR(writer,theme/event) denotes that the writer of a given document is positive toward theme of an event or event itself (and similarly for NEGP AIR).	15
2.2	Sentiment implicature rulesets of Deng and Wiebe (2015a). Information in the first column is explicitly stated in text and implies sentiment in the second column.	21
2.3	The recap of Deng and Wiebe’s work on sentiment inference with emphasis on the following: dependence on existing systems for capturing explicit sentiment, usage of anaphora resolution, and focus on discourse beyond a single sentence.	24
2.4	MPQA 2.0 topic categories.	26
2.5	Features for coreference resolution in Strube et al. (2002).	43
2.6	Features for shell noun resolution in Kolhatkar et al. (2013b).	44
2.7	Features for pronoun p and candidate v for the pronoun resolution in Jauhar et al. (2015). Features marked with • were selected, and those marked with - were discarded by feature selection.	46
3.1	Output of the AllenNLP SRL demo (He et al., 2017) for the sentence <i>Australia said it feared violence if voters thought the election had be stolen</i> . . .	58
3.2	Oversampling for the Z&X-STL model (Zhou and Xu, 2015). The sentence <i>Australia said it feared violence if voters thought the election had be stolen</i> is oversampled six times for six different predicates that occur in it. Each row illustrates the label sequence of the corresponding predicate.	61
3.3	Datasets with the number of SRL predicates/ORL opinions in train, dev & test set, size of label inventory.	62
3.4	ORL 10-fold CV results.	64

3.5	ORL repeated 4-fold CV results.	65
3.6	Randomly sampled hyperparameters for comparing MTL models.	67
3.7	The dev examples for which both models (FS-MTL, Z&X-STL) <i>correctly</i> predict the <i>holder</i> in 6/8 trials.	68
3.8	Statistics of <i>holder</i> prediction.	68
3.9	The dev examples for which both models (FS-MTL, Z&X-STL) <i>correctly</i> predict the <i>target</i> in 6/8 trials.	69
3.10	Statistics of <i>target</i> prediction.	69
3.11	The dev examples for which both models (FS-MTL, Z&X-STL) <i>incorrectly</i> predict the <i>holder</i> in 6/8 trials.	70
3.12	The dev examples for which both models (FS-MTL, Z&X-STL) <i>incorrectly</i> predict the <i>target</i> in 6/8 trials.	70
3.13	The dev examples for which the FS-MTL model <i>correctly</i> predicts the <i>holder</i> in 6/8 trials, whereas the Z&X-STL model predicts <i>incorrectly</i> in 6/8 trials.	72
3.14	The dev examples for which the FS-MTL model <i>correctly</i> predicts the <i>target</i> in 6/8 trials, whereas the Z&X-STL model predicts <i>incorrectly</i> in 6/8 trials.	72
4.1	The embedding sentence types (column 1), the head of SBAR (column 2), and anaphoric expressions they induce (column 3).	82
4.2	Data statistics: number of anaphors. AnaphS _{-d} denotes the sentence that occurs $d - 1$ positions (in number of intervening sentences) before the anaphoric sentence. The requirement "antecedent \in window" indicates that the shown number is the number of anaphors whose antecedent occurs in this window. The symbol ★ indicates that we exclude anaphors whose antecedent's syntactic tag label is not in the set {S, VP, ROOT, SBAR, SBARQ}. Difference between rows (1–3) and (4–6) shows how many anaphors can not be resolved with our models since their antecedent is not an S-like syntactic constituent.	87
4.3	Datasets used in different training data types: silver vs. gold vs. mixed.	88
4.4	Distance probabilities calculated from a balanced subset of the development parts of the ASN and CoNLL-12 corpora.	88
4.5	An overview of the evaluation configurations for the baselines and corresponding tables.	92
4.6	An overview of the evaluation configurations for our MR-LSTM models and corresponding tables.	93
4.7	Comparison between the training on the silver data to the training on the gold data.	93

4.8	The best training strategies between gold, silver, and mixed. The window $\{\text{AnaphS}_{-i}, \dots, \text{AnaphS}\}$ is denoted by W_i . The "mixed" in bold indicates the cases where the training on the mixture of the gold and silver datasets results in the best performance.	95
4.9	Comparison between the MR-LSTM with and without the VERB feature. The symbol ✓ denotes the cases when MR-LSTM benefits from the VERB training and ✗ when it does not.	96
4.10	Comparison between variants of the MR-LSTM on the test part of the ASN corpus (shell noun <i>decision</i>) and the reported result in Kolhatkar (2015). . .	97
4.11	The results of the baselines ($BL_{\text{dist,sent}}$, $BL_{\text{dist,tag}}$, BL_{tag}) in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$. See Section 4.5.1 on page 92 for other details.	99
4.12	MR-LSTM results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers that are marked in bold indicate the best result for the corresponding evaluation dataset across different training data types: gold, silver, mixed. See Section 4.5.2 on page 92 for other details.	100
4.13	MR-LSTM without the VERB feature results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in bold indicate the cases when the result obtained with the $MR-LSTM_{-VERB}$ model is better than the result obtained with the full MR-LSTM model. See Section 4.5.3 on page 95 for other details.	101
4.14	MR-LSTM without the DISTANCE feature results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$. See Section 4.5.3 on page 95 for other details.	102
4.15	The results of the baselines ($BL_{\text{dist,sent}}$, $BL_{\text{dist,tag}}$, BL_{tag}) in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_2=\{\text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. A number marked in bold indicates that any variant of our MR-LSTM model does not beat this baseline result. See Section 4.5.1 on page 92 for other details.	103

