Incremental Coreference Resolution for German

Thesis
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of Doctor of Philosophy

by

Don Tuggener

Accepted in the spring semester 2016
on the recommendation of the doctoral committee:

Prof. Dr. Martin Volk (main advisor)

PD Dr. Gerold Schneider

Zurich, 2016

"Choose a job you love, and you will never have to work a day in your life."

Confucius

Abstract

The main contributions of this thesis are as follows:

1. We introduce a general model for coreference and explore its application to German.
    - The model features an incremental discourse processing algorithm which allows it to coherently address issues caused by underspecification of mentions, which is an especially pressing problem regarding certain German pronouns.
    - We introduce novel features relevant for the resolution of German pronouns. A subset of these features are made accessible through the incremental architecture of the discourse processing model.
    - In evaluation, we show that the coreference model combined with our features provides new state-of-the-art results for coreference and pronoun resolution for German.
2. We elaborate on the evaluation of coreference and pronoun resolution.
    - We discuss evaluation from the view of prospective downstream applications that benefit from coreference resolution as a preprocessing component. Addressing the shortcomings of the general evaluation framework in this regard, we introduce an alternative framework, the Application Related Coreference Scores (ARCS).
    - The ARCS framework enables a thorough comparison of different system outputs and the quantification of their similarities and differences beyond the common coreference evaluation. We demonstrate how the framework is applied to state-of-the-art coreference systems. This provides a method to track specific differences in system outputs, which assists researchers in comparing their approaches to related work in detail.
3. We explore semantics for pronoun resolution.
    - Within the introduced coreference model, we explore distributional approaches to estimate the compatibility of an antecedent candidate and the occurrence context of a pronoun. We compare a state-of-the-art approach for word embeddings to syntactic co-occurrence profiles to this end.
    - In comparison to related work, we extend the notion of context and thereby increase the applicability of our approach. We find that a combination of both compatibility models, coupled with the coreference model, provides a large potential for improving pronoun resolution performance.

We make available all our resources, including a web demo of the system, at:

`http://pub.cl.uzh.ch/purl/coreference-resolution`

Abstract

Die wichtigsten Beiträge der vorliegenden Arbeit sind folgende:

1. Die Arbeit führt ein generelles Modell zur Koreferenzauflösung ein und exploriert dessen Anwendung auf die deutsche Sprache.
    - Das Modell verfügt über einen inkrementellen Algorithmus zur Diskursverarbeitung, der es erlaubt, Unterspezifizierung von Erwähnung von Entitäten auf kohärente Weise zu behandeln, was ein besonderes Problem beim Verarbeiten von deutschen Pronomen darstellt.
    - Ein Set an neuen Merkmalen für die Auflösung von deutschen Pronomen wird eingeführt. Ein Teil dieser Merkmale wird durch die inkrementelle Architektur des Algorithmus zur Diskursverarbeitung zugänglich.
    - In der Evaluation wird gezeigt, dass das Koreferenzmodell, gekoppelt mit den neuen Merkmalen, neue *state-of-the-art*-Resultate für Koreferenzauflösung für das Deutsche erreicht.
2. Die Arbeit behandelt die Evaluation von Koreferenz- und Pronomenauflösung.
    - Die gängige Evaluation wird aus der Perspektive von Anwendungen beleuchtet, die von Koreferenzauflösung als Vorverarbeitungsschritt profitieren. Ein alternativer Evaluationsansatz (ARCS) wird vorgeschlagen, der Defizite in der gängigen Evaluation aufnimmt.
    - Der vorgeschlagene Evaluationsansatz ermöglicht einen eingehenden Vergleich von Systemen und quantifiziert Gemeinsamkeiten und Unterschiede von Systemausgaben in einer Weise, die tiefer greift als die gängige Evaluation. Dieser Ansatz ermöglicht es Forschenden, ihre Systeme detailliert und unter verschiedenen Gesichtspunkten mit anderen zu vergleichen.
3. Die Arbeit untersucht den Einbezug von Semantik in die Auflösung von Pronomen.
    - Innerhalb des eingeführten Koreferenzmodells wird distributionelle Semantik als Ansatz zur Bestimmung der Kompatibilität eines Antezedenskandidaten und dem Kontext eines Pronomens exploriert. Zu diesem Zweck wird ein *state-of-the-art*-Ansatz zur Berechnung von Vektorrepräsentationen von Wörtern mit syntaktischen Kookurrenzprofilen von Wörtern verglichen.
    - Im Vergleich zu verwandten Arbeiten wird die Definition des Pronomenkontexts erweitert und dadurch die Anwendbarkeit des Ansatzes erhöht. Die Evaluation zeigt, dass die Kombination beider Kompatibilitätsmodelle, gekoppelt mit dem eingeführten Koreferenzmodell, ein hohes Potenzial für die Verbesserung der Resultate bezüglich Pronomenauflösung bietet.

Die Ressourcen, die ihm Rahmen der Arbeit erstellt wurden, inklusive einer Web-Demo des Systems, sind zugänglich unter:

`http://pub.cl.uzh.ch/purl/coreference-resolution`

Acknowledgements

Being able to do whatever you want must be considered a privilege. At the Institute of Computational Linguistics of the University of Zurich, I felt that I was granted this privilege while I investigated the subject of this dissertation. For entrusting me with this privilege, I am entirely grateful to Dr. Manfred Klenner, the main supervisor of this thesis, and Prof. Dr. Martin Volk, the head of the institute. I am indebted to Manfred Klenner for supporting my research from the get-go, which started with an innocent little PROLOG programming project for pronoun resolution and has led to the completion of this thesis. His input and feedback were invaluable to this thesis. Martin Volk, by seeing and connecting the dots at the right moment, provided the financial (but also moral) support for this thesis. His motion to support collaborations of students and senior members of the institute made this thesis possible.

Being able to do whatever you want also has a sort of negative connotation, and perhaps rightfully so. This thesis was mainly written detached of any major project. This comes with the advantages of freedom of project deadlines and milestones, and independence from other project members. It also comes with the disadvantages of freedom of project deadlines and milestones, and independence from other project members - it requires a considerable amount of self-discipline and self-motivation. I am grateful to my family, especially my wife Tamara and our two PhD production babies, Siro and Mona, for keeping me motivated and determined. With that in mind, I dedicate this thesis to them.

I would also like to thank everyone at the Institute of Computational Linguistics at the University of Zurich for the companionship while researching and writing this thesis. A special thanks goes to Dr. Simon Clematide, who introduced me to *wapiti*, and Johannes Graën who kept the institute servers running.

Finally, I thank PD Dr. Gerold Schneider and Prof. Dr. Manfred Stede who agreed to review this thesis and who accepted my invitation into the PhD committee.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the coreference phenomenon and discusses our motivation and goals for this thesis. We present a linguistic account of coreference and discuss how this account is transferred into gold standard annotations for research in Computational Linguistics. We outline the common tasks that coreference resolution systems have to perform in a shared task setting, which is the standard platform for research in coreference resolution in Computational Linguistics.

1.1 Problem description

Coreference is a natural language phenomenon that occurs when different, potentially varying linguistic forms are used in a discourse to refer to the same extra-linguistic entity. For example, a discourse might introduce the person Barack Obama by the name, i.e. “Barack Obama”, but for the next mention it will manifest a nominal description, like “the first black president in the history of the U.S.”, and subsequently simply a pronoun, i.e. “he”. That is, entities are generally introduced into discourse by a nominal description that includes a common noun or a name. But when they occur subsequently, the linguistic form that is used to mention them is often changed.

This poses a problem for many applications in Computational Linguistics (CL) and Natural Language Processing (NLP). Applications that identify entity occurrences purely based on their nominal descriptions will miss occurrences that deviate from this initial form. Coreference resolution aims at identifying and mapping these different linguistic forms to unique identifiers so that the occurrences of each entity in a discourse can be tracked.

Coreference resolution is a well-established research area in Computational Linguistics. Since the Message Understanding Conferences (MUC) in the 90ties, several shared tasks have tackled the problem, and to date, all major CL conferences have featured publications on the topic. While most work in the field evolves around the English language, two shared tasks in recent years have also included other languages, such as Spanish, Arabic, Chinese, Dutch, Catalan, Italian, and German. However, these shared tasks aimed at finding coreference models that perform well on all the languages, with minimal effort put into adapting the models to the specifics of each language.

By contrast, this thesis focuses on coreference resolution for the German language and pays attention to its specifics w.r.t. pronouns. Compared to e.g. English, two important aspects in this regard are i) certain German pronouns are underspecified regarding their morphological properties, and ii) certain German pronouns can be used to refer to both animate and inanimate entities. This thesis proposes an approach that addresses the issues that arise from these differences.

1.2 Research questions and hypotheses

We outline three hypotheses to be investigated throughout this thesis.

Underspecification of German pronouns. In the German language, certain pronouns are morphologically underspecified when viewed in isolation. Their morphological properties only become evident when used in discourse, i.e. when their referents are identified. This arguably renders the task of automated resolution of German pronouns more difficult than in English, since more candidates have to be considered as potential referents. Given the morphological underspecification of these pronouns, filtering based on morphological agreement licenses a larger number of potential referents. In turn, a larger number of potential referents increases the chance of picking an incorrect one.

Furthermore, if the morphological properties of resolved (and, thereby disambiguated) pronouns are not kept track of, subsequent coreference decisions involving these pronouns are bound to yield conflicting interpretations of their morphological properties. These conflicting interpretations, in turn, yield coreference chains with inconsistent morphological features.

For example, the German possessive pronoun *sein* is used to refer to entities with either neutral gender (corresponding to the English pronoun *its*) or masculine gender (corresponding to *his*). A coreference model that processes pairs of antecedents and pronouns in isolation might resolve an instance of *sein* to an entity with neutral gender, e.g. *Berlin*, as in the sentence “Berlin feiert sein Jubiläum” (Berlin celebrates its anniversary). The model might also resolve a subsequent, masculine personal pronoun, i.e. *er (he)* to the possessive pronoun *sein*, because the morphological properties of *sein* are underspecified when viewed in isolation, and the two pronouns are compatible in this view. When these two independently formed pairs of antecedent and anaphor are merged to form a coreference chain, it features incoherent morphological properties, i.e. {*Berlin - sein - er*}.

Related work on German pronoun and coreference resolution (almost) exclusively features such a pair-wise model. This thesis thus sets out to devise a coreference model for German which remedies these conflicting interpretations of underspecified pronouns. Our main hypothesis states that propagating taken coreference decisions to subsequent coreference decisions remedies the problem of arising inconsistencies within coreference chains. Furthermore, we hypothesise that ensuring consistency in coreference chains improves coreference resolution performance, and pronoun resolution for German in particular.

Evaluation of coreference and pronoun resolution. The commonly used framework for coreference evaluation models the linguistic mentions of entities as generic items. Coreference chains are interpreted as unordered sets of these items. Evaluation then compares the mention clustering in a system output to the clustering in a gold standard.

Since the mentions no longer have any distinguishable linguistic properties and the linear order of their appearance in discourse is lost, the metrics are hard to interpret for downstream applications that benefit from coreference resolution. For example, the metrics cannot answer the question how well a system resolves pronouns compared to another system, or how well a system links pronouns to their nominal antecedents. Furthermore, the metrics do not provide an interpretable explanation to downstream applications for the better performance of a system when compared to a system with lower performance.

We hypothesise that keeping the linguistic properties of the mentions and the linear order of their appearance in discourse yields a metric that answers these questions and which is interpretable for downstream applications and can be adjusted to different requirements that downstream applications have towards coreference resolution. We propose such a metric and present how it is applied to provide answers to the questions above.

Semantics for pronoun resolution. Most approaches to pronoun resolution rely on features that encode the salience of entities to determine an antecedent for a pronoun. While these approaches are successful, one outstanding question is whether semantics can improve them. Related work so far has produced mixed answers.

Incorporating semantics into pronoun resolution in related work signified that the selectional preferences of the verb governing a pronoun were considered w.r.t. the antecedent entities. The underlying rationale was that an antecedent has to be compatible with the selectional preference of the verb governing the pronoun. However, the mixed results in related work suggests that selectional preferences of verbs are a poor means to incorporate semantics into pronoun resolution.

We hypothesise that the utility of selectional preferences can be improved if the syntactic co-arguments of the pronouns are taken into account when modeling contexts of pronouns. We argue that not all verbs feature a selection of arguments that is narrow enough to derive clear preference towards a specific antecedent candidate. Including the syntactic co-argument of the pronoun should thus help to narrow down the selection preference of the pronoun's context.

1.3 Thesis outline

Chapter 1 gives a brief linguistic introduction to the phenomena of coreference and anaphora and outlines the potential benefit of coreference and pronoun resolution from the view of downstream applications. The common steps in a coreference resolution pipeline and the accompanying nomenclature are introduced.

Chapter 2 introduces the discourse model commonly used in coreference resolution, the mention-pair model. We discuss its weaknesses and survey theoretical and empirical improvements. We devise our incremental entity-mention model that addresses the problem of underspecification of certain German pronouns and exemplify its theoretical advantages over the mention-pair model.

Chapter 3 discusses evaluation of coreference and pronoun resolution. We point out issues in the commonly used evaluation framework from the perspective of downstream applications. We introduce an alternative evaluation framework that supports the view of potential downstream applications that benefit form coreference resolution as a preprocessing step.

Chapter 4 surveys related work on German coreference and pronoun resolution in different regards.

Chapter 5 empirically validates the theoretical claims of the advantages of the entity-mention model over related work. We also compare different machine learning frameworks that correspond to different antecedent selection strategies.

Chapter 6 explores the utility of distributional models to determine the compatibility of an antecedent candidate and a pronoun's context. We propose a graph-based model of syntactic co-occurrence and compare it to a state-of-the-art approach to word embeddings.

Chapter 7 summarizes our findings regarding our hypotheses and addresses issues to be investigated in future work.

1.4 Coreference and anaphoricity

We introduce coreference and anaphoricity by shedding light on the linguistic background in the following section. We do so with a focus on corpora that feature coreference annotation for research in Computational Linguistics (CL) and Natural Language Processing (NLP) in order to understand and motivate the kind of annotation in these corpora. We then explore coreference from the perspective of CL and NLP applications and discuss how they benefit from coreference resolution as a preprocessing component. Thereafter, we discuss how automatic approaches to coreference resolution tackle the problem by exploring the basic architecture of the common coreference resolution pipeline.