4.16	MR-LSTM results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_2 = \{AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$. Numbers that are marked in bold indicate the best result for the corresponding evaluation dataset across different training data types: gold, silver, mixed. The $BL_{dist,tag}$ results that are better than all MR-LSTM results for the corresponding evaluation dataset are <u>underlined</u> . See Section 4.5.2 on page 92 for other details.	104
4.17	MR-LSTM without the VERB feature results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_2 = \{AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$. Numbers marked in bold indicate the cases when the result obtained with the $MR-LSTM_{-VERB}$ model is better than the result obtained with the full MR-LSTM model. See Section 4.5.3 on page 95 for other details.	105
4.18	MR-LSTM without the DISTANCE feature results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_2 = \{AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$. See Section 4.5.3 on page 95 for other details.	106
4.19	The results of the baselines ($BL_{dist,sent}$, $BL_{dist,tag}$, BL_{tag}) in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_3 = \{AnaphS_{-3}, AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$. A number marked in bold indicates that any variant of our MR-LSTM model does not beat this baseline result. See Section 4.5.1 on page 92 for other details. . . .	107
4.20	MR-LSTM results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_3 = \{AnaphS_{-3}, AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$. Numbers that are marked in bold indicate the best result for the corresponding evaluation dataset across different training data types: gold, silver, mixed. Baseline results that are better than all MR-LSTM results for the corresponding evaluation dataset are <u>underlined</u> . See Section 4.5.2 on page 92 for other details.	108
4.21	MR-LSTM without the VERB feature results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_3 = \{AnaphS_{-3}, AnaphS_{-2}, AnaphS_{-1}, AnaphS\}$. Numbers marked in bold indicate the cases when the result obtained with the $MR-LSTM_{-VERB}$ model is better than the result obtained with the full MR-LSTM model. See Section 4.5.3 on page 95 for other details.	109

4.22	MR-LSTM without the DISTANCE feature results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_3=\{\text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. See Section 4.5.3 on page 95 for other details.	110
4.23	The results of the baselines ($\text{BL}_{\text{dist,sent}}$, $\text{BL}_{\text{dist,tag}}$, BL_{tag}) in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_4=\{\text{AnaphS}_{-4}, \text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. A number marked in bold indicates that any variant of our MR-LSTM model does not beat this baseline result. See Section 4.5.1 on page 92 for other details.	111
4.24	MR-LSTM results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_4=\{\text{AnaphS}_{-4}, \text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers that are marked in bold indicate the best result for the corresponding evaluation dataset across different training data types: gold, silver, mixed. Baseline results that are better than all MR-LSTM results for the corresponding evaluation dataset are <u>underlined</u> . See Section 4.5.2 on page 92 for other details.	112
4.25	MR-LSTM without the VERB feature results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_4=\{\text{Anaph}_{-4}, \text{AnaphS}_{-3}, \text{Anaph}_{-2}, \text{Anaph}_{-1}, \text{AnaphS}\}$. Numbers marked in bold indicate the cases when the result obtained with the $\text{MR-LSTM}_{\text{-VERB}}$ model is better than the result obtained with the full MR-LSTM model. See Section 4.5.3 on page 95 for other details.	113
4.26	MR-LSTM without the DISTANCE feature results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_4=\{\text{AnaphS}_{-4}, \text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. See Section 4.5.3 on page 95 for other details.	114
4.27	Comparison between the MR-LSTM with and without adversarial training. The symbol ✓ denotes the cases when adversarial training is beneficial and ✗ when it is not.	117
4.28	The best training strategies between gold, silver, mixed, and mixed with adversarial training. The window $\{\text{AnaphS}_{-i}, \dots, \text{AnaphS}\}$ is denoted by W_i . The "mix-adv" in bold indicates the cases where adversarial training results in the best performance between all training strategies.	117

4.29	MR-LSTM with adversarial training results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in bold indicate the cases when the adversarial training resulted in a better performance than gold, silver, or mixed training (Table 4.12 on page 100).	118
4.30	MR-LSTM with adversarial training results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_2=\{\text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in bold indicate the cases when the adversarial training resulted in a better performance than gold, silver, or mixed training (Table 4.16 on page 104).	118
4.31	MR-LSTM with adversarial training results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_3=\{\text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in bold indicate the cases when the adversarial training resulted in a better performance than gold, silver, or mixed training (Table 4.20 on page 108).	119
4.32	MR-LSTM with adversarial training results in resolution of anaphors that have an S-like antecedent that occurs in a sentences from the window $W_4=\{\text{AnaphS}_{-4}, \text{AnaphS}_{-3}, \text{AnaphS}_{-2}, \text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in bold indicate the cases when the adversarial training resulted in a better performance than gold, silver, or mixed training (Table 4.24 on page 112).	119
4.33	Comparison between the MR-LSTM with and without multi-task learning. The symbol  denotes the cases when MR-LSTM benefits from multi-task learning and  when it does not.	122
4.34	The best training strategies between all possibilities: gold, silver, mixed, mixed with adversarial training, and mixed with multi-task learning. The window $\{\text{AnaphS}_{-i}, \dots, \text{AnaphS}\}$ is denoted by W_i . The "mix-ntl" in bold indicates the cases where multi-task learning results in the best performance between all training strategies.	122
4.35	MR-LSTM with multi-task learning results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in bold indicate the cases when the multi-task learning resulted in a better performance than gold, silver, or mixed training (Table 4.12 on page 100), or adversarial training (Table 4.29 on page 118).	124

4.36	MR-LSTM with multi-task learning results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in bold indicate the cases when the multi-task learning resulted in a better performance than gold, silver, or mixed training (Table 4.16 on page 104), or adversarial training (Table 4.30 on page 118).	125
4.37	MR-LSTM with multi-task learning results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in bold indicate the cases when the multi-task learning resulted in a better performance than gold, silver, or mixed training (Table 4.20 on page 108), or adversarial training (Table 4.31 on page 119).	126
4.38	MR-LSTM with multi-task learning results in resolution of anaphors that have an S-like antecedent that occurs in one of the sentences in the window $W_1=\{\text{AnaphS}_{-1}, \text{AnaphS}\}$. Numbers marked in bold indicate the cases when the multi-task learning resulted in a better performance than gold, silver, or mixed training (Table 4.24 on page 112), or adversarial training (Table 4.32 on page 119).	127
A.1	Statistics of the ORL (MPQA) data for 4-fold CV.	140
A.2	Statistics of the ORL (MPQA) data for 10-fold CV.	141
B.1	Shell noun resolution results.	144
B.2	Architecture ablation for <i>reason</i>	144
C.1	Tables with manual evaluations for (up to) 10 randomly drawn examples for different embedding types (arguments and adjuncts) in Trial 1.	146
C.2	Tables with manual evaluations for (up to) 10 randomly drawn examples for different embedding types (arguments and adjuncts) in Trial 2.	146
C.3	The first trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>0</i> as the head of the SBAR clause.	148
C.4	The first trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>NONE</i> as the head of the SBAR clause.	150
C.5	The first trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>that</i> as the head of the SBAR clause.	152
C.6	The first trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>if</i> as the head of the SBAR clause.	154

C.7	The first trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>as</i> as the head of the SBAR clause.	156
C.8	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>0</i> as the head of the SBAR clause.	158
C.9	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>NONE</i> as the head of the SBAR clause.	160
C.10	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>whether</i> as the head of the SBAR clause.	162
C.11	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>because</i> as the head of the SBAR clause.	164
C.12	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>since-prp</i> as the head of the SBAR clause.	166
C.13	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>as-prp</i> as the head of the SBAR clause.	168
C.14	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>if</i> as the head of the SBAR clause.	170
C.15	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>after</i> as the head of the SBAR clause.	172
C.16	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>since-tmp</i> as the head of the SBAR clause.	174
C.17	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>until</i> as the head of the SBAR clause.	176
C.18	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>while</i> as the head of the SBAR clause.	178
C.19	Second trial of the manual evaluation of the data extracted with the VC-SS-Extract method and <i>as</i> as the head of the SBAR clause.	179

List of Abbreviations

AAR Abstract Anaphora Resolution	5, 8, 9, 24, 39, 42, 73, 115, 116
ASNs Anaphoric Shell Nouns	37, 39, 41
CNNs Convolutional Neural Networks	47, 49, 53
CRF Conditional Random Field	3, 4, 28, 29, 31, 32
CSNs Cataphoric Shell Nouns	37, 39, 40, 42
DSEs Direct Subjective Elements	26, 117
ELUs Exponential Linear Units	76, 77
ESEs Expressive Subjective Elements	26
FFNNs Feed-Forward Neural Networks	48
FGOA Fine-Grained Opinion Analysis	3, 4, 8, 33, 56
FS-MTL Fully Shared MTL model	53, 60, 64, 65, 70, 80
H-MTL Hierarchical MTL model	53, 60, 64, 65, 114
ILP Integer Linear Programming	12, 18–20, 24, 29, 30
LBP Loopy Belief Propagation	12–14, 16, 24
LSTM Long Short-Term Memory	49, 50, 53, 54, 60, 62–65, 76, 78, 80
MPQA Multi-Perspective Question Answering Corpus	1, 3, 4, 7–9, 17, 22, 25, 27–29, 32, 56–58, 61, 65, 70, 72, 117

MTL Multi-Task Learning .	4, 7–9, 46, 53, 56, 57, 59, 60, 62–66, 72, 92, 93, 97, 114, 115, 142
ORL Opinion Role Labeling	4, 7, 8, 56–58, 60–63, 65, 70, 72, 114–117
PSL Probabilistic Soft Logic	12, 20–24
RNNs Recurrent Neural Networks	47, 49, 53
SP-MTL Shared-Private MTL model	54, 60, 65, 114
SRL Semantic Role Labeling	4, 8, 9, 18, 29, 30, 56–62, 65, 70, 72, 114, 115

References

- Aditya, S., Yang, Y., Baral, C., and Aloimonos, Y. (2018). Combining Knowledge and Reasoning through Probabilistic Soft Logic for Image Puzzle Solving. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, Monterey, USA.
- Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649. Association for Computational Linguistics.
- Allen, J. and Heeman, P. (1995). TRAINS Spoken Dialog Corpus LDC95S25. *Linguistic Data Consortium*.
- Anand, P. and Hardt, D. (2016). Antecedent selection for sluicing: Structure and content. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1243. Association for Computational Linguistics.
- Anand, P. and McCloskey, J. (2015). Annotating the Implicit Content of Sluices. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 178–187. Association for Computational Linguistics.
- Angelidis, S. and Lapata, M. (2018). Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686. Association for Computational Linguistics.
- Artstein, R. and Poesio, M. (2006). Identifying Reference to Abstract Objects in Dialogue. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, pages 56–63.
- Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.
- Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. (2018). Relational Inductive Biases, Deep Learning, and Graph Networks. *arXiv preprint arXiv:1806.01261*.
- Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning Long-Term Dependencies With Gradient Descent Is Difficult. *IEEE transactions on neural networks*, 5(2):157–166.
- Bikel, D. M., Miller, S., Schwartz, R., and Weischedel, R. (1997). Nymble: a High-Performance Learning Name-finder. In *Fifth Conference on Applied Natural Language Processing*, pages 194–201, Washington, USA.