1.4.1 A brief linguistic introduction

Text Linguistics states that texts generally feature the properties of being cohesive and coherent (Halliday and Hasan, 1976, De Beaugrande and Dressler, 1981, inter alia). Cohesion subsumes features of a text's surface which glue together the sequence of utterances in the text. For example, the use of discourse connectors such as *because* and *but* signals on the text surface that two utterances are connected logically or rhetorically. Re-occurrence of words is another such cohesive device, because re-occurrence of words implies re-occurrence of the concepts or entities they denote. In turn, the re-occurrence of concepts and entities are indicators of a text's semantics. If concepts and entities are shared across utterances, it can be argued that the utterances are semantically related, which is an aspect of coherence.

Re-occurrence of concepts and entities manifests on the text surface through linguistic mentions. Coreference signifies that these different linguistic mentions, with potentially varying surface forms, denote the same underlying concepts and entities. Thus, the phenomena of coreference and anaphora are tightly interwoven with cohesion (aspects of a text's surface) and coherence (aspects of a text's meaning).

We exemplify the main devices of cohesion and coherence related to coreference and anaphora and discuss how they are incorporated into coreferentially annotated corpora for Computational Linguistic research.

1.4.1.1 Re-occurrence of words

A simple example of re-occurrence is given by Linke et al. (2004), p. 245, translated to English as example 1.

(1) Yesterday, I watched a **bird** building a nest. The **bird** was tiny, but it managed to fetch rather large twigs. Of all places, the **bird** had picked the roller shutter box for its nest.

However, re-occurrence does not necessarily imply coreference, as shown in example 2 (Linke et al., 2004, p. 246).

(2) My **mother** is terribly anxious and always assumes the worst. Anna's **mother** is more endurable: She lets her daughter go out alone in the evening. I'd rather have a **mother** like that.

In example 2, the word **mother** occurs three times. However, each occurrence denotes a different extra-linguistic referent. A coreferentially annotated corpus would thus not link the three occurrences of **mother** in the example above. By contrast, the occurrences of the word **bird** in example 1 all denote the same extra-linguistic bird and would therefore be annotated as coreferent in a gold standard corpus.

Simple re-occurrence of words to refer to the same extra-linguistic referent is often conceived as stylistically unsatisfactory (Linke et al., 2004), which gives rise to the next cohesive device.

1.4.1.2 Word substitution

Another means to repeatedly refer to discourse entities is word substitution. Substitution occurs when the words to mentioning a discourse entity are replaced, compared to an earlier mention. For example, the mention of the person Barack Obama can be realized by a descriptive noun phrase like "the president of the U.S.". Substitution generally triggers or introduces facets of meaning so far not expressed in the discourse (e.g. that Obama is the president of the U.S.). Substitution generally implies coreference, since the substituting words denote the same underlying entity as the substituted words. Therefore, substitution is generally annotated in coreference corpora.

Semantically, the relation between a substituting and a substituted expression is that of IDENTITY, an *is-a* relation (as in "Obama" *is-a* "president of the U.S."). It is important to distinguish this relation from meronymy (the *part/member-of* relation). For example, "car" and "driver" can be in a meronymic relation, which is another device of coherence. However, since "car" and "driver" do not denote the same underlying entity, they would not be annotated as being coreferent in a gold standard. The phenomenon of bridging (sometimes referred to as associative anaphora or bridging anaphora), which subsumes meronymy, but excludes coreference, is a separate but closely related research area to coreference resolution. Terminology is often combined to describe the two phenomena. For example, Versley (2010) uses the term *coreferent bridging* to refer to substitution.

1.4.1.3 Pronominalization

Finally, pronominalization describes the phenomenon of referring to a discourse entity by a pronoun, which can be seen as a special case of substitution. An important difference from re-occurring and substituting expressions is that pronouns cannot be interpreted without an antecedent, i.e. a previous mention of an entity that clearly identifies it, such as a named entity mention (e.g. "Barack Obama"). That is, we can infer the underlying entity based on the expression "the first black president of the U.S.". However, an isolated occurrence of "he" cannot be understood without any discourse context.

Linguistics makes a distinction between the phenomena of anaphora, which captures the property of an expression to rely on another, previous expression for interpretation, and coreference, which generally denotes that two expressions refer to the same underlying entity. However, the term "anaphoric" is commonly overloaded with the meaning of "coreferent" in the Computational Linguistic literature (Ng, 2010, Björkelund and Kuhn, 2014, inter alia). For example, when a substituting noun phrase (such as "the president") is called *anaphoric*, the implied meaning is that there is a previous mention of the entity to the left of the substitution (an antecedent, i.e. "Barack Obama").

It is noteworthy that not all pronouns are anaphoric. The third person pronoun *it* can be used in a non-referential (pleonastic) fashion in expressions like "It is raining" or "It is clear that [...]". In such utterances, *it* does not refer to any previously mentioned entity. Obviously, such pleonastic uses of *it* are not annotated in coreference corpora and a challenging task for coreference and pronoun resolution systems is to differentiate pleonastic from anaphoric uses of *it*.

Additionally, pronouns can be anaphoric but not coreferent. Consider the following example, taken from the TüBa-D/Z coreference annotation manual (Naumann, 2007, p. 6):

(3) **Nobody** likes to lose **their** job.

The possessive pronoun "their" in the above example is anaphoric to "Nobody", which, in turn, does not refer to any specific real-world entity. Therefore, the common definition of coreference does not apply. Corpora pursue different strategies to handle these cases of anaphoric expressions. One of them is to distinguish between anaphoric and bound pronouns, where the latter group captures the cases of anaphoric but non-coreferential pronouns.

Having outlined the basic linguistic background of coreference and anaphora, we point to the thesis of Wunsch (2010) which provides a broad overview on the linguistic literature mainly focused on the English tradition and the salience-based view on discourse. Furthermore, the thesis of Versley (2010) excellently ties Computational Linguistic research with linguistic theory on coreference, with a focus on noun coreference and its semantic aspects. Complementary to Wunsch (2010) and Versley (2010), we now turn to exemplifying how the linguistic phenomena affect common applications in Computational Linguistics and NLP.

1.4.2 Significance for subsequent applications in CL and NLP

It is widely accepted in Computational Linguistics that coreference resolution is an important preprocessing step for many subsequent, higher-level applications. We exemplify this for four of such applications and show how they benefit from coreference resolution as a preprocessing component.

- A typical area where coreference resolution is deemed useful is **Information Retrieval**. In Information Retrieval, the relevance of documents for a given query is weighted based on the words that occur in the documents by applying a measure such as TF-IDF. However, if words (e.g. “Obama”) occur in a document in a pronominalized form (“he”) or are referred to by a common noun description (“president”), their within-document term frequency will not be incremented. Salient discourse entities are likely to be referred to by substituting expression such as pronouns. Measuring term frequency of words to denote their importance for a document given a query is therefore impaired by not being able to recognize and count substituting expressions of entities. Therefore, performing coreference resolution before measuring TF-IDF can improve the relevance of the returned documents (Dalton et al., 2011, e.g.).
- In **Information Extraction**, pronouns and nominal descriptions of entities can yield problems in template filling tasks. When pronouns are inserted into template slots, they leave them underspecified as the underlying entities cannot be inferred without the proper antecedents. Similarly, substituting expressions pose a problem for relation extraction tasks, e.g. extraction of protein-protein interactions. If the arguments of such interactions are pronominalized or are referred by nominal

descriptors, extracting them is of no value for subsequent tasks (e.g. "This inhibits the interaction with the protein"). The BioNLP 2011 shared task (Kim et al., 2012) combined coreference resolution and event extraction to address this issue.

- Pronouns can yield errors in **Machine Translation** when languages are involved where nouns have conflicting grammatical gender. For example, in German *Mond (moon)* has masculine gender and the pronoun *er* is used to refer to it. The French translation *lune* has feminine gender and is pronominalized by *elle*. When translating between these two languages, the pronoun in the source language will be translated into a pronoun in the target language which is morphologically incompatible with its antecedent (*Mond* → *sie*; lune* → *il**). Knowing to which antecedent a pronoun refers enables a Machine Translation system to insert the correct pronoun in the target language. The recently held DiscoMT 2015 shared task[1] tackled this problem.

- **Sentiment Analysis** is affected by coreference in target-specific analysis. In target-specific Sentiment Analysis, the sentiment towards a specific target entity (e.g. "Barack Obama") is investigated. Here, we encounter the same problem as in Information Retrieval. If coreference resolution is not applied, the sentiment system will miss any context where the entity is mentioned by a pronoun or a nominal descriptor. Therefore, coreference resolution has the potential to increase the Recall of the target-specific contexts and to help deliver a more broadly supported analysis regarding the target entity (Nicolov et al., 2008, e.g.).

There are more areas in Computational Linguistics where coreference resolution is useful, such as summarization and dialogue modeling, and we have only touched on a few. The examples show that coreference resolution plays an important role as a preprocessing component for applications in these areas. Having outlined the problem from the perspective of such applications, we turn to the basic processing pipeline shared by most approaches to coreference resolution.

1.5 Architecture of automated coreference resolution systems

In this section, we introduce the generic architecture of most coreference resolution pipelines. Since most research on coreference in Computational Linguistics evolves around shared task data sets, we outline the pipeline in accordance with the requirements and settings of these tasks. We do so with an anticipatory glimpse at evaluation

[1] `https://www.idiap.ch/workshop/DiscoMT/shared-task`

of coreference systems, as we will discuss the matter of evaluation in greater detail in chapter 3.

While coreference resolution can be viewed as the task of establishing coreference links among noun phrases and pronouns, coreference systems have to accomplish a set of intermediate tasks. Evaluation is not immune to errors occurring during these steps, which makes it arguably hard to assess system performance w.r.t. the task of actually linking the noun phrases (and pronouns) coreferentially. Therefore, we will point out the effects that each of the steps in the general coreference pipeline have on system performance.

1.5.1 Gold standard annotation

Generally, coreference systems are built on the basis of a gold standard corpus which features coreference annotation. The overall goal of a coreference system is to reproduce the manually added coreference annotation automatically. In evaluation, the difference between the gold annotation and the system annotation constitutes the basis of determining the quality of the system output.

Table 1.1 shows an excerpt of such a gold standard, i.e. the TüBa-D/Z corpus (Telljohann et al., 2004), a treebank for German that contains coreference annotation. It features a CoNLL-style format which verticalizes the text and adds token annotations horizontally as columns. The last column denotes coreference between spans of tokens, where identical numerical IDs indicate membership in the same coreference chain.

Table 1.1 shows that the coreference annotation not only spans the syntactic head of coreferring NPs, but their full projections.[2] We see that in this segment, *die Arbeiterwohlfahrt Bremen* and *ihren* corefer (both having the numeric ID 0 in the coreference column). Further we see that *ihren langjährigen Geschäftsführer Hans Taake* (coreference ID 4) corefers with another NP outside the segment.

To achieve the automatic annotation of coreference in the documents in a test set, a coreference system has to accomplish a set of tasks, all of which influence its output and, subsequently, its performance in evaluation. Based on our example segment, we discuss these tasks and introduce the common nomenclature.

[2] The example actually shows dependency parses. To extract projections in the constituency sense, all tokens depending on a noun or name token are gathered recursively.

tok. ID	**lexeme**	**lemma**	**PoS**	**morph.**	**gov. ID**	**gram. role**	**NE class**	**coref.**
1	Im	in	APPRART	dsm	3	pp	-	-
2	Januar	Januar	NN	dsm	1	pn	-	-
3	hat	haben%aux	VAFIN	3sis	0	root	-	-
4	die	die	ART	nsf	5	det	-	(0
5	Arbeiterwohlfahrt	Arbeiterwohlfahrt	NN	nsf	3	subj	ORG	-
6	Bremen	Bremen	NE	nsn	5	app	GPE	0)
7	ihren	ihr	PPOSAT	asm	9	det	-	(0)|(4
8	langjährigen	langjährig	ADJA	asm	9	attr	-	-
9	Geschäftsführer	Geschäftsführer	NN	asm	13	obja	-	-
10	Hans	Hans	NE	asm	9	app	PER	-
11	Taake	Taake	NE	asm	10	app	PER	4)
12	fristlos	fristlos	ADJD	–	13	adv	-	-
13	entlassen	entlassen	VVPP	–	3	aux	-	-

Table 1.1: Excerpt from the TüBa-D/Z in a CoNLL-style format for the segment *Im Januar hat die Arbeiterwohlfahrt Bremen ihren langjährigen Geschäftsführer Hans Taake fristlos entlassen (In January, the Worker Welfare Association Bremen has laid off its long-term CEO Hans Taake without notice).*

1.5.2 Preprocessing

First, the input text undergoes preprocessing, i.e. it is annotated with linguistic features. This generally involves, at the very least, part-of-speech tagging and syntactic parsing, as exemplified in table 1.1. Named entity recognition, animacy detection, and semantic class labeling are further possible markup steps.

Corpora often provide gold annotations, i.e. annotation that is manually added by human experts, for the preprocessing steps. Evaluation can then be divided into several settings, i.e. a setting where the coreference systems make use of the gold preprocessing information (called the **gold setting**) vs. using automated tools such as dependency parsers for preprocessing (referred to as the **regular setting**).

In the regular setting, systems would have to automatically convert our example segment *Im Januar hat die Arbeiterwohlfahrt Bremen ihren langjährigen Geschäftsführer Hans Taake fristlos entlassen* into the CoNLL format (or any other suitable intermediate format) and automatically produce the annotations in the columns shown in table 1.1. Obviously, evaluation of coreference does not consider any other column than the last one. The regular setting denotes that systems are not given any other information than the texts themselves.[3]

[3] Systems can also be given the same automated preprocessing, as in the SemEval 2010 shared task (Recasens et al., 2010)

Comparison of the regular and gold settings gives insight into the impact of using real preprocessing and indicates what performance is to be expected if systems are applied in a real-world setting. The gold setting, by contrast, is insightful when trying to establish which coreference resolution strategy is more successful than others, because system performance is not affected by preprocessing errors.[4]

Coreference resolution approaches differ regarding the features they apply and the preprocessing components they rely on. Therefore, shared tasks often feature a **closed setting**, where only information provided by the shared task data may be used as the basis for feature extraction. Conversely, the **open setting** allows participants to incorporate any external resources or tools into their systems. The closed setting is thus also aimed at identifying successful coreference resolution strategies by canceling out the impact of additional resources, while the open setting identifies the most successful coreference resolution system overall.[5]

1.5.3 Markable extraction

After token annotation in preprocessing, the noun phrases and pronouns that will be considered to take part in coreference relations (the so-called **markables** (Hirshman and Chinchor, 1998, Poesio, 2004)) have to be identified and extracted. This is called the markable extraction step. Most approaches rely on the syntactic analysis of the sentences to identify the relevant NPs and their boundaries. In our example segment, such an approach extracts the following four markables: *[Januar], [die Arbeiterwohlfahrt Bremen], [ihren], [ihren langjährigen Geschäftsführer Hans Taake].* Note that pronouns (in our case the possessive pronoun *[ihren]*) are also markables, although they do not (necessarily) denote an individual NP.