- Bingel, J. and Sjøgaard, A. (2017). Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.
- Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (2010). A High-Performance Syntactic and Semantic Dependency Parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 33–36. Association for Computational Linguistics.
- Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., and Reynar, J. (2008). Building a Sentiment Summarizer for Local Service Reviews. In *Proceedings of the WWW Workshop on NLP Challenges in the Information Explosion Era (NLPiX)*, volume 14, pages 339–348.
- Botley, S. P. (2006). Indirect Anaphora: Testing the Limits of Corpus-Based Linguistics. *International Journal of Corpus Linguistics*, 11(1):73–112.
- Bowman, S. R., Gauthier, J., Rastogi, A., Gupta, R., Manning, C. D., and Potts, C. (2016). A Fast Unified Model for Parsing and Sentence Understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1477, Berlin, Germany.
- Breck, E., Choi, Y., and Cardie, C. (2007). Identifying Expressions of Opinion in Context. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*, volume 7, pages 2683–2688.
- Byron, D. (2003). Annotation of Pronouns and Their Antecedents: A Comparison of Two Domains. Technical report, The University of Rochester, Computer Science Department.
- Byron, D. (2004). *Resolving Pronominal Reference to Abstract Entities*. PhD thesis, University of Rochester, Department of Computer Science.
- Cai, L. and Wang, W. Y. (2018). KBGAN: Adversarial Learning for Knowledge Graph Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1470–1480. Association for Computational Linguistics.
- Carreras, X. and Màrquez, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164. Association for Computational Linguistics.
- Caruana, R. (1997). Multitask Learning. *Machine learning*, 28(1):41–75.
- Chen, D. and Manning, C. (2014). A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750. Association for Computational Linguistics.

- Chen, X. and Cardie, C. (2018). Multinomial Adversarial Networks for Multi-Domain Text Classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1226–1240. Association for Computational Linguistics.
- Chen, X., Shi, Z., Qiu, X., and Huang, X. (2017). Adversarial Multi-Criteria Learning for Chinese Word Segmentation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1193–1203, Vancouver, Canada. Association for Computational Linguistics.
- Chen, Y., Xu, L., Liu, K., Zeng, D., and Zhao, J. (2015). Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176. Association for Computational Linguistics.
- Chen, Y.-C. and Bansal, M. (2018). Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686. Association for Computational Linguistics.
- Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111. Association for Computational Linguistics.
- Choi, E., Rashkin, H., Zettlemoyer, L., and Choi, Y. (2016). Document-level Sentiment Inference with Social, Faction, and Discourse Context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 333–343. Association for Computational Linguistics.
- Choi, Y., Breck, E., and Cardie, C. (2006). Joint Extraction of Entities and Relations for Opinion Recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Sydney, Australia.
- Choi, Y. and Cardie, C. (2010). Hierarchical Sequential Learning for Extracting Opinions and Their Attributes. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 269–274. Association for Computational Linguistics.
- Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Vancouver, Canada.
- Choi, Y., Deng, L., and Wiebe, J. (2014). Lexical Acquisition for Opinion Inference: A Sense-Level Lexicon of Benefactive and Malefactive Events. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 107–112. Association for Computational Linguistics.

- Choi, Y. and Wiebe, J. (2014). +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191. Association for Computational Linguistics.
- Choi, Y., Wiebe, J., and Mihalcea, R. (2017). Coarse-Grained +/-Effect Word Sense Disambiguation for Implicit Sentiment Analysis. *IEEE Transactions on Affective Computing*, 8(4):471–479.
- Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 539–546. IEEE.
- Clark, K. and Manning, C. D. (2015). Entity-Centric Coreference Resolution with Model Stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415. Association for Computational Linguistics.
- Clark, K. and Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653. Association for Computational Linguistics.
- Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2016). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico.
- Collins, E., Augenstein, I., and Riedel, S. (2017). A Supervised Approach to Extractive Summarisation of Scientific Papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 195–205. Association for Computational Linguistics.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural Language Processing (Almost) From Scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.
- Consten, M., Knees, M., and Schwarz-Friesel, M. (2007). The Function of Complex Anaphors in Texts. *Anaphors in Texts*, pages 81–102.
- Das, A., Yenala, H., Chinnakotla, M. K., and Shrivastava, M. (2016). Together We Stand: Siamese Networks for Similar Question Retrieval. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berling, Germany.
- Davis, W. (2014). Implicature. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2014 edition.
- Deng, L., Choi, Y., and Wiebe, J. (2013a). Benefactive/Malefactive Event and Writer Attitude Annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125. Association for Computational Linguistics.

- Deng, L., Hinton, G., and Kingsbury, B. (2013b). New Types Of Deep Neural Network Learning For Speech Recognition And Related Applications: An Overview. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8599–8603. IEEE.
- Deng, L. and Wiebe, J. (2014). Sentiment Propagation via Implicature Constraints. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 377–385. Association for Computational Linguistics.
- Deng, L. and Wiebe, J. (2015a). Joint Prediction for Entity/Event-Level Sentiment Analysis using Probabilistic Soft Logic Models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 179–189. Association for Computational Linguistics.
- Deng, L. and Wiebe, J. (2015b). MPQA 3.0: An Entity/Event-Level Sentiment Corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328. Association for Computational Linguistics.
- Deng, L., Wiebe, J., and Choi, Y. (2014). Joint Inference and Disambiguation of Implicit Sentiments via Implicature Constraints. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 79–88. Dublin City University and Association for Computational Linguistics.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.
- Dipper, S., Rieger, C., Seiss, M., and Zinsmeister, H. (2011). Abstract Anaphors in German and English. In *Discourse Anaphora and Anaphor Resolution Colloquium*, pages 96–107. Springer.
- Dipper, S. and Zinsmeister, H. (2009). Annotating Discourse Anaphora. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 166–169. Association for Computational Linguistics.
- Dipper, S. and Zinsmeister, H. (2012). Annotating Abstract Anaphora. *Language Resources and Evaluation*, 46(1):37–52.
- Dong, L., Wei, F., Zhou, M., and Xu, K. (2015). Question Answering over Freebase with Multi-Column Convolutional Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 260–269. Association for Computational Linguistics.
- Dozat, T. and Manning, C. D. (2017). Deep Biaffine Attention for Neural Dependency Parsing. In *Proceedings of the 15th International Conference on Learning Representations*.
- Duong, L., Cohn, T., Bird, S., and Cook, P. (2015). Low Resource Dependency Parsing: Cross-lingual Parameter Sharing in a Neural Network Parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850. Association for Computational Linguistics.