This step is of major importance, as the general coreference evaluation framework requires systems to perfectly match boundaries (i.e. the span of tokens) of coreferent NPs in order to count as correctly resolved. Providing NP boundaries not conforming with the gold annotation has the potential to impact the overall system performance in evaluation significantly.[6] Performance for this step is also often plagued by inconsistent

[4] However, this evaluation is still plagued by the problem of matching NP projections between system outputs and the gold standard, cf. the next section.

[5] Past shared task have shown that not all systems will participate in all settings. It is therefore not always easy to determine a clear winner.

[6] For example, Pradhan et al. (2011) conducted a post-task evaluation of the CoNLL 2011 shared task in a setting where only the syntactic heads of coreferring markables needed to be identified and found that some systems, in particular our submission (Klenner and Tuggener, 2011b), climbed up the result ranks. Our results improved from 51.77% to 55.28% average F-score, raising our position from 4th to 3rd rank in the open setting (5 participants) and from 9th to 5th rank in the closed setting (18 participants).

annotation of NP boundaries in the gold standard (e.g. including or excluding PP attachments and relative clauses). Commonly, annotation guidelines for coreference state that the maximal projections of the syntactic heads of NPs should be considered as markable boundaries. However, correctly marking the maximal projection of syntactic heads manually is a challenging and error-prone task.

Given that the correct identification of the NP boundaries is of such substantial importance for performance evaluation, it remains unclear why gold standards for coreference research feature NP boundaries that extend beyond the syntactic head. We argue that it should not be considered the task of coreference systems to correctly identify NP boundaries, which is rather the aim of syntactic parsing or chunking. Thus, including the task of identifying NP boundaries arguably obscures comparison and performance calculation of coreference resolution systems.[7]

Some systems filter the extracted markables to weed out NPs that are unlikely to enter coreference relations. This task is referred to as **anaphoricity determination** or **mention detection** and has been shown to improve system performance in some cases (Recasens et al., 2013, inter alia). The filtered non-referring markables are called **singletons**. In our running example, the markable *[Januar]* constitutes a singleton.

Most coreference resolution approaches and corpora containing coreference annotation consider as markables the noun phrases and pronouns. It is noteworthy that clauses or full sentences can also be coreferent, like in the following example:

(4) The laptop's CPU overheated .
[That] was unfortunate.

Since there is almost no training data available, only a few approaches exist that also include clauses as markables. While the most commonly used English corpus for coreference resolution, OntoNotes (Pradhan et al., 2013), contains a few coreferring clause instances, there are, to our knowledge, no German corpora containing mentions that denote clauses.

[7] Also, from the perspective of downstream applications it is not necessary for a coreference resolution system to provide extended NPs, because the downstream application can itself decide what kind of extensions it prefers for the coreferent NPs (including or excluding PPs and relative clauses) and produce them freely based on the syntactic heads. We outline in section 5.1.2.2 how we address this problem.

1.5.4 Markable resolution

Finally, coreference systems establish coreference relations between the extracted markables and merge them into coreference chains in order to form the **coreference partition**, i.e. the set of coreference chains in a document. To do so, systems rely on a discourse processing strategy. This strategy traverses the markables and evaluates potential coreference links. The strategy usually filters markable pairs based on morphosyntactic agreement and distance constraints (e.g. markables have to match regarding their number and gender properties and cannot be more than a set number of sentences apart). A classifier then determines whether two (or more) markables should be labeled as coreferent based on a feature set.

In our running example, such a discourse processing strategy would suggest linking the markables *[die Arbeiterwohlfahrt Bremen]* and *[ihre]*, since they are morphologically compatible and positioned very close to each other. It would, however, not suggest linking *[Januar]* and *[ihre]*, since the two markables are morphologically incompatible.

In pair-wise discourse processing strategies, such as the mention-pair model[8], whose basic workings we have just outlined, pairs of anaphors and antecedents are linked in a first step. A second step, the clustering step, is needed to merge the found pairs into coreference chains given the transitivity property of coreference: If markables A and B corefer, and B and C corefer, then A and C corefer.

Markables that have been put in coreference chains become **mentions** of the underlying discourse entities, and unresolved markables remain **singletons**, since they are the only mention of their entity in the discourse at hand. Mentions in the system output are called **system mentions** and those in the gold standard **gold mentions**. Markables that have been resolved by a system, but are not annotated as coreferent in the gold standard, are called **spurious system mentions** or **twinless system mentions** (Stoyanov et al., 2009). Shared tasks commonly feature a **mention detection score**, which measures how well systems divide markables into singletons and mentions.[9]

Shared tasks sometimes also offer a **gold mention setting**, where participants are given the boundaries of the mentions in the gold standard (i.e. the parentheses in the last column of table 1.1, but not the coreference chain IDs). The systems then only need to establish the correct links between the gold mentions, which eradicates the problem of identifying which markables should be considered for resolution. Naturally, performance figures for these evaluation runs are much higher compared to the other runs (Pradhan

[8] Cf. 2.1.

[9] Recall measures how many of the gold mentions are deemed coreferent by a system, regardless of the correctness of the produced coreference links. In turn, Precision quantifies how many of the system mentions correspond to gold mentions.

et al., 2011, inter alia), since an important subproblem in the coreference pipeline is assumed to be solved.

The markable resolution step is what we deem the actual core task of the automated coreference resolution process, and it constitutes what most of the research on coreference in Computational Linguistics investigates. Naturally, this is also the area where systems differ the most. However, as indicated above, other steps in the processing pipeline can heavily influence a system's performance in the common evaluation framework.

Having developed an understanding of the coreference resolution task, we next focus on the discourse processing strategy, as it is an crucial part of the coreference pipeline and an area to which this thesis contributes.

1.6 Chapter summary

This chapter outlined the goals of this theses and introduced the coreference phenomenon. We presented a Text Linguistics account of coreference and discussed the implications of coreference for an exemplary set of applications in Computational Linguists.

We discussed a common pipeline for coreference resolution systems and showed how the contained steps affect evaluation. Additionally, we introduced the common nomenclature encountered in shared tasks on coreference resolution.

Chapter 2

Discourse processing models for coreference resolution

In this section, we focus on the discourse processing models that approaches to coreference resolution apply to process the markables. We first discuss the predominant model and its shortcomings, and then overview conceptual and empirical improvements. We introduce our coreference model that addresses the issue of underspecification of mentions, with a focus on German pronouns.

We note that there is a large diversity in the experimental setup in the related work discussed below. Thus, it is difficult to assess which approach performs best in general.[1] Nonetheless, we will indicate performance scores in cases where a baseline is compared to an extension in the same experimental setup.

To facilitate our discussion, we overload the term *mention* to denote NPs and pronouns (potentially) partaking in coreference relations.[2]

2.1 Mention-pair model

The most prevalent model for establishing coreference between mentions is the so called mention-pair model introduced by Aone and Bennett (1995) and McCarthy and Lehnert (1995), but popularized in its variant presented in Soon et al. (2001). The model's name

[1] Some of the approaches discussed below resolve gold mentions only. Given the gold mentions, a system only needs to establish the correct coreference links between them. That is, the system does not need to decide which NPs it should resolve (cf. section 1.5.4). Other systems process all markables in a more realistic setting, which raises the task complexity. Furthermore, the approaches evaluate on different test sets collected from different corpora, sometimes using different evaluation metrics. Therefore, we cannot compare their scores directly.

[2] For a concise explanation of the nomenclature, cf. section 1.5.4.

implies its basic workings: It creates pairs of mentions and represents them in feature vectors. A binary classifier then decides for each pair instance whether it should be labeled as coreferent.

Algorithm: Mention-pair Training

```
Input: Markables, gold coreference partition
Output: Pair instances $Pairs$
1: for $m_i \in Markables$ do
2:     if $m_i \in CorefPartition$ then
3:         for $m_j \in Markables$ do
4:             if $j < i \wedge coref(m_j, m_i)$ then
5:                 $Pairs \oplus \{m_j, m_i, positive\}$
6:                 break
7:             else if $j < i \wedge \neg coref(m_j, m_i)$ then
8:                 $Pairs \oplus \{m_j, m_i, negative\}$
9: return $Pairs$
```

Algorithm: Mention-pair Testing (Closest first)

```
Input: Markables
Output: Coreference partition
1: for $m_i \in Markables$ do
2:     for $m_j \in reversed(Markables)$ do
3:         if $j < i$ then
4:             $class \leftarrow classify(m_j, m_i)$
5:             if $class == positive$ then
6:                 $PositivePairs \oplus \{m_j, m_i\}$
7:                 break
8: $CorefPartition \leftarrow trans\_merge(PositivePairs)$
9: return $CorefPartition$
```

Table 2.1: Mention-pair algorithms for creating training instances (left) and for resolving markables (right).

Table 2.1 shows the algorithms for creating training and testing instances as proposed by Soon et al. To obtain training instances (left algorithm), a gold mention m_i (line 2 determines that the mention is coreferent according to the gold standard) is paired with the immediate antecedent in its coreference chain to create a positive instance (line 5). Negative instances are formed by pairing m_i with all mentions (including singletons) of other entities on the way to the closest antecedent of m_i (lines 7-8). Soon et al. trained a binary decision tree on these training instances. Subsequent work has explored other machine learning frameworks.

In our running example (*Die Staatsanwaltschaft [...] Im Januar hat die Arbeiterwohlfahrt Bremen ihren langjährigen Geschäftsführer Hans Taake fristlos entlassen*), consider m_i to be the possessive pronoun [*ihre*]. The algorithm would iterate the previous (morphologically compatible) mentions from right to left, i.e. [$die\ Arbeiterwohlfahrt\ Bremen$] and [$Staatsanwaltschaft$]. It would, however, stop at [$die\ Arbeiterwohlfahrt\ Bremen$], since it is the closest antecedent according to the gold standard. The algorithm would thus only create one positive instance, i.e. the pair [$die\ Arbeiterwohlfahrt\ Bremen - ihre$]. If there were any intervening mentions of other entities, these would be used to form negative instances.

When establishing coreference relations (right algorithm in table 2.1), the mentions are traversed from left to right and for each an antecedent (i.e. a preceding mention) is sought, again in a backward-looking manner (lines 2-7). That is, each mention is paired with preceding mentions (sorted by proximity) until a pair is classified as positive. This is called the *closest-first* heuristic. The *best-first* heuristic pairs a mention with all preceding mentions. The positive pair with the highest score then yields the antecedent for the mention at hand. Once all mentions have been traversed, the positive pairs are transitively merged (line 8) to form the desired coreference chains.

Consider again our running example and [*ihre*] as m_i. The algorithm would again start to iterate the preceding mentions in a right-to-left manner and pair the pronoun with them to create the vector representations. Using the closest-first heuristic, if the classifier labeled the first pair [$die\ Arbeiterwohlfahrt\ Bremen - ihre$] as positive, it would append the pair to the list of positive pairs and stop. For the best-first heuristic, the algorithm would also consider the pair [$Staatsanwaltschaft{-}ihre$]. The pair labeled positive with the highest score would then be appended to the set of positive pairs.

Despite using what Soon et al. called a shallow feature set[3], the approach yielded competitive results compared to the mainly rule-driven coreference resolution approaches participating in the MUC-6 and MUC-7 coreference shared tasks (Hirshman and Chinchor, 1998). Furthermore, the system has often been re-implemented and extended. For instance, most systems in the CoNLL shared tasks relied on a mention-pair architecture (Pradhan et al., 2011, 2012, p. 23; 22), despite its commonly known weaknesses, which we discuss next.

2.1.1 Issues of the mention-pair model

Research evolving around the mention-pair model has revealed several conceptual weaknesses. The main issue of the model lies within its local confinement regarding the coreference decisions. All decisions are kept local during the iteration over the mentions, i.e. no information is propagated to subsequent decisions. This runs counter to the transitive nature of the coreference phenomenon and yields several problems.

2.1.1.1 Underspecification of antecedent candidates

During resolution, the local confinement of the mention-pair model is prone to lead to inconsistent coreference sets when the pairs of coreferring mentions found locally are merged (Klenner and Ailloud, 2008, Raghunathan et al., 2010, Ng, 2010, Klenner and Tuggener, 2011a, inter alia). The merge operation (line 8 in the right algorithm in table 2.1) exploits the transitive nature of coreference: If mention A is coreferent with mention B, and mention B is coreferent with mention C, then A and C have to be coreferent. Thus, systems implementing the mention-pair model use the transitive closure to merge the local pairs [$A - B$] and [$B - C$] into a coreference chain [$A - B - C$].

[3] Here, we focus on the discourse models and leave out the discussion of feature sets. For an overview of a feature set like the one used in Soon et al., see section 4.2.

However, this approach suffers from underspecification of mentions in local contexts. For example, assume we have processed the following three mentions: [*Bill Clinton*], [*Clinton*], [*she*]. We have established the following positive pair-wise decisions: [*Bill Clinton – Clinton*] and [*Clinton – she*]. The transitive closure will construct the following coreference chain: [*Bill Clinton – Clinton – she*], which is obviously inconsistent, since [*Bill Clinton*] and [*she*] are exclusive. However, since the [*Clinton*] mention is morphologically underspecified when viewed in isolation, it is a valid local antecedent candidate for the pronoun [*she*].

This is particularly problematic in combination with the morphological underspecification of certain German pronouns, e.g. *sein (its/his)*. A classifier might label the following pairs as coreferent: [*Berlin – sein*], [*sein – er*] ([*Berlin – its/his*], [*its/his – he*]). This would yield the chain [*Berlin – sein – er*], since the incompatibility of [*Berlin*] and [*er*] is not evident to the greedy merge operation which only considers positive pairs and has no knowledge of negative evidence. As outlined in the introduction in section 1.2, this particular problem comprises one of the main interests of this theses.

2.1.1.2 Redundant instances and skewed training sets

A second major issue of the mention-pair model is the generally large number of instances and the imbalance of positive and negative ones. For training, the method for creating pair instances shown in table 2.1 leads to a skewed set, since the closest antecedent for a given mention can be quite far away. Collecting all intermediate mentions of other entities as negative instances yields many such negative instances. For example, Soon et al. (2001) reported that only between 4-7% of the instances in their training set were positive. This imbalance biases the trained classifier towards negative classification (Ng, 2010, inter alia), which can leave e.g. third person pronouns unresolved if no pair of antecedent candidate and pronoun is classified as positive (Hinrichs et al., 2005, Wunsch, 2010).