- Eckert, M. and Strube, M. (2000). Dialogue Acts, Synchronizing Units, and Anaphora Resolution. *Journal of Semantics*, 17(1):51–89.
- Elman, J. L. (1990). Finding Structure In Time. *Cognitive science*, 14(2):179–211.
- Falke, T., Meyer, C. M., and Gurevych, I. (2017). Concept-Map-Based Multi-Document Summarization using Concept Coreference Resolution and Global Importance Optimization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 801–811. Asian Federation of Natural Language Processing.
- Feng, S., Kang, J. S., Kuznetsova, P., and Choi, Y. (2013). Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774–1784. Association for Computational Linguistics.
- Filippova, K., Alfonseca, E., Colmenares, C. A., Kaiser, L., and Vinyals, O. (2015). Sentence Compression by Deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368. Association for Computational Linguistics.
- Fillmore, C. J., Johnson, C. R., and Petruck, M. R. (2003). Background to FrameNet. *International journal of lexicography*, 16(3):235–250.
- Gal, Y. and Ghahramani, Z. J. C. (2016). A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems (NIPS)*.
- Ganin, Y. and Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. In *International Conference on Machine Learning (ICML)*, pages 1180–1189.
- Gehring, J., Auli, M., Grangier, D., and Dauphin, Y. (2017). A Convolutional Encoder Model for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135. Association for Computational Linguistics.
- Goldberg, Y. (2017). Neural Network Methods For Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.
- Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep Learning*, volume 1. MIT press Cambridge.
- Graves, A. and Schmidhuber, J. (2005). Framewise Phoneme Classification With Bidirectional LSTM And Other Neural Network Architectures. *Neural Networks*, 18:602–610.
- Gui, T., Zhang, Q., Huang, H., Peng, M., and Huang, X. (2017). Part-of-Speech Tagging for Twitter with Adversarial Neural Networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2401–2410, Copenhagen, Denmark. Association for Computational Linguistics.

- Gundel, J. K., Hedberg, N., and Zacharski, R. (2004). Demonstrative Pronouns in Natural Discourse. In *Proceedings of the 5th Discourse Anaphora and Anaphora Resolution Colloquium (DAARC)*, pages 81–86, Sao Miguel, Portugal.
- Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015). Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16. Association for Computational Linguistics.
- Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.
- Hashimoto, K., xiong, c., Tsuruoka, Y., and Socher, R. (2017). A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933. Association for Computational Linguistics.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1026–1034.
- He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep Semantic Role Labeling: What Works and What’s Next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483. Association for Computational Linguistics.
- Henaff, M., Szlam, A., and LeCun, Y. (2016). Recurrent Orthogonal Networks and Long-Memory Tasks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2034–2042.
- Hochreiter, S. (1991). Untersuchungen Zu Dynamischen Neuronalen Netzen. *Diploma, Technische Universität München*, 91(1).
- Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural computation*, 9(8):1735–1780.
- Holtzman, A., Buys, J., Forbes, M., Bosselut, A., Golub, D., and Choi, Y. (2018). Learning to Write with Cooperative Discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649. Association for Computational Linguistics.
- Hou, Y., Markert, K., and Strube, M. (2018). Unrestricted Bridging Resolution. *Computational Linguistics*, 44(2):237–284.
- Hovy, E. (2011). Invited Keynote: What are Subjectivity, Sentiment, and Affect? In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, page 1. Asian Federation of Natural Language Processing.
- Howard, J. and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics.

- Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168–177, Seattle, USA. ACM.
- Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv:1508.01991*.
- Huddleston, R. and Pullum, G. (2002). *The Cambridge grammar of the English language*. Cambridge University Press.
- Irsoy, O. and Cardie, C. (2014a). Deep Recursive Neural Networks for Compositionality in Language. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2096–2104.
- Irsoy, O. and Cardie, C. (2014b). Opinion Mining with Deep Recurrent Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 720–728. Association for Computational Linguistics.
- Jacob, N. and Gurevych, I. (2010). Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045. Association for Computational Linguistics.
- Jauhar, S. K., Guerra, R., González Pellicer, E., and Recasens, M. (2015). Resolving Discourse-Deictic Pronouns: A Two-Stage Approach to Do It. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 299–308, Denver, Colorado.
- Johansson, R. and Moschitti, A. (2013). Relational Features in Fine-grained Opinion Analysis. *Computational Linguistics*, 39(3):473–509.
- Joty, S., Nakov, P., Màrquez, L., and Jaradat, I. (2017). Cross-language Learning with Adversarial Neural Networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 226–237, Vancouver, Canada. Association for Computational Linguistics.
- Jozefowicz, R., Zaremba, W., and Sutskever, I. (2015). An Empirical Exploration of Recurrent Network Architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 2342–2350.
- Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A Convolutional Neural Network For Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 655–665. Association for Computational Linguistics.
- Kang, D., Khot, T., Sabharwal, A., and Hovy, E. (2018). AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2418–2428. Association for Computational Linguistics.

- Katiyar, A. and Cardie, C. (2016). Investigating LSTMs for Joint Extraction of Opinion Entities and Relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929. Association for Computational Linguistics.
- Kim, S.-M. and Hovy, E. (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8. Association for Computational Linguistics.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.
- Kim, Y.-B., Stratos, K., and Kim, D. (2017). Adversarial Adaptation of Synthetic or Stale Data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1297–1307, Vancouver, Canada. Association for Computational Linguistics.
- Kingma, D. and Ba, J. (2015). Adam: A method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego.
- Klenner, M. (2015). Verb-centered Sentiment Inference with Description Logics. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 134–139. Association for Computational Linguistics.
- Klenner, M. and Amsler, M. (2016). Sentiframes: A Resource for Verb-centered German Sentiment Inference. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*.
- Kolhatkar, V. (2015). *Resolving Shell Nouns*. PhD thesis, University of Toronto.
- Kolhatkar, V., Roussel, A., Dipper, S., and Zinsmeister, H. (2018). Anaphora with non-nominal antecedents in computational linguistics: a survey. *Computational Linguistics*, 44(3):547–612.
- Kolhatkar, V., Zinsmeister, H., and Hirst, G. (2013a). Annotating Anaphoric Shell Nouns with their Antecedents. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 112–121. Association for Computational Linguistics.
- Kolhatkar, V., Zinsmeister, H., and Hirst, G. (2013b). Interpreting Anaphoric Shell Nouns using Antecedents of Cataphoric Shell Nouns as Training Data. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 300–310, Seattle, Washington, USA.
- Kurita, S., Kawahara, D., and Kurohashi, S. (2018). Neural Adversarial Training for Semi-supervised Japanese Predicate-argument Structure Analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 474–484. Association for Computational Linguistics.
- Lazarsfeld, P. F. and Merton, R. K. (1954). Friendship As a Social Process: A Substantive and Methodological Analysis. *Freedom and control in modern society*, 18(1):18–66.

- Le Nagard, R. and Koehn, P. (2010). Aiding Pronoun Translation with Co-Reference Resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261. Association for Computational Linguistics.
- Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics*, 39(4).
- Lee, K., He, L., Lewis, M., and Zettlemoyer, L. S. (2017). End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 188–197.
- Lerman, K., Blair-Goldensohn, S., and McDonald, R. (2009). Sentiment Summarization: Evaluating and Learning User Preferences. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, pages 514–522. Association for Computational Linguistics.
- Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D. (2017). Adversarial Learning for Neural Dialogue Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2147–2159, Copenhagen, Denmark. Association for Computational Linguistics.
- Liu, B. (2015). *Sentiment Analysis: Mining Sentiments, Opinions, and Emotions*. Cambridge University Press.
- Liu, K., Xu, H. L., Liu, Y., and Zhao, J. (2013a). Opinion Target Extraction Using Partially-Supervised Word Alignment Model. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2134–2140, Beijing, China.
- Liu, K., Xu, L., and Zhao, J. (2013b). Syntactic Patterns versus Word Alignment: Extracting Opinion Targets from Online Reviews. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1754–1763. Association for Computational Linguistics.
- Liu, K., Xu, L., and Zhao, J. (2014). Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 314–324. Association for Computational Linguistics.
- Liu, P., Joty, S., and Meng, H. (2015). Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443. Association for Computational Linguistics.
- Liu, P., Qiu, X., and Huang, X. (2017). Adversarial Multi-task Learning for Text Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.
- Lu, J. and Ng, V. (2017). Joint Learning for Event Coreference Resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 90–101. Association for Computational Linguistics.

- Lu, Y., Zhai, C., and Sundaresan, N. (2009). Rated Aspect Summarization of Short Comments. In *Proceedings of the 18th International Conference on World Wide Web*, pages 131–140, Madrid, Spain.
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60. Association for Computational Linguistics.
- Marasović, A., Born, L., Opitz, J., and Frank, A. (2017). A mention-ranking model for abstract anaphora resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 221–232. Association for Computational Linguistics.
- Marasović, A. and Frank, A. (2016). Multilingual Modal Sense Classification using a Convolutional Neural Network. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 111–120. Association for Computational Linguistics.
- Marasović, A. and Frank, A. (2018). SRL4ORL: Improving Opinion Role Labeling Using Multi-Task Learning with Semantic Role Labeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 583–594. Association for Computational Linguistics.
- Marasović, A., Zhou, M., Palmer, A., and Frank, A. (2016). Modal Sense Classification At Large: Paraphrase-Driven Sense Projection, Semantically Enriched Classification Models and Cross-Genre Evaluations. In *Linguistic Issues in Language Technology, Special issue on Modality in Natural Language Understanding*, volume 14 (2), Stanford, CA. CSLI Publications.
- Marcheggiani, D. and Titov, I. (2017). Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515. Association for Computational Linguistics.
- Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.
- Massey Jr, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American statistical Association*, 46(253):68–78.
- Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. (2007). Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180, Banff, Canada. ACM.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation Of Word Representations In Vector Space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751. Association for Computational Linguistics.