Since transitive coreference links between mentions are established after classification, i.e. during the merge step, the model yields many redundant instances. Consider that we have the two coreference chains: *[Bill Clinton - Clinton - President - he]* and *[Angela Merkel - Merkel - Chancellor - she]* and we want to resolve the pronouns. Since all mentions of each coreference chain are generally accessible, the mention-pair model potentially pairs the pronouns with all of them, except for the first mentions that have clear morphological properties on their own (given the approach implements morphological agreement as a hard filter), as shown in figure 2.1.

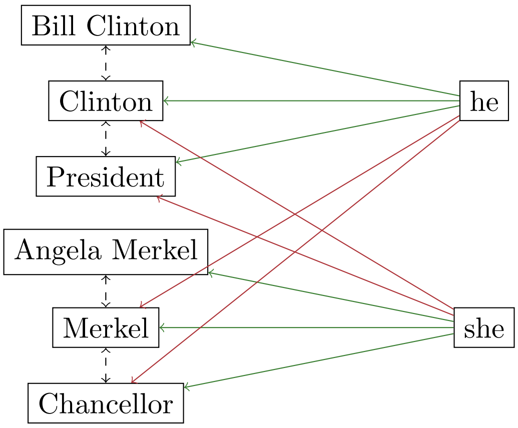

FIGURE 2.1: Example of redundant pairs formed by the mention-pair model for pronouns. Solid arrows denote pairs considered by the model, green ones positive instances and red ones negative instances. Dashed arrows signify coreference links invisible to the model.

The mention-pair model could thus create five pairs for each of the pronouns, i.e. ten in total. All pairs but one per entity formed by the mention-pair model can be considered (at least implicitly) redundant, since they denote the same underlying entity. The pairs that denote morphologically incompatible entities can be regarded as irrelevant, since they should not be considered for resolution.

2.2 Overcoming the limitations of the mention-pair model

A substantial body of research on coreference resolution in the past decade has focused on addressing the inherent issues of the mention-pair model outlined in the previous section. One important aspect introduced in these approaches is incremental discourse processing. Instead of separating local pair-wise decisions from the transitive merge when reaching the document end, incremental models try to combine the two steps, which makes it possible to propagate information from one decision to another. Thus, all detected previous mentions of an entity that acts as an antecedent candidate are accessible when a subsequent anaphor is resolved. This ameliorates the issue of local underspecification of antecedent candidates, since morphological (but also other) features from all previous mentions of an antecedent entity can be queried.

In the following sections, we review approaches that address the aforementioned weaknesses of the mention-pair model. We do so anticipating our incremental entity-mention model for German coreference and pronoun resolution, which combines and extends several notions of the approaches discussed below.

2.2.1 Mention ranking model

The mention ranking model (Denis and Baldridge, 2007, 2008) is an extension of the method for creating training and testing instances in the mention-pair model. The main difference is that the mention ranking approach does not classify pairs of antecedent candidates and anaphors in isolation, but rather applies a ranking of all candidates for a given anaphor simultaneously. The candidate with the highest rank is then selected as the antecedent. This eliminates the need for a best-first or closest-first selection in the mention-pair model in the case that multiple candidates are classified positively as antecedents. That is, there is always exactly one antecedent per anaphor, i.e. the highest ranked candidate.

Training instances are now comprised of an anaphor and a set of antecedent candidates with exactly one correct antecedent. Weights for features are learned based on the competition of the correct antecedent and all incorrect ones. The model for ranking candidates is expressed as a probability distribution over the candidates (Denis and Baldridge, 2007):

$$P(m_j, m_l) = \frac{exp(\sum_{i=1}^{n} \lambda_i f_i(m_j, m_l))}{\sum_k exp(\sum_{i=1}^{n} \lambda_i f_i(m_k, m_l))} \tag{2.1}$$

where m_l denotes an anaphor and m_j one of the candidates, and $\lambda_i f_i(m_k, m_l)$ signifies the weighted features of the pair. The point-wise probability of a candidate to denote the correct antecedent $P(m_j, m_l)$ is calculated by normalizing the sum of the weights of the candidate by the sum of the weight sums of all other candidates and their features. Therefore, the model can be said to express competition between multiple antecedent candidates directly, which overcomes the isolated view on single pairs of candidates and anaphors in the mention-pair model. This extends the twin-candidate model proposed by Yang et al. (2003), where an instance is composed of an anaphor and two of the competing antecedent candidates.

Martschat and Strube (2015) recently showed that the mention ranking approach outperforms not only a mention-pair competitor, but also an approach that models coreference sets with latent tree structures and applies structured perceptron, which has become increasingly popular recently (Björkelund and Kuhn, 2014, Fernandes et al., 2014). However, the mention ranking approach still shares with the mention-pair model the flaw of local confinement, i.e. coreference decisions are kept local, and entity-level information is not accumulated and propagated to subsequent decisions.

2.2.2 Mention clustering models

Mention clustering models recast coreference resolution as an incremental clustering task instead of modelling it as a binary classification problem. One of the first of such approaches was presented in Cardie et al. (1999). The approach of Cardie et al. first put all mentions in their own coreference set, i.e. they were treated as singletons. Working in the reversed text direction, mentions were compared to preceding ones and incrementally clustered into coreference sets w.r.t. a distance metric based on features commonly used in coreference resolution.

The approach featured one crucial operation. When two clusters (each containing one or more mentions) were considered for merging, compatibility between all the mentions in both clusters was asserted. If two mentions from the two clusters did not agree in e.g. number and gender, the merge was prevented. Doing so, the problem of contradicting morphological properties (but also semantic properties, such as animacy) in coreference chains, as present in the mention-pair model, was avoided.

In evaluation, the approach yielded a ranking in the middle field compared to contemporary approaches, despite its simple feature set. However, Cardie et al. did not compare the model directly to a mention-pair approach in the same experimental setup.

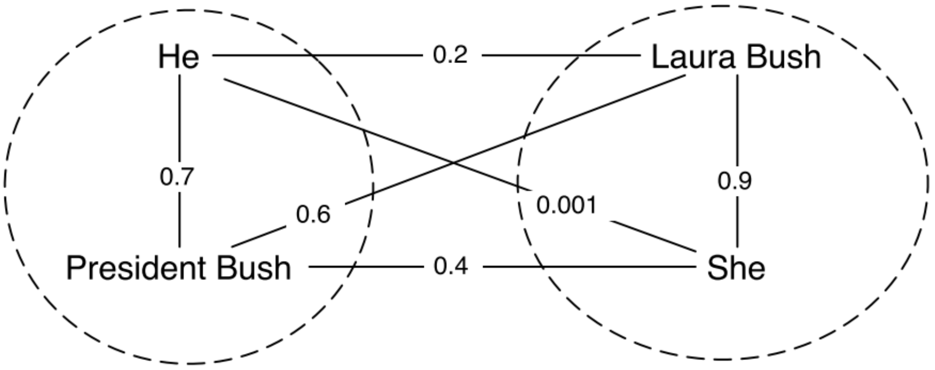

Figure 2.2: Modelling coreference as a graph. Vertices denote mentions, and edges denote potential pair-wise coreference relations and their weights. The circled entity clusters indicate gold coreference clusters. Example due to Culotta et al. (2007).

Another model of mention clustering is presented in graph partitioning approaches to coreference (Nicolae and Nicolae, 2006, Culotta et al., 2007, Cai and Strube, 2010a, inter alia). Mentions are stored as vertices (i.e. nodes), and edges between them signify potential coreference relations. Initially all mentions are connected through edges. The edges carry weights based on the binary and unary features of the connected mention pairs, as shown in figure 2.2. A graph cut algorithm then cuts edges based on their weights to extract the coreference partition from the graph. In figure 2.2, a cut algorithm would ideally cut all edges which connect the two circled clusters. The stopping criterion for the cut algorithm is determined empirically with the help of machine learning. Once the cutting algorithm has stopped, the established subgraphs or vertex clusters denote coreference sets.

Nicolae and Nicolae (2006) applied the *BestCut* method to remove edges from the graph and achieved state-of-the-art results in evaluation. However, they needed to treat pronouns separately, i.e. pronouns were not included in the graph, but attached according to the highest ranked pair-wise decision. Also, Nicolae and Nicolae did not include any cluster-level features, unlike e.g. Cardie et al. (1999). The approach of Nicolae and Nicolae did however outperform a mention-pair baseline and the Luo et al. (2004) entity-mention system.[4]

Cai and Strube (2010a) extended this approach by introducing hyperedges. Hyperedges denote features spanning (possibly) multiple mentions (i.e. nodes in the graph). For example, a hyperedge denoting head string match connects all mention nodes in the graph whose heads match (e.g. *[US President Barack Obama - Barack Obama - Obama]*). Spectral clustering was applied to cluster these subhypergraphs formed by the hyperedges into coreference clusters. However, like Nicolae and Nicolae (2006), the approach did not feature any means to enforce consistency regarding gender etc. in the emerging coreference clusters, and Cai and Strube noted that they found such inconsistencies in the coreference chains in their system output. Despite this drawback, the system outperformed two strong mention-pair baselines and ranked among the top systems in the CoNLL 2011 shared task (Pradhan et al., 2011).

Culotta et al. (2007) applied first-order logic to capture cluster- (or subgraph-)level features of arbitrary clusters of potentially coreferring mentions. First, a feature encoded how many of the mention pairs (*All, Most-True, Most-False*) in an arbitrarily generated cluster shared a particular feature (such as a WordNet class or gender agreement). A second feature encoded how many of the mention pairs in the cluster were coreferent (*All-True, Most-True, Most-False*), as well as whether the maximum and minimum pairwise scores were above a given threshold. Additionally, cluster size and the distribution of mention types were used as features. The latter is particularly interesting, since it tries to model regularities in mention type distributions in coreference sets, which Culotta et al. speculated would help prevent the formation of sets comprised of pronouns only. Learning and inference then aimed at finding relations between feature sharing and cluster purity (e.g. *Most-True* for feature sharing and coreference). This approach outperformed a mention-pair baseline in evaluation by large margins. However, Culotta et al. did not evaluate the impact of the cluster-level features separately.

[4] Cf. the next section 2.2.3.

An advantage of the clustering- and graph-based approaches over the mention-pair model is that they perform coreference resolution in a single step, i.e. once the cutting or clustering algorithm has finished, the coreference partition is established. Furthermore, graph-based models consider multiple coreference links at once, instead of processing individual pairs of markables in isolation. The cut algorithm also adheres to the transitivity and exclusiveness restrictions of the coreference relations, as it places a vertex in only one subgraph or cluster. Clustering approaches also feature the benefit of having access to all mentions of the incrementally established cluster to determine compatibility with an anaphor at hand, which ameliorates underspecification of certain mentions.

A related approach to clustering which avoids inconsistencies in coreference sets enforces global constraints during the merging of classified pair instances in the mention-pair model. An additional layer, e.g. Integer Linear Programming (ILP), guides the pair merging step and ensures transitivity and exclusiveness (Finkel and Manning, 2008, Klenner and Ailloud, 2008, Denis and Baldridge, 2009, Klenner and Ailloud, 2009). The weights of the pair-wise decisions serve as input, and the ILP layer optimizes the clustering given the global coreference constraints. This has been shown to improve performance, but requires considerable engineering and computational effort. Additionally, the approach has the drawback of still relying on the pair generation mechanics of the mention-pair model.

2.2.3 Entity-mention models

The entity-mention model extends the mention-pair model by combining pairwise decisions with the transitive merge of resolved mentions in a single step. That is, pair-wise decisions regarding coreference of two mentions are stored and made accessible for subsequent decisions.

Luo et al. (2004) introduced such an approach and combined it with global optimization of the coreference partition. The possible partitions of the mentions at a given point in discourse were represented in a tree, i.e. the Bell tree. Figure 2.3 depicts such a tree for three mentions m_i, m_j, m_k. For each mention, the approach decided whether it should be attached to an existing coreference chain or if it should start its own chain. In figure 2.3, the mention m_k is processed in nodes 2 and 3. Node 2 and its leafs represents the decision whether m_k should be merged into the existing chain (leaf 4) or if it should start its own (leaf 5). To decide that, m_k is paired for each feature with either m_i or m_j in node 2, depending on which pair yields the better score for each feature.

That is, to decide whether a mention should be linked to a given antecedent chain, it has access to all mentions in that chain, and the maximum pair-wise score is taken. For each feature, the highest scoring pair is determined anew by traversing all mentions in the antecedent chain. For example, for distance features, the most recent mention of the antecedent entity is accessed. For string matching features, the most similar one to the potential anaphor is used to calculate the feature value and extract the respective weight.

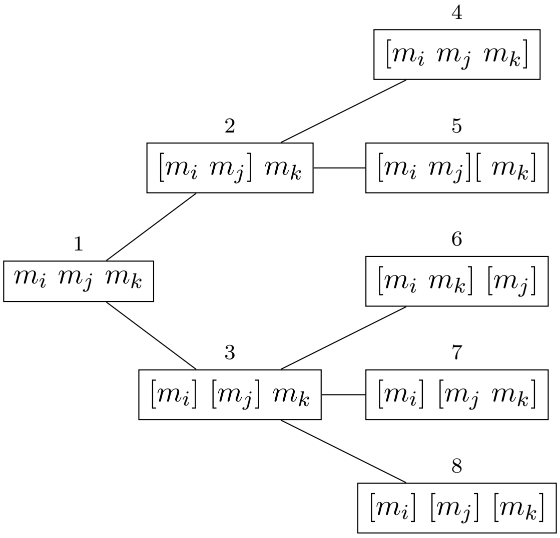

Figure 2.3: Example Bell tree for three markables as modeled in the entity-mention model in Luo et al. (2004).

The leaf nodes of the Bell tree each denote one of the possible partitions of the mentions. The aim is to identify the one that is most probable. Leaf node 4 shows a partition that puts all mentions in one coreference set, leaf 8 a partition that puts all mentions in their own clusters, and the leaf nodes in between (5-7) denote all possible permutations. Since the search space is large when all mention partitions are considered, Luo et al. applied pruning based on heuristics and beam search. Furthermore, they did not create branches that would attach mentions to entities with incompatible type.[5]

Somewhat disappointing, Luo et al. found that the entity-mention model did not outperform a mention-pair competitor in evaluation. However, the mention-pair model used more features. More relevant to our interest, Luo et al., p. 6 noted that the mention-pair model created coreference sets with incompatible pronouns, e.g. putting *he* and *she* in the same chain. These errors were not found in the output of the entity-mention model.

Yang et al. (2004a) examined an entity-mention model for coreference resolution in the biomedical domain. Like in Luo et al. (2004), coreference clusters were formed incrementally in a left-to-right pass over the mentions. Each mention was compared to previous clusters[6] in a pair-wise fashion, where only one pair was created per cluster.

[5] The semantic entity type was provided by the ACE gold standard.

[6] Singletons formed their own clusters.

For clusters containing multiple mentions, the score for merging the current mention was given by the maximal pair-wise score over all mentions in the antecedent cluster, similar to Luo et al. (2004). While Cardie et al. (1999) used morphological agreement as a hard constraint, Yang et al. turned it into cluster-level features. For example, a feature encoded whether all members of an antecedent cluster agreed in number with the mention at hand. Other cluster-level features were constructed, such as the number of mentions in the antecedent cluster.

Evaluation showed that the entity-mention model outperformed the mention-pair variant by a significant margin and raised F-score from 78.9% to 81.7%. However, Yang et al. found that from the features on the cluster-level, only the string-matching features improved performance, i.e. the compatibility-related features had little effect. Later, Yang et al. (2008a) applied Inductive Logic Programming to learn coreference rules for linking entities and mentions and applied them on top of their earlier entity-mention model, which further improved performance.

Of particular interest for us, Yang et al. (2004b) employed an entity-mention-style approach for pronoun resolution in English. The main idea of the approach was to access features of the antecedent (*ante-of-cand$_i$*) of an antecedent candidate (*cand$_i$*) of a pronoun (given the candidate (*cand$_i$*) was already in a coreference chain). Training and testing instances were formed like in the mention-pair model. However, in the case the antecedent candidate (*cand$_i$*) had itself an antecedent (*ante-of-cand$_i$*), features describing its antecedent (*ante-of-cand$_i$*) were incorporated into the mention-pair instance. Evaluation showed that pronoun resolution performance increases up to almost 5 percentage points in success rate (from 70.00% to 74.40%) compared to a standard mention-pair model, highlighting that keeping track of previous decisions, a main feature of the entity-mention model, is also important for pronoun resolution.

Daumé et al. (2005) proposed a system that jointly modeled mention detection[7] and coreference resolution in an entity-mention approach. One advantage over other approaches was that the type of the antecedent mention under scrutiny depended on the type of the anaphoric mention. That is, if the anaphoric mention was a name mention (i.e. its head token was a named entity) the coreference chain of the antecedent candidate was queried for a name mention. If none was found, a nominal mention was sought. Failing again, the mention that produced the highest pair-wise score was selected as the representative mention of the chain. Analogously, specific criteria were applied when the anaphoric mention at hand was a noun or a pronoun. Daumé et al. argued that this has the advantage that pairs are favored which facilitate learning and inferring coreference.

[7] The task of identifying which of the markables are actually in coreference sets (cf. section 1.5.3).

Daumé et al. applied an impressively large, diverse, and heavily lexicalized feature set. In particular, they explored many novel global coreference partition-level features, such as number of entities detected so far, entity-to-mention ratio, and size of the potential coreference chain. These features proved to be beneficial in evaluation. However, Daumé et al. noted that their system underperformed regarding pronoun resolution.

The winning system of the CoNLL 2011 shared task was a rule-based entity-mention approach which became part of the Stanford CoreNLP pipeline.[8] Raghunathan et al. (2010) introduced this system, which incrementally applies a battery of coreference sieves in a precision-first manner. Each sieve consists of rules, such as string matching of mention heads, to determine coreference between mentions. The output of a sieve served as the input to a subsequent, less precise coreference rule etc. That is, after passing the first sieve, some coreference chains are partially formed. This partial partition of the mentions then passes the next sieve which merges chains or adds new ones. Despite its arguable simplicity, the system won the CoNLL 2011 shared task.

Rahman and Ng (2009) presented a combination of the mention ranking approach and the entity-mention model. It adapted the mention ranking's method of ranking all antecedents candidates for a mention at once and implemented the incremental creation of coreference sets of the entity-mention model. That is, after selecting an antecedent for a mention, the mention was appended to the antecedent's coreference chain immediately. Subsequent mentions then accessed all the mentions, like in Luo et al. (2004), to determine whether they should be merged with the chain. Given this combination, Rahman and Ng presented their approach as a cluster ranking model. In evaluation, the approach outperformed the competing models which represented the components of the cluster ranking model, i.e. a mention-pair model, a mention ranker, and an entity-mention model.

In conclusion, related work has presented a set of conceptual improvements over the mention-pair coreference model. However, these conceptual improvements did not always carry over to a better performance in evaluation. Also, despite its weaknesses, the mention-pair remains the most applied coreference model, as can be seen from the entries to the CoNLL shared tasks (Pradhan et al., 2011, 2012, p. 23; 22). One possible reason for the persistence of the mention-pair model might be its arguable simplicity. We saw that the entity-mention models, by comparison, implement complex strategies to determine antecedents which require numerous comparisons and queries of all previous mentions of an antecedent entity. In this light, we propose our own incremental entity-mention model that features a simple and efficient method of bookkeeping entity-level information.

[8] `http://stanfordnlp.github.io/CoreNLP/`

2.3 Our incremental entity-mention model for German coreference resolution

In this section, we discuss our incremental entity-mention model for coreference resolution. Klenner and Tuggener (2010) introduced a first version of the model for German coreference resolution. In Klenner and Tuggener (2011a), we extended the system to tackle English coreference resolution. This system participated in the CoNLL 2011 shared task on English coreference (Pradhan et al., 2011).

In Tuggener and Klenner (2014), we explored the model especially for German pronoun resolution. Here, we discuss the model on the basis of Tuggener and Klenner (2014) and show how it is tailored for handling underspecification of certain German pronouns in the light of the models discussed in the previous sections.

As stated earlier, certain German pronouns are underspecified regarding their number and gender attributes, e.g. the personal *sie (she/they)* and the possessive pronoun *ihr (her/their)*, or simply their gender, i.e. the possessive pronoun *sein (his/its)*. This means that both plural and singular/feminine antecedent candidates have to be considered for *sie* and *ihr*, and both masculine and neutral candidates have to be licensed for *sein*, which leads to a large number of candidates.[9] Additionally, if resolved instances of these pronouns are not disambiguated, they can act as antecedent candidates for subsequent pronouns which are incompatible with their antecedents. This leads to incoherent coreference chains when the pairs of antecedents and pronouns are merged, as we will demonstrate.

In the following exposition, we represent mentions in the form $[lexeme]^{morph.}_{entityID}$, where identical values for *entityID* signifies coreference, and $*$ indicates underspecified values. Consider the three mentions $[Berlin]^{neut.}_1$, $[sein]^*_1$, $[er]^{masc.}_2$ (*Berlin*, *its/his*, *he*). When resolving the possessive pronoun $[sein]^*_1$, a mention-pair model would generate the pair $[Berlin]^{neut.}_1 - [sein]^*_1$. Next, when resolving $[\mathrm{er}]^{masc.}_2$, the model would not generate the pair $[Berlin]^{neut.}_1 - [er]^{masc.}_2$, assuming that it applies morphological agreement as a hard filter. However, it would consider the pair $[sein]^*_1 - [er]^{masc.}_2$. If all licensed pairs were classified as positive, the transitive closure would yield the coreference chain $[Berlin - sein - er]$, which is obviously inconsistent, since $[Berlin]^{neut.}_1$ and $[er]^{masc.}_2$ are exclusive. However, the mention-pair model has no means to propagate this exclusiveness from one pair decision to another.

To address this issue, we introduced our entity-mention model which features incremental disambiguation. While we here focus our discussion on German pronouns, it

[9] We quantify these numbers in section 5.3.1.

is noteworthy that the model constitutes a general coreference resolution model for all mention types and is applicable to other languages (Klenner and Tuggener, 2010, 2011a). Algorithm 1 outlines the model based on Tuggener and Klenner (2014).[10]

Algorithm 1 Incremental entity-mention model

```
Input: Markables
Output: Coreference partition
 1: for $m_i \in Markables$ do
 2:     for $e_k \in CorefPartition$ do
 3:         if $compatible(e_{k_n}, m_i)$ then          ▷ $e_{k_n}$ is the most recent mention of entity $e_k$
 4:             $Candidates \oplus e_{k_n}$
 5:     for $m_j \in BufferList$ do
 6:         if $compatible(m_j, m_i)$ then
 7:             $Candidates \oplus m_j$
 8:     $ante \leftarrow get\_best(Candidates)$
 9:     if $ante \neq \varnothing$ then                   ▷ An antecedent has been identified
10:         $ante, m_i \leftarrow disambiguate(ante, m_i)$   ▷ Propagate animacy, NE class, morphology
11:         if $\exists e_k \in CorefPartition : ante \in e_k$ then   ▷ Antecedent is part of a coref. chain
12:             $e_k \oplus m_i$
13:         else
14:             $CorefPartition \oplus \{ante \oplus m_i\}$      ▷ Open new coreference chain
15:             $BufferList \ominus ante$                 ▷ Remove antecedent from buffer list
16:     else
17:         $BufferList \oplus m_i$                      ▷ No antecedent, append $m_i$ to buffer list
18: return $CorefPartition$
```

Before walking through the algorithm, we note that the model uses it for both the training and testing mode. The difference between the two modes lies in the function *get_best*(*Candidates*) (line 8). In training mode, the function accesses the gold standard to identify the correct antecedent among the candidates and returns it. Also, it creates training instances or updates the weights of the applicable features of the candidates, depending on the machine learning framework. During testing, the function scores all candidates and returns the one with the highest score as the predicted antecedent.

The algorithm traverses the markables in a left-to-right pass (main loop lines 1-17) and establishes the coreference partition incrementally, i.e. once finished, there are no further steps required to produce or refine the partition.

For each markable m_i, we gather compatible antecedent candidates from the coreference partition (lines 2-4) and the buffer list (lines 5-7).[11] The buffer list contains markables that have been traversed earlier but have not been resolved to an antecedent. The list

[10] In the previous sections, we have overloaded the term *mention* to subsume markables. Here, we return to the distinct terms to give a concise presentation of the algorithm. The term *markable* denotes all NPs and pronouns considered to be potentially partaking in coreference relations, while *mention* refers to markables that are actually resolved or annotated as being coreferent in a gold standard.

[11] Compatibility is defined based on the markable type, i.e. PoS tag of m_i. Cf. section 5.2 for a more specific description where we implement the model. We here restrain the discussion to the conceptual aspects.

mainly contains nominal markables which can serve as antecedents, since we always resolve pronouns when there is at least one compatible candidate. Obviously, the first markable in a document will always be put in the buffer list, since there are no previous markables to refer to. If a markable from the buffer list is selected as antecedent (lines 13-15), the pair is appended to the coreference partition (line 14), i.e. it becomes one of the entities denoted by e_k for the next markable m_{i+1} in the iteration. The denoted entity e_k is then only accessible through its last mention e_{k_n} which is now m_i. That is, the *ante* markable is now no longer accessible for subsequent reference, which prevents the generation of redundant pairs (as *ante* and m_i denote the same entity). All of *ante*'s relevant features have been projected onto m_i (line 10), and *ante* is removed from the buffer list (line 15).

Furthermore, the coreference partition is queried for compatible antecedent candidates (lines 2-4). Only the last mention e_{k_n} of the mentions of a specific entity e_k in the coreference partition is accessible, the other mentions are hidden, as mentioned above. If an antecedent candidate from such an entity e_k in the coreference partition is selected (line 11), m_i is directly appended to the coreference chain of e_k (line 12).

The restricted accessibility of mentions outlined above is one major difference to other entity-mention models discussed in the previous section. Related entity-mention models (but also clustering-based models) query all previous mentions of an entity to decide whether a markable at hand should be resolved to it. By contrast, we project all known features of the entity to its most recent mention. We then only need to query this last mention to decide whether to resolve a subsequent markable to the denoted entity.

To exemplify the algorithm and highlight its differences to the mention-pair model, recall our example from section 1.5, i.e.:

> Im Januar hat die Arbeiterwohlfahrt Bremen$_1$ ihren$_1$ langjährigen Geschäftsführer Hans Taake fristlos entlassen.
>
> In January, the Worker Welfare Association Bremen$_1$ has laid off its$_1$ long-term CEO Hans Taake without notice.

In this example, the possessive pronoun "ihren" ("her/their") is underspecified. The algorithm first resolves the possessive pronoun $[ihre]_1^*$ to the antecedent $[die\ Arbeiterwohl-fahrt\ Bremen]_1^{Sg.}$ and projects the morphological properties of the antecedent (feminine, singular) onto the pronoun. Also, the antecedent has the named entity class ORG (organisation), and preprocessing has determined it to be an inanimate entity. These three features (number/gender, named entity class, animacy) are projected onto $[ihre]_1^*$, which disambiguates its morphological properties and augments its semantic features (line 10 in algorithm 1). $[ihre]_1^{Sg.}$ is now the last mention of the $[Arbeiterwohlfahrt]_1^{Sg.}$ entity and thus the only accessible mention for subsequent reference, since the markable $[Arbeiterwohlfahrt]_1^{Sg.}$ is removed from the buffer list (line 15).

Having disambiguated this last mention, we prevent the entity from being selected as an antecedent candidate for a subsequently occurring, incompatible pronoun markables, e.g. a plural instance of $[sie]_*^{Pl.}$ (they).[12] Without the projection of the antecedent morphology, a plural instance of $[sie]_*^{Pl.}$ would consider $[ihre]_1^*$ as a candidate, which is what happens in a mention-pair model, since earlier decisions are not considered. If a coreference link were established between the two markables, the coreference chain would feature conflicting morphological properties.

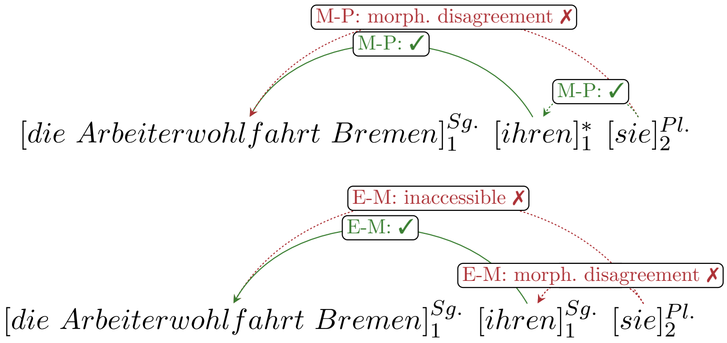
M-P: morph. disagreement ✗
M-P: ✓
M-P: ✓
$[die\ Arbeiterwohlfahrt\ Bremen]_1^{Sg.}$ $[ihren]_1^*$ $[sie]_2^{Pl.}$
E-M: inaccessible ✗
E-M: ✓
E-M: morph. disagreement ✗
$[die\ Arbeiterwohlfahrt\ Bremen]_1^{Sg.}$ $[ihren]_1^{Sg.}$ $[sie]_2^{Pl.}$

Figure 2.4: Differences in pair generation mechanics in the mention-pair (M-P; above) and entity-mention (E-M; below) model. Solid arcs denote established decisions. Red dotted arcs signify potential links not considered, green dotted arcs considered links.