- Minervini, P., Demeester, T., Rocktäschel, T., and Riedel, S. (2017). Adversarial Sets for Regularising Neural Link Predictors. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*.
- Mitkov, R. (2014). *Anaphora Resolution*. Routledge. pp. 4–5.
- Moosavi, N. S. and Strube, M. (2018). Using Linguistic Features to Improve the Generalization Capability of Neural Coreference Resolvers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 193–203. Association for Computational Linguistics.
- Mrkšić, N., Ó Séaghdha, D., Wen, T.-H., Thomson, B., and Young, S. (2017). Neural Belief Tracker: Data-Driven Dialogue State Tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788. Association for Computational Linguistics.
- Mueller, J. and Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the 13th Conference on Artificial Intelligence (AAAI)*, pages 2786–2792, Phoenix, Arizona.
- Müller, C. (2007). Resolving It, This, and That in Unrestricted Multi-Party Dialog. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 816–823. Association for Computational Linguistics.
- Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada. Association for Computational Linguistics.
- Navarretta, C. and Olsen, S. (2008). Annotating Abstract Pronominal Anaphora in the DAD Project. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*.
- Neculoiu, P., Versteegh, M., and Rotaru, M. (2016). Learning Text Similarity with Siamese Recurrent Networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, Berlin, Germany.
- Nie, A., Bennett, E. D., and Goodman, N. D. (2017). DisSent: Sentence Representation Learning from Explicit Discourse Relations. *arXiv preprint arXiv:1710.04334*.
- Orăsan, C. (2007). Pronominal Anaphora Resolution for Text Summarisation. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 430–436.
- Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1).
- Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.
- Pappuswamy, U., Jordan, P. W., and VanLehn, K. (2005). Resolving discourse deictic anaphors in tutorial dialogues. *Constraints in Discourse*, 103:95–102.

- Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the Difficulty of Training Recurrent Neural Networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1310–1318.
- Pearl, J. (1982). Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach. In *Proceedings of the American Association of Artificial Intelligence National Conference on AI*, pages 133–136, Pittsburgh, PA, USA.
- Peng, H., Thomson, S., Swayamdipta, S., and Smith, N. A. (2018). Learning joint semantic parsers from disjoint data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1492–1502. Association for Computational Linguistics.
- Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018a). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.
- Peters, M., Neumann, M., Zettlemoyer, L., and Yih, W.-t. (2018b). Dissecting Contextual Word Embeddings: Architecture and Representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509. Association for Computational Linguistics.
- Pham, N.-Q., Kruszewski, G., and Boleda, G. (2016). Convolutional Neural Network Language Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1153–1162. Association for Computational Linguistics.
- Poesio, M. and Artstein, R. (2008). Anaphoric Annotation in the ARRAU Corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*.
- Poesio, M., Grishina, Y., Kolhatkar, V., Moosavi, N., Roesiger, I., Roussel, A., Simonjetz, F., Uma, A., Uryupina, O., Yu, J., and Zinsmeister, H. (2018). Anaphora Resolution with the ARRAU Corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22. Association for Computational Linguistics.
- Popescu, A.-M. and Etzioni, O. (2005). Extracting Product Features and Opinions from Reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.
- Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

- Pradhan, S. S., Ramshaw, L., Weischedel, R., MacBride, J., and Micciulla, L. (2007). Unrestricted Coreference: Identifying Entities and Events in OntoNotes. In *Proceedings of the International Conference on Semantic Computing, ICSC '07*, pages 446–453, Washington, DC, USA. IEEE Computer Society.
- Qin, L., Zhang, Z., Zhao, H., Hu, Z., and Xing, E. (2017). Adversarial Connective-exploiting Networks for Implicit Discourse Relation Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.
- Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.
- Quarteroni, S. (2007). *Advanced Techniques for Personalized, Interactive Question Answering*. PhD thesis, University of York.
- Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London and New York: Longman.
- Rashkin, H., Singh, S., and Choi, Y. (2016). Connotation Frames: A Data-Driven Investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321. Association for Computational Linguistics.
- Recasens, M. (2008). Discourse Deixis and Coreference: Evidence from AnCora. In *Proceedings of the Second Workshop on Anaphora Resolution*, pages 73–82, Bergen, Norway.
- Recasens, M., de Marneffe, M.-C., and Potts, C. (2013). The Life and Death of Discourse Entities: Identifying Singleton Mentions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 627–633. Association for Computational Linguistics.
- Reimers, N. and Gurevych, I. (2017). Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.
- Riloff, E. (1996). An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains. *Artificial Intelligence*, 85:101–134.
- Rønning, O., Hardt, D., and Søgaard, A. (2018a). Linguistic Representations in Multi-Task Neural Networks for Ellipsis Resolution. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 66–73. Association for Computational Linguistics.
- Rønning, O., Hardt, D., and Søgaard, A. (2018b). Sluice Resolution without Hand-Crafted Features over Brittle Syntax Trees. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 236–241. Association for Computational Linguistics.

- Roth, D. and Yih, W.-T. (2004). A Linear Programming Formulation for Global Inference in Natural Language Tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, Boston, USA.
- Ruder, S. (2017). An Overview Of Multi-Task Learning In Deep Neural Networks. *arXiv preprint arXiv:1706.05098*.
- Ruder, S., Vulić, I., and Søgaard, A. (2017). A Survey of Cross-Lingual Word Embedding Models. *The Journal of Artificial Intelligence Research*.
- Ruppenhofer, J. and Brandes, J. (2016). Effect functors for opinion inference. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 2879 – 2887, Paris.
- Ruppenhofer, J., Somasundaran, S., and Wiebe, J. (2008). Finding the Sources and Targets of Subjective Expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 2781–2788, Marrakech, Morocco.
- Sandhaus, E. (2008). The New York Times Annotated Corpus. *Linguistic Data Consortium*, 6(12).
- Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*, volume 34. Walter de Gruyter.
- Semeniuta, S., Severyn, A., and Barth, E. (2017). A Hybrid Convolutional Variational Autoencoder for Text Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 627–637. Association for Computational Linguistics.
- Shen, H. and Sarkar, A. (2005). Voting Between Multiple Data Representations for Text Chunking. In *Proceedings of the Conference of the Canadian Society for Computational Studies of Intelligence*, pages 389–400. Springer.
- Shnarch, E., Alzate, C., Dankin, L., Gleize, M., Hou, Y., Choshen, L., Aharonov, R., and Slonim, N. (2018). Will it Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605. Association for Computational Linguistics.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics.
- Søgaard, A. and Goldberg, Y. (2016). Deep Multi-Task Learning With Low Level Tasks Supervised At Lower Layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235. Association for Computational Linguistics.
- Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of machine learning research*, 15(1):1929–1958.