Figure 2.4 illustrates these different pair generation mechanics in the models. Both models establish a correct link from $[ihren]_1^*$ to $[die\ Arbeiterwohlfahrt\ Bremen]_1^{Sg.}$, since the markables are morphologically compatible. When processing the markable $[sie]_2^{Pl.}$, neither model considers a link to $[die\ Arbeiterwohlfahrt\ Bremen]_1^{Sg.}$, but for different reasons. In the mention-pair model, the link is discarded because the markables are morphologically incompatible. The entity-mention model does not consider the link, because the markable $[die\ Arbeiterwohlfahrt\ Bremen]_1^{Sg.}$ is not anymore a member of the buffer list after $[ihren]_1^*$ has been resolved to it. Thereafter, only the $[ihren]_1^{Sg.}$ mention, the last mention of the $[die\ Arbeiterwohlfahrt\ Bremen]_1^{Sg.}$ entity, is accessible for subsequent reference as an antecedent candidate. However, the entity-mention model does not generate the mention $[ihren]_1^{Sg.}$ as an antecedent candidate for $[sie]_2^{Pl.}$, since $[ihren]_1^*$ has inherited the morphological properties of its antecedent (singular/feminine) and is therefore incompatible with $[sie]_2^{Pl.}$. The mention-pair model, by contrast, considers the link from $[sie]_2^{Pl.}$ to $[ihren]_1^*$, because $[ihren]_1^*$ is still morphologically underspecified after its resolution to $[die\ Arbeiterwohlfahrt\ Bremen]_1^{Sg.}$. If the link is realized, $[sie]_2^{Pl.}$

[12] *Sie* is itself morphologically underspecified (she/they). However, *sie* can be disambiguated by the morphology of the governing verb in the case that the pronoun is in the subject position.

becomes a member of the coreference chain, i.e. $[die\ Arbeiterwohlfahrt\ Bremen - ihre - sie]_1^{Sg./Pl.}$, after the transitive closure, which is morphologically incoherent.

Besides ensuring coherence of morphological and semantic properties, the entity-mention model also affects the occurrence count of features during training and testing, since pronoun antecedent candidates now carry the semantic features of the entities they denote, e.g. named entity class and animacy. For example, the $[ihre]_1^*$ mention above features the named entity class ORG after having been resolved to the $[Arbeiterwohlfahrt]_1^{Sg.}$ entity. When encountering a subsequent, compatible personal pronoun, i.e. $[sie]_*^{Sg.}$, the information that $[ihre]_1^*$ denotes an organisation entity might affect the likelihood of being selected as the antecedent of $[sie]_*^{Sg.}$, given that there is another candidate which denotes e.g. a person entity etc. That is, propagating a feature value, like named entity class, between mentions affects the distribution of the value, which in turn affects how machine learning frameworks weight it.

There is a caveat, however. It is possible that an underspecified pronoun is resolved to an incorrect antecedent which then projects its morphological properties onto it. The pronoun then carries incorrect morphological features, which prevents it from becoming the correct antecedent for subsequent pronouns in the chain. Consider an extension and a modification to our previous example, i.e. resolving the mentions $[Berlin]_1^{neut.}$, $[Der\ Mann]_2^{masc.}$, $[sein]_2^*$, $[er]_2^{masc.}$. If the entity-mention model incorrectly resolves $[sein]_2^*$ to $[Berlin]_1^{neut.}$ instead of to $[Der\ Mann]_2^{masc.}$, the now incorrectly disambiguated mention $[sein]_1^{neutr.}$ can no longer act as an antecedent for $[er]_2^{masc.}$ due to falsely produced morphological incompatibility. Furthermore, if the mention $[Der\ Mann]_2^{masc.}$ is more than three sentences away from $[er]_2^{masc.}$, the denoted entity, which is the correct antecedent entity, is no longer accessible for $[er]_1^{masc.}$ due to distance constraints. Therefore, $[er]_2^{masc.}$ would be resolved to some other incorrect but compatible candidate, if available. Empirical evaluation in section 5.4 will assess whether this problem occurs frequently enough to hamper performance of the model significantly, or whether the overall benefits of the entity-mention model outweigh the drawbacks of the mention-pair model, despite the danger of such cascaded errors.

A notable aspect of the incremental entity-mention model is that it mimics human cognitive processes of reading, arguably at least to a certain degree, and certainly to larger degree than other models, e.g. the mention-pair model (Ng, 2010, Klenner and Tuggener, 2011a, Webster and Curran, 2014). For example, Webster and Curran (2014) rephrase the entity-mention model in the nomenclature of the shift-reduce algorithm commonly known from syntax parsing. The stack holds entities. When a markable is processed, a classifier decides whether the markable should be appended to an existing entity (the reduce operation), or if it should denote a new entity (the shift operation).[13] The stack structure can be seen as a model of human short-term memory, and entities nested deeper in the stack (or short-term memory) are arguably harder to retrieve or access for reference.

We argue that the entity-mention model at least captures two aspects of human reading behaviour: i) The coreference relations in a document are created in a single pass over the text, and ii) the model traverses text in a left-to-right manner.

The mention-pair model can be considered a two-step architecture. In a first pass, local pairs are formed and classified. The transitive merge of the positive pairs is a second pass needed to realize the transitive nature of the coreference relations. The entity-mention model, by contrast, forms the coreference partition in a single pass over the markables, which is more closely related to human cognitive processes when reading texts.

In the mention-pair model, the order in which the markables are processed during pair generation is of no significance, since all previous markables are generally considered as antecedent candidates for any markable. This is also true for graph and clustering models for coreference, because they do not take into account the original discourse ordering of the markables. The entity-mention model, by contrast, processes the markables in text direction, i.e. form left to right, which more closely represents human reading behaviour.

Another advantage of our model over the mention-pair model, but also over related entity-mention models, is that it does not consider redundant pairs of markables. To illustrate this, we revisit our example from figure 2.1. Table 2.2 juxtaposes the pair generation mechanics of the mention-pair and entity-mention model for the example. As discussed in section 2.1.1.2, the mention pair model (left) produces many redundant pairs which all denote the same entity, i.e. ten pairs in total for the two pronouns. The entity-mention model (right) propagates morphological information (↓+masc.sg.) from mention to mention and therefore avoids creating invalid pairs. In the example, the model thus only creates two pairs of which both are relevant. We argue that this model is a more plausible depiction of human cognitive processes, since it seems unlikely that humans access all individual previous mentions of an antecedent candidate entity when resolving an anaphor.

While it is in general not necessary to have a coreference model that is closely related to human cognitive processes, especially from an engineering perspective, we argue that it is desirable from a theoretical viewpoint; especially if the relatedness yields a model that overcomes deficiencies of another, more unrelated one.

[13] This is analogous to deciding whether a markable should be put on the buffer list or if it should be resolved to an antecedent which is already a member of a coreference chain in our model.

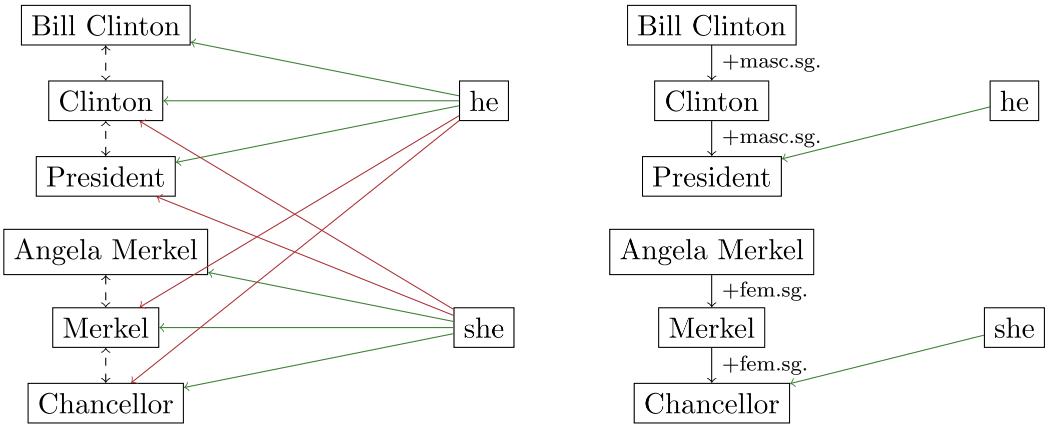

Table 2.2: Example of pairs formed by the mention-pair (left) and entity-mention model (right). Red arrows denote negative pairs, green ones show positive ones. Dashed arrows signify coreference links invisible to the model, and solid black arrows designate coreference links previously established.

2.4 Chapter summary

This chapter presented and discussed the most prevalent discourse processing model for coreference resolution, namely the mention-pair model. We iterated its commonly known weaknesses and demonstrated in particular how they affect the task of German pronoun resolution.

We surveyed related work that conceptually addresses these weaknesses, including clustering and graph-based coreference models. We discussed our entity-mention model which is tailored to cope with the shortcomings of the mention-pair model w.r.t. the underspecification of mentions, and demonstrated its benefits for German pronoun resolution.

Chapter 3

Coreference resolution evaluation

In this chapter, we discuss how coreference resolution systems are evaluated empirically. We outline the general evaluation framework and point to commonly acknowledged problems. The pros and cons of the common evaluation framework as introduced in the CoNLL shared tasks (Pradhan et al., 2011, 2012) have been discussed extensively, and modifications of the metrics have been implemented (Cai and Strube, 2010b) and reverted again (Pradhan et al., 2014).

We argue that the general evaluation framework for coreference resolution is geared towards comparing system responses in a closed-world assumption, i.e. detached from the potential integration of the systems into CL and NLP pipelines. In Tuggener (2014), we proposed the ARCS (Application-Related Coreference Scores) evaluation framework which addresses this issue. This chapter overviews both evaluation frameworks and proposes several extensions to ARCS.

3.1 Issues in evaluation from the perspective of higher-level applications

In order to evaluate coreference systems, their output is mapped onto a manual annotation of coreference in a gold standard. The difference between the manual annotation (the key) and the system output (the response) is quantified based on a distance metric.

The commonly used evaluation framework for coreference resolution was introduced in Denis and Baldridge (2009) and adapted by the CoNLL shared tasks (Pradhan et al., 2011, 2012). It consists of applying five acknowledged coreference metrics and then averaging the F-scores of three of them to determine the overall best performing system. We briefly overview the five metrics, but not extensively, since they all share a common principle which makes it arguably difficult to interpret them from the view of downstream applications.

- **MUC**: The MUC metric (Vilain et al., 1995) compares links between mentions in the key chains to links in the response chains. Recall is calculated based on the number of links that have to be inserted into the response to obtain a key chain, and Precision is determined by the number of links that have to be deleted in a response chain to obtain a key chain.
- **B-CUBED**: For Recall, B-CUBED (Bagga and Baldwin, 1998) evaluates each mention in the key by mapping it to one of the mentions in the response and then measuring the amount of overlapping mentions in the corresponding key and response chains. To calculate Precision, the role of key and response are inverted.
- **CEAFM/CEAFE**: The CEAF metrics (Luo, 2005) are based on the B-CUBED metric but impose restrictions on the alignment of the key and response chains to enable a more coherent comparison. CEAFM is a mention-centric measure, while CEAFE evaluates entities.
- **BLANC**: The BLANC metric (Recasens and Hovy, 2010) uses the Rand Index clustering algorithm to align key and response chains and then evaluates coreference and non-coreference links.

The common principle shared by all these metrics is that they view coreference chains as unordered sets of generic items (Chen and Ng, 2013) and approach evaluation as a clustering evaluation task. This principle is questionable, since a) coreference chains are not unordered. The sequence of the mentions in a chain is established by the occurrence sequence of the mentions in discourse and therefore follows a natural, linear structure. And b) mentions are not generic items, but linguistic objects with linguistic properties, such as PoS, syntactic label etc.

Given this arguably problematic principle, three main issues arise from the perspective of downstream applications:

- **Interpretability**: The meaning of the metrics cannot easily be conveyed in natural language, because of the rather complex algorithms used to determine the scores. For example, how can a 56.78% Recall in BLANC be interpreted for a Sentiment Analysis system? Additionally, averaging F-scores further obscures the meaning of the scores.

- **Informativeness**: As the metrics do not differentiate mention types, e.g. based on part-of-speech tags of the mentions, we cannot gain more fine-grained insight into why a system performs better than another. That is, does one system resolve pronouns better than the other, or does it simply feature a more precise markable boundary identification approach? Which system resolves noun mentions better?
- **Differentiability**: The metrics quantify the goodness of each system response. However, they cannot tell us how different two given system outputs are or how much they overlap. We do not know whether two systems with very similar F-scores resolve the same set of mentions or resolve completely different mentions. For example, given a set of mentions $[A - B - C - D - E]$, one system might correctly produce the links between mentions $[A - B - C]$, while another system might produce the links $[C - D - E]$. The MUC metric, for instance, will rank both responses equally, because each of them finds two out of four coreference links. However, the outputs do not overlap at all. On the contrary, they are maximally dissimilar.

Another problem is that the metrics show discrepancies regarding the ranking of systems (Holen, 2013), i.e. they have been shown to produce different rankings over the same set of system responses. To handle this problem, the CoNLL shared tasks took the average of MUC (a link-based metric), B-CUBED (mention-based), and CEAFE (entity-based)[1] to determine the winner.

3.2 The ARCS evaluation framework for coreference resolution

Given the issues in coreference evaluation discussed in the previous section, we proposed the **A**pplication-**R**elated **C**oreference **S**cores (ARCS) in Tuggener (2014). It is important to note that we do no claim to have presented a solution that solves all issues in coreference evaluation. Neither are our proposed metrics intended to replace the common evaluation framework. Our proposition is, as an alternative to the common coreference evaluation, to abandon the idea of evaluating systems in a general or universal setting and to look at them from the perspective of downstream applications and then provide measures that are interpretable and relevant from their view.

[1] The so-called MELA (**M**ention, **E**ntity, and **L**ink **A**verage) metric originally introduced by Denis and Baldridge (2009).

The main goal of the ARCS metrics is to address the three issues of the common coreference metrics we outlined in the previous section: **interpretability**, **informativeness**, and **differentiability**.

In order to achieve **interpretability**, we keep the commonly known nomenclature used in Information Retrieval from where coreference evaluation borrows its terms. That is, we use Recall and Precision, True Positives, False Positives and so forth as the basis of our scores.

In Information Retrieval, Recall denotes how many of the relevant documents are retrieved by a system, i.e. $Recall = \frac{TP}{TP+FN}$. Thus, the perspective is that of the gold standard. Precision, in turn, is the perspective of the system, i.e. how many of the retrieved documents are relevant, $Precision = \frac{TP}{TP+FP}$. To adapt the nomenclature and its meaning to coreference, we want Recall to denote how many of the coreferent mentions in the gold standard are identified and resolved correctly by a system, and Precision to signify how many of the mentions resolved by the system overall are correct. Thus, if a system resolves many of the gold mentions correctly, Recall is high, and if a system does not hallucinate or invent too many mentions[2], Precision is high.

To achieve these definitions of Recall and Precision, we adapt the common classes TP, FP, TN, FN from Information Retrieval in a straight-forward fashion to coreference.

- **TP**: True positive; a gold mention correctly resolved by a system.
- **FP**: False positive; a markable resolved by a system that is not annotated as coreferent in the gold standard.
- **FN**: False negative; a gold mention not resolved by a system.
- **TN**: True negative; a markable not resolved by a system that does not have coreference annotation in the key[3].

However, we need a novel error class for those cases where a system correctly decides to resolve a gold mention, but attaches it to an incorrect antecedent candidate. These cases are not included in the inventory above. We name this error class $Wrong\ Linkage$ (WL) and then define $Recall = \frac{TP}{TP+FN+\mathbf{WL}}$ and $Precision = \frac{TP}{TP+FP+\mathbf{WL}}$. The

[2] I.e. so-called spurious or twinless system mentions, cf. section 1.5.4

[3] Note that we do not use the TN class, which would signify singletons, i.e. NPs not partaking in coreference resolution and which a system (correctly) does not resolve. Whether to include singletons in coreference evaluation has been disputed and shown to greatly affect evaluation (Kübler and Zhekova, 2011). The argument for including singletons is that coreference systems should be rewarded for not resolving non-coreferent markables. Our view is that downstream applications have no use for singletons and therefore systems should not be rewarded for identifying them. We thus discard the TN class.

Recall denominator extends over all mentions in the key, and the Precision denominator denotes all mentions in the response. Recall now has the desired meaning of *how many of the gold mentions are resolved correctly* and Precision accurately signifies *how many of the response mentions are correct.* We believe these definitions to be straight-forward enough to serve **interpretability**.

Note that, like Chen and Ng (2013), we have adapted a link-based view to evaluate coreference. For each mention, we examine how it is linked to an antecedent and classify it accordingly. This comes with an additional benefit. Since we investigate each mention individually, we have access to all its linguistic properties and are able to evaluate systems with regards to any of these properties. For example, we can assess system performance regarding the PoS tags of the mentions and compare systems by their pronoun and noun resolution strengths. Or we can diversify further and analyze how well different systems resolve different pronoun lemmas of a particular pronoun type, as demonstrated in Tuggener and Klenner (2014).

Another benefit of introducing the wrong linkage class is that we can analyze in more detail whether a system performs well because it features a strong mention resolution strategy (by analyzing the true positive and wrong linkage distributions) or because it is effective in determining anaphoricity of mentions (by looking at the false negative and false positive classes counts).[4] Given these possibilities, our evaluation provides different levels of **informativeness**.

Having outlined the nomenclature, the crucial task remains to determine what “resolved correctly” means regarding a gold mention. The question is synonymous to asking what is a correct antecedent. In Tuggener (2014), we clustered different applications that benefit from coreference resolution into three groups, where each application within a group has similar requirements regarding the definition of a correct antecedent. At this point, we refer to the paper for more details on the application groups. We note, however, that all our scores take into account the linear order of the occurrence of the mentions in discourse. That is, one score requires mentions to link to their immediate antecedent in the key chain (ARCS immediate antecedent), one metric requires that the closest nominal antecedent of mentions is correct (ARCS inferred antecedent), and one score requires all mentions to link to the first nominal mention in a chain (ARCS anchor mention). The bare-bone algorithm for comparing a key to a response within ARCS is outlined in the next section (algorithm 2), where we focus on the last point of interest mentioned above, namely **differentiability**.

[4] We exemplify this method of analysis in the next section 3.2.1.

3.2.1 Quantification of the difference between system responses

The CoNLL shared tasks revealed that many of the top ranked systems score similarly, despite deploying profoundly different approaches to coreference resolution. For example, the system that ranked third in 2011 only scored 0.03 points lower in average F-score compared to the second best system. It also performed best on the BLANC and B-CUBED metric, but, since BLANC was not included in the average F-score, the system ranked lower than the first two.

Given that the top ranking systems score very similarly on the F-score average, an interesting question is whether these systems actually resolve a similar set of markables. The performance difference then would stem from one system resolving slightly more of the commonly resolved mentions correctly. Alternatively, the systems generally resolve rather different sets of mentions, and the best performing system resolves more mentions correctly. To assess this, we propose an approach to quantify the overlap of two (and potentially more) system responses, which complements the common evaluation.

Using the class inventory introduced in the previous section, we compare two system responses S_1 and S_2 to a key K. We classify each gold mention m_i of a gold entity $e_k \in K$ in S_1 and S_2 and determine if its classification $c(m_i) \in \{TP, FN, FP, WL\}$ is the same in both responses. To classify m_i, here, we assess whether m_i in S_1 or S_2 shares an antecedent with m_i in K, respectively.[5] Additionally, if m_i denotes a false positive in S_1, we check whether it is also a false positive in S_2, and vice versa, because the overlap of spurious system mentions should also be considered in a similarity estimate of system responses.

We first outline how mentions are classified in the ARCS framework given a key and one system response and then extend it to two responses. Algorithm 2 presents the algorithm for classifying the mentions given a key K and a system response S. Neglecting the linguistic subtleties[6], we call mentions anaphoric if they are not the first mention of a chain. Note further that we only evaluate anaphoric mentions (anaphoric as in the definition just stated), as indicated by lines 2 and 10. We use the false negative class to label key mentions that are either not in the response (line 4) or which are the first mention in the response but not in the key (line 5). Here, as stated above, we only

[5] We realize that this can be argued to be a weak criterion for evaluating mentions in coreference sets. However, for our purposes of comparison here, we argue that it is sufficient because we do not aim at establishing the "goodness" of the systems, but at quantifying the differences of their outputs. Also, it is a method often adapted in evaluating pronoun resolution within coreference in a pair-wise fashion (cf. section 3.4). Note that we could require any criterion presented in the ARCS framework regarding the correctness of the antecedent.

[6] cf. section 1.4.1.3

Algorithm 2 ARCS mention classification algorithm

```
**Input:** Key $K$, System response $S$
**Output:** Mention class $c_{m_i} \forall m \in K \cup S$
 1: for $e_k \in K$ do
 2:     for $m_i \in e_k$ where $index(m_i, e_k) > 0$ do
 3:         $e_s \leftarrow get\_entity(m_i, S)$
 4:         if $e_s == \varnothing$ then $c_{m_i} \leftarrow FN$
 5:         if $index(m_i, e_s) == 0$ then $c_{m_i} \leftarrow FN$
 6:         if $\exists m_j \in e_s : j < i \wedge m_j \in e_k$ then $c_{m_i} \leftarrow TP$
 7:         else $c_{m_i} \leftarrow WL$
 8:         return $c_{m_i}$
 9: for $e_s \in S$ do
10:     for $m_i \in e_s$ where $index(m_i, e_s) > 0$ do
11:         $e_k \leftarrow get\_entity(m_i, K)$
12:         if $e_k == \varnothing$ then $c_{m_i} \leftarrow FP$
13:         if $index(m_i, e_k) == 0$ then $c_{m_i} \leftarrow FP$
14:         else $c_{m_i} \leftarrow TN$
15:         return $c_{m_i}$
```

require that the response chain of a mention contains an antecedent that is also in the key chain in order for the mention to be counted as a true positive (line 6).

We also want to include system mentions in the comparison, i.e. markables resolved by the system that are not in the key (false positives). To do so, we iterate over all system mentions and see if they are in the key (lines 11-12) or if their index in the key chain is zero, which means that the mention is the chain starter in the key. In this case, the system has resolved the markable to an antecedent and deemed it anaphoric, and therefore we count it as a false positive.

When comparing a key to two system responses, i.e. S_1 and S_2, the difference in the algorithm consists of processing lines 1-8, i.e. the key mention classification, for both responses and recording whether the classification of m_i differs in S_1 and S_2. Furthermore, lines 9-15 are applied to both responses to identify shared and non-shared false positives.

Finally, we calculate the difference of S_1 and S_2 as the percentage of mentions that are classified differently in S_1 and S_2, where the set of considered mentions is the union of all mentions in the K, S_1, and S_2:

$$diff(S_1, S_2) = \frac{|\forall m_i \in \{K \cup S_1 \cup S_2\} \text{ where } c_{m_i} \text{ in } S_1 \neq c_{m_i} \text{ in } S_2|}{|\forall m_i \in \{K \cup S_1 \cup S_2\}|} \tag{3.1}$$

We apply our difference metric to four publicly available coreference systems and their responses for the CoNLL 2012 shared task test to demonstrate the use of the metric. The systems are the following:

- **Stanford CoreNLP** (Lee et al., 2013) is a rule-based (or rather sieve-based) coreference system that won the CoNLL shared task 2011 and which is probably the most widely used and popular coreference system for English to date.
- **IMSCoref** (Björkelund and Farkas, 2012) ranked 2nd in the CoNLL shared task 2012 and uses resolver stacking in a machine learning setting.
- **Berkeley Coreference** (Durrett and Klein, 2013) achieved a substantial performance improvement over the state-of-the art in a CoNLL 2012 post task evaluation. One of its novelties was the heavily lexicalized feature set.
- **HOTCoref** (Björkelund and Kuhn, 2014) is, to our knowledge, the best performing freely available coreference resolution system for English.[7] It promotes the view of coreference sets as trees and models coreference as a structure prediction problem. Also, it makes use of latent antecedent features.[8]

Table 3.1 reports the performance of the systems using the reference scorer[9] (Pradhan et al., 2014) for the commonly used metrics. We apply our difference metric to the systems in a pair-wise fashion, i.e. one response acts as S_1 and another as S_2. The overview is given in table 3.2.

The first surprising observation we make is that the system outputs actually seem to be quite different, i.e. the percentage of mentions classified differently is high. This magnitude of difference is not present in the evaluation of the systems using the common metrics in table 3.1. For example, the HOTCoref system outperforms the Berkeley system by 2.69 points in average F-score (64.31 vs. 61.62), but the systems process 28.15% of the mentions in the pool ($\{K \cup S_1 \cup S_2\}$) differently. We further see that the best and second best performing systems are most similar (BERK vs. HOTCoref; 28.15% difference), while the best and worst performing system are most different (STAN vs. HOTCoref; 37.60% difference).

We are also able to measure system differences regarding certain properties of the mentions, such as PoS tags. An interesting question is how different the systems are w.r.t.

[7] However, the Stanford and Berkeley systems are the only two that can be run on raw text, i.e. IMS and HOTCoref require preprocessed and preformatted input, while Stanford and Berkeley are real end-to-end coreference resolution systems.

[8] Given that an antecedent candidate of a mention is already part of a coreference set, other mentions of that chain can be accessed for generating features.

[9] `http://conll.github.io/reference-coreference-scorers/` Note that the scores differ from those on the website of the CoNLL shared tasks because of the changes to the scorer.

the resolution of noun markables compared to resolving pronouns. We observe that the systems differ more regarding the noun mentions (table 3.3) than compared to pronouns (table 3.4). For example, the HOTCoref and Berkeley systems process 33.81% of the noun mentions differently, while they only differ in 22.44% of the pronoun instances.

	MUC	B^3	CEAFE	$F\phi$
STAN	64.81	52.74	49.51	55.69
IMS	67.58	54.47	50.21	57.42
BERK	71.19	58.29	55.39	61.62
HOTCoref	**73.29**	**61.36**	**58.30**	**64.31**

Table 3.1: F-scores of coreference systems on the CoNLL 2012 test set.

	IMS	BERK	HOTCoref
STAN	35.20	36.31	**37.60**
IMS		36.09	33.27
BERK			**28.15**

Table 3.2: Percentages of all mentions classified differently.

	IMS	BERK	HOTCoref
STAN	41.78	43.87	43.24
IMS		**45.28**	41.84
BERK			**33.81**

Table 3.3: Percentages of nominal mentions classified differently.

	IMS	BERK	HOTCoref
STAN	28.16	28.47	**31.94**
IMS		26.10	24.14
BERK			**22.44**

Table 3.4: Percentages of pronouns classified differently.

Overall, we conclude that the analysed systems for coreference resolution in English differ more strongly than the averaged F-scores would suggest w.r.t. how they process the mentions.

Given our inventory of mentions classifications, we are also able to understand in what regard two systems differ, and, in turn, why one system performs better than another. To do so, we analyse the composition of the difference between two systems, i.e. we investigate the mentions processed differently. We track each mention and its specific classification in S_1 and S_2, respectively. For example, we count how often a mention that is classified as wrong linkage (WL) in S_1 is classified as true positive (TP) in S_2, i.e. the count of $WL \rightarrow TP$ transitions from S_1 to S_2. If we encounter many of these $WL \rightarrow TP$ transitions, we conclude that S_2 performs better than S_1 because it features a better strategy to link mentions to the correct antecedents. On the other hand, if we encounter many false positive to true negative transitions, $FP \rightarrow TN$, we conclude that S_2 performs better because it is superior in identifying which markables to resolve, i.e. it produces fewer spurious mentions (false positives), etc.

We apply this comparison to the best and worst performing systems (HOTCoref vs. Stanford) in table 3.5 and to the best and second best performing systems (Berkeley vs. HOTCoref) in table 3.6, respectively.[10] The transition column (**Trans.**) indicates the different types of mention classifications, e.g. how many mentions that are classified as

[10] Note that the overall counts of pronouns and nominal mentions is not the same in the two tables because the mention pool $\{K \cup S_1 \cup S_2\}$ is not the same for the two comparisons, based on the different spurious system mentions.

wrong linkage in the Stanford response become true positives in the HOTCoref response (wl→tp) etc. The transitions are divided into corrections (↑) and errors (↓). Note that not all transitions in the error category render correct decisions incorrect. Some transitions turn one error into another, e.g. fn→wl. The last column (**% Trans.**) shows the percentage of a transition given all transitions for pronoun and noun mentions, respectively.