- Steinberger, J., Kabadjov, M., Poesio, M., and Sanchez-Graillet, O. (2005). Improving LSA-based Summarization with Anaphora Resolution. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.
- Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.
- Stoyanov, V. and Cardie, C. (2008). Topic Identification for Fine-Grained Opinion Analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 817–824. Coling 2008 Organizing Committee.
- Stoyanov, V. and Cardie, C. (2011). Automatically Creating General-Purpose Opinion Summaries from Text. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 202–209. Association for Computational Linguistics.
- Strube, M., Rapp, S., and Müller, C. (2002). The Influence of Minimum Edit Distance on Reference Resolution. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL)*, Beijing, China.
- Tang, D., Qin, B., and Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432. Association for Computational Linguistics.
- Thorne, J. and Vlachos, A. (2018). Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359. Association for Computational Linguistics.
- Tiedemann, J. (2009). News from OPUS-A collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 237–248.
- Uryupina, O., Artstein, R., Bristot, A., Cavicchio, F., Rodriguez, K. J., and Poesio, M. (2016). ARRAU: Linguistically-Motivated Annotation of Anaphoric Descriptions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 2058–2062, Portoroz.
- Vicedo, J. L. and Ferrández, A. (2008). Coreference in Q & A. In *Advances in Open Domain Question Answering*, pages 71–96. Springer.
- Vieira, R., Salmon-Alt, S., Gasperin, C., Schang, E., and Othero, G. (2005). Coreference and Anaphoric Relations of Demonstrative Noun Phrases in Multilingual Corpus. *Anaphora Processing: Linguistic, Cognitive and Computational Modeling*, pages 385–403.

- Wang, L. and Ling, W. (2016). Neural Network-Based Abstract Generation for Opinions and Arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57. Association for Computational Linguistics.
- Webber, B. L. (1988). Discourse Deixis: Reference to Discourse Segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*.
- Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language resources and evaluation*, 39(2-3):165–210.
- Wiegand, M., Bocionek, C., and Ruppenhofer, J. (2016a). Opinion Holder and Target Extraction on Opinion Compounds—A Linguistic Approach. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 800–810. Association for Computational Linguistics.
- Wiegand, M. and Ruppenhofer, J. (2015). Opinion Holder and Target Extraction based on the Induction of Verbal Categories. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 215–225. Association for Computational Linguistics.
- Wiegand, M., Schulder, M., and Ruppenhofer, J. (2015). Opinion Holder and Target Extraction for Verb-based Opinion Predicates—The Problem is Not Solved. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 148–155. Association for Computational Linguistics.
- Wiegand, M., Schulder, M., and Ruppenhofer, J. (2016b). Separating Actor-View from Speaker-View Opinion Expressions using Linguistic Features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 778–788. Association for Computational Linguistics.
- Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.
- Wilson, T. A. (2008). *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. PhD thesis, University of Pittsburgh.
- Wiseman, S. J., Rush, A. M., Shieber, S. M., and Weston, J. (2015). Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL)*, Beijing, China.
- Wu, Y., Bamman, D., and Russell, S. (2017). Adversarial Training for Relation Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1779–1784, Copenhagen, Denmark. Association for Computational Linguistics.

- Xue, W. and Li, T. (2018). Aspect Based Sentiment Analysis with Gated Convolutional Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523. Association for Computational Linguistics.
- Yang, B. and Cardie, C. (2012). Extracting Opinion Expressions with semi-Markov Conditional Random Fields. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1335–1345. Association for Computational Linguistics.
- Yang, B. and Cardie, C. (2013). Joint Inference for Fine-grained Opinion Extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649. Association for Computational Linguistics.
- Yang, B. and Cardie, C. (2014a). Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 325–335. Association for Computational Linguistics.
- Yang, B. and Cardie, C. (2014b). Joint Modeling of Opinion Expression Extraction and Attribute Classification. *Transactions of the Association for Computational Linguistics*, 2:505–516.
- Yang, Y. and Hospedales, T. M. (2017). Trace Norm Regularised Deep Multi-task Learning. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Yasunaga, M., Kasai, J., and Radev, D. (2018). Robust Multilingual Part-of-Speech Tagging via Adversarial Training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 976–986. Association for Computational Linguistics.
- Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005). Constructing Free Energy Approximations and Generalized Belief Propagation algorithms. *IEEE Transactions on information theory*, 51(7):2282–2312.
- Yih, W.-t., He, X., and Meek, C. (2014). Semantic Parsing for Single-Relation Question Answering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 643–648. Association for Computational Linguistics.
- Yu, X. and Vu, N. T. (2017). Character Composition Model with Convolutional Neural Networks for Dependency Parsing on Morphologically Rich Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–678. Association for Computational Linguistics.
- Zhang, M., Liu, Y., Luan, H., and Sun, M. (2017). Adversarial Training for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

- Zhang, Y. and Wallace, B. (2017). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 253–263. Asian Federation of Natural Language Processing.
- Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014). Facial Landmark Detection By Deep Multi-Task Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 94–108. Springer.
- Zhao, J., Liu, K., and Wang, G. (2008). Adding Redundant Features for CRFs-based Sentence Sentiment Classification. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 117–126. Association for Computational Linguistics.
- Zhou, J. and Xu, W. (2015). End-To-End Learning Of Semantic Role Labeling Using Recurrent Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1127–1137. Association for Computational Linguistics.
- Zopf, M., Botschen, T., Falke, T., Marasović, A., Mihaylov, T., P.V.S., A., Mencía, E. L., Fürnkranz, J., and Frank, A. (2018). What's Important in a Text? An Extensive Evaluation of Linguistic Annotations for Summarization. In *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security*.