-/↑/↓	**Trans.**	**Count**	**% Trans.**
***Pronouns**: 9158 mentions* ***Changed**: 2925 (31.94%)*			
↑	**wl→tp**	**736**	**25.16%**
	fp→tn	**726**	**24.82%**
	fn→tp	339	11.59%
% of changes: 61.57%			
↓	tp→fn	375	12.82%
	tp→wl	324	11.08%
	wl→fn	162	5.54%
	tn→fp	148	5.06%
	fn→wl	115	3.93%
% of changes: 38.43%			
***Nominal mentions**: 9203 mentions* ***Changed**: 3979 (43.24%)*			
↑	**fp→tn**	**1452**	**36.49%**
	fn→tp	789	19.83%
	wl→tp	228	5.73%
% of changes: 62.05%			
↓	tn→fp	561	14.10%
	tp→fn	389	9.78%
	wl→fn	246	6.18%
	fn→wl	209	5.25%
	tp→wl	105	2.64%
% of changes: 37.95%			

Table 3.5: Mention classification transitions when comparing Stanford → HOTCoref.

-/↑/↓	**Trans.**	**Count**	**% Trans.**
***Pronouns**: 8842 mentions* ***Changed**: 1984 (22.44%)*			
↑	fp→tn	410	20.67%
	wl→tp	355	17.89%
	fn→tp	210	10.58%
% of changes: 49.14%			
↓	tp→wl	360	18.15%
	tp→fn	349	17.59%
	tn→fp	114	5.75%
	wl→fn	110	5.54%
	fn→wl	76	3.83%
% of changes: 50.86%			
***Nominal mentions**: 8935 mentions* ***Changed**: 3021 (33.81%)*			
↑	**fp→tn**	**1254**	**41.51%**
	fn→tp	372	12.31%
	wl→tp	151	5.00%
% of changes: 58.82%			
↓	tp→fn	434	14.37%
	tn→fp	412	13.64%
	wl→fn	140	4.63%
	tp→wl	136	4.50%
	fn→wl	122	4.04%
% of changes: 41.18%			

Table 3.6: Mention classification transitions when comparing Berkeley → HOTCoref.

Table 3.5 shows that the HOTCoref response introduces more corrections than errors for pronouns (↑ 61.57% of changes; ↓ 38.43% of changes) and for nominal mentions (↑ 62.05% of changes; ↓ 37.95% of changes) given the Stanford response. That is, the ratio for introducing corrections compared to errors is roughly the same for nominal mentions and pronouns. However, the distributions of transitions differ between nominal mentions and pronouns. The majority of transitions that HOTCoref introduces w.r.t. pronouns is wrongly linked pronouns to true positives (wl→tp; 25.16%), closely followed by turning false positives into true negatives (fp→tn; 24.82%). This indicates that HOTCoref both performs better in resolving pronouns to a correct antecedent and in determining the anaphoricity status of pronouns (which is especially essential for the pronoun *it*). Regarding the nominal mentions, we observe that HOTCoref corrects many of the false positive mentions (fp→tn; 36.49%) that Stanford produces. To a smaller degree, it is able to resolve nouns that Stanford does not recognize to be coreferent (fn→tp; 19.83%).

In table 3.6, which compares the two best performing systems, we observe that the HOTCoref response corrects about the same amount of pronoun resolutions of the Berkeley system as it renders incorrect (↑ 49.14%; ↓ 50.85%). This indicates that HOTCoref does not outperform the Berkeley system by a wide margin w.r.t. pronouns. To substantiate this observation, we score the performance of the systems in pronoun resolution using the same criterion as for estimating their differences, i.e. if the response chain of a pronoun contains an antecedent present in the key chain, it its counted as correct. Table 3.7 gives the results and the counts of the mention classifications.

	Rec.	**Prec.**	**F1**	**Acc.**	**TP**	**WL**	**FN**	**FP**
Personal pronouns (PRP; 6166 mentions)								
STAN	74.02	68.98	71.41	81.18	4564	1058	544	994
IMS	77.07	76.61	76.84	**86.86**	4752	**719**	695	732
BERK	**79.52**	76.24	77.84	86.20	**4903**	785	**478**	743
HOTCoref	77.70	**79.96**	**78.81**	86.64	4791	739	636	**462**
Possessive pronouns (PRP$; 1721 mentions)								
STAN	72.23	71.77	72.00	77.54	1243	360	118	129
IMS	79.78	79.27	79.53	86.73	1373	210	138	149
BERK	**82.74**	**82.55**	**82.65**	**87.52**	**1424**	**203**	**94**	98
HOTCoref	80.88	82.12	81.50	86.35	1392	220	109	**83**

TABLE 3.7: Pronoun resolution performance of state-of-the-art systems for English.

We see that, excluding the Stanford system, performance does not vary strongly overall. The difference between HOTCoref and the Berkeley system mainly stems form the comparably low false positive counts of HOTCoref (PRP 462; PRP$ 83) which gives it high Precision. On the other hand, the Berkeley system has the highest Recall for both pronoun types because of its low false negative and high true positive counts. We also see that the two systems do not vary strongly regarding the wrong linkage counts, which indicates that they perform similarly well when identifying antecedents of anaphoric pronouns. This is substantiated by Accuracy (Acc.), which is calculated by $\frac{TP}{TP + WL}$, i.e. where we cancel out the anaphoricity detection problem to determine the resolution performance on gold pronouns.[11] We see that the Accuracy of the top three systems varies very little for both pronoun types, which further indicates that the differences in F-score stem from anaphoricity detection to a substantial margin.

Going back to the comparison of HOTCoref and Berkely in table 3.6, we see that the main differences between the two systems regarding nominal mentions again stems from the anaphoricity identification performance. That is, the difference is caused by the fp→tn transition to a large degree (41.51% of all changes regarding the nominal mentions).

[11] Section 3.4.3 on pronoun resolution evaluation will expand on this metric in more detail.

In conclusion, our comparisons indicate that the best performing system, HOTCoref, outperforms the others mainly due to its superior ability to determine the anaphoricity status of mentions, rather than performing particularly better in identifying correct antecedents of gold mentions.

3.2.2 Assessment of feature potential

Apart from using the difference metric to quantify the difference of two coreference systems, as presented in the previous section, the metric can also be applied to quantify the change introduced by a feature when performing the task of feature engineering for a single system.

Common metrics measure the overall improvement or degradation of performance when features are introduced into a feature set, e.g. in a feature ablation experiment. However, the common evaluation process does not quantify the amount of change that the addition of a feature induces in the output. For example, two new features might increase performance by a similar but small margin. However, one feature reverts many correct decisions from the previous feature set and at the same time corrects many faulty previous decisions. The other new feature, by contrast, only affects few previous decisions and introduces only few new correct ones. But as the ratio of correct and wrong decisions stays the same overall, the performance metrics do not reflect the different amount of change in the system outputs. In other words, one might argue that the common evaluation framework, which consist of observing changes in Recall and Precision, only looks at the proverbial tip of the iceberg when comparing two system outputs.

Obviously, improving performance is the ultimate goal when introducing novel features. But a feature that introduces many new correct decisions while simultaneously introducing many faulty decisions indicates that the feature is worth investigating further, as opposed to a feature that only marginally changes system output or only introduces errors. The difference metric introduced above can directly be applied to a baseline response and a modified feature set response to monitor change.

3.2.3 Comparison of multiple system responses

Using our overlap approach, we are able to compare several system responses and identify easy or hard to resolve gold mentions. That is, we look for gold mentions that are classified as true positives (TP) or wrong linkages (WL) in all responses to identify mentions that are easy or hard to resolve, or mentions with false negative (FN) or false positive (FP) classifications in all responses, which indicates mentions where the anaphoricity detection proves difficult or might even indicate errors in the gold annotation (e.g. if all systems indicate that a mention should be coreferent). Table 3.8 presents a selection of such mentions from the CoNLL 2012 test set.

Relation	**Antecedent - Mention**	**Mention class**
	Easy cases	
Str. match	*[the stock reform] ...* ***[this stock reform]***	True positive
Str. match	*[the accident] ...* ***[the accident]***	True positive
Easy PRP	*[the friend] first sent me an SMS, Uh-huh. saying* ***[he]*** *would ...*	True positive
Easy PRP	*[this emergency repair worker] said that* ***[he]*** *was there ...*	True positive
Easy PRP$	*thanking [citizens] for* ***[their]*** *cooperation...*	True positive
	Hard cases	
Cataphora	*To express [its] determination,* ***[the Chinese securities regulatory department]*** *...*	False negative
Cataphora	*Thank [you]* ***[everyone]*** *...*	False negative
Nominal	*[Focus Today] ...* ***[our program]*** *...*	False negative
Nominal	*[a road cave-in accident that happened in Beijing over the holiday] ...* ***[the road caving in]*** *...*	False negative
Date	*[January 3] ...* ***[the day of the accident]*** *...*	False negative
Date	*[the day of the accident] ...* ***[yesterday]*** *...*	False negative
Count	*[the two honorable guests] ...* ***[both of you]*** *...*	False negative
Count	*[both of you] ...* ***[the two of you]*** *...*	False negative
Dir. Speech	*[Yang Yang, a host of Beijing Traffic Radio Station] ...* ***[you]*** *...*	Wrong linkage
Hard PRP	*[an SMS like this one] ...* ***[it]*** *did not give people ...*	Wrong linkage
Orth.	*[Chaoyang Road] ...* ***[this road]*** *...*	Wrong linkage
Discourse	*[the neighborhood] ...* ***[this place]*** *...*	Wrong linkage

Table 3.8: Examples of easy and hard to resolve mentions identified by the comparison of multiple system responses.

While the examples of mentions classified as true positive (TP) and false negative (FN) are self-explanatory (TP means all systems resolve the mentions correctly; FN indicates that none of the systems resolves the mentions), the selected WL mentions at the bottom of the table need elaboration. In the direct speech example (Dir. speech), all systems link *you* to another preceding *you* pronoun and fail to catch the switch of addressee. For the *it* pronoun in the difficult case (Hard PRP), the systems fail to correctly determine the mention boundaries of the antecedent. HOTCoref and Berkeley chose *[this one]* as the antecedent span, while Stanford and IMS mark *[an SMS]* as antecedent. In the wrong linkages related to the *Chaoyang Road* mention, all system link *this road* to an NP further back which string matches perfectly, overlooking the partially matching antecedent which contains an uppercase version of the NP head. The last example stems from a similar problem, i.e. *this place* is linked to a perfectly matching NP preceding the antecedent.

In contrast to the error analysis approach presented in Kummerfeld and Klein (2013) which categorizes and quantifies errors on the entity level, our approach is able to identify and extract specific mentions that are resolved incorrectly by a set of systems. Given these mentions, new directions in coreference resolution might be identified. Although far from being quantitatively substantiated, table 3.8 indicates that all systems seem to have difficulties with mentions denoting counts of entities, and date or time related mentions. We leave such systematic explorations to future work.

In the comparative system output analysis, we are able to calculate an upper bound for Recall under the assumption of perfect system combination. Given that the differences between systems proved to be large, we would hope that combining their output increases performance compared to the individual responses. To do so, we assume an oracle that picks for each mention in the key among the four system responses one that resolves it correctly. That is, we count each gold mention as a true positive if at least one system has resolved it correctly. Again using the antecedent criterion above (i.e. key and response chain for a mention need to share at least one antecedent), we calculate Recall for the best performing system HOTCoref by dividing the number of true positives by the number of all key mentions, which gives us 10'620/15'232, i.e. 69.72%. If we use the oracle approach, we reach a Recall of 12'540/15'232, i.e. 82.33%. The large difference of 12.61 percentage points i) suggest that system output combination seems to be a fruitful direction to pursue for future work, and ii) is another indicator for the difference of how each mention is processed by the systems, and that their output is complementary to a certain degree.

3.3 Error analysis in coreference resolution

Another view on evaluation of coreference resolution is error analysis. Instead of rewarding correctly resolved mentions, error analysis focuses on the errors a system makes and tries to group and quantify them. As Kummerfeld and Klein (2013) noted, most papers on coreference resolution that performed error analysis did so by looking at erroneous pair-wise decisions and by discussing a handful of examples. Here, we overview error classification frameworks that feature a systematic approach and discuss their implications for higher-level applications.

Uryupina (2007, 2008) conducted error analysis on a single system and proposed three main error classes:

1. Gold standard annotation errors
2. Errors propagated from preprocessing
3. Markable linking errors

Errors were classified manually in her approach and then quantified to provide an overview of the error type distribution for her system. This analysis helps to identify problems in the gold standard annotation, in the preprocessing tools, and in the mention resolution strategy of the coreference system at hand. However, due to the amount of involved manual labour, it is a rather unpractical approach for comparative error analysis on a larger scale. Also, several of the issues addresses in this approach can (by now) be analyzed by other means. For evaluating the impact of the preprocessing tools, gold standard annotations can be used to assume perfect preprocessing. System performance degradation can then be directly quantified when using real preprocessing, as e.g. in Klenner et al. (2010). This approach was also chosen in the CoNLL shared tasks (Pradhan et al., 2011, 2012) (and earlier by SemEval (Recasens et al., 2010)) to quantify the impact of imperfect preprocessing information. Automatically identifying errors in the gold standard annotation is more tricky. In section 3.2.3 we outlined an approach to identify them using an ensemble of coreference resolution systems. In section 3.4.3, we will show how evaluating local decisions can avoid annotation problems by only considering cases where the system is actually able to make the correct decision. Finally, for analyzing problems in the mention resolution strategy, we have proposed the ARCS evaluation framework in section 3.2, which, although conceptualized as scores, can be used for error analysis and tracking.

Holen (2013) introduced an error analysis inventory for evaluating system outputs on the entity level (i.e. coreference chains). For a small number of documents and sentences, she analyzed how often 1) key and response entities match perfectly, 2) response entities are partial or response mentions have NP boundaries that are too short, 3) response entities merge multiple key entities into one (i.e. conflate two coreference chains), and 4) no response entity is given or the entity is missing an informative mention (where e.g. a full NP containing a proper name is more informative than a pronoun). The downside of this approach is, again, that it requires manual labor by experts.

Kummerfeld and Klein (2013) addressed this issue and presented an automated framework for analyzing and comparing system responses on a large scale. They adapted the error inventory put forward by Holen (2013). More specifically, they investigated and quantified the kind of transformations (split entity; spurious mentions etc.) which the entities in a system response need to undergo in order to produce the key entities.

Finally, the work in Martschat and Strube (2014, 2015) presented coreference error analysis based on a tree-structure view on coreference sets. Entities are represented as spanning trees (i.e. a graph) and Recall and Precision errors are identified by missing and spurious links in the key and response graphs. A particular strength of this error analysis framework is that it is able to categorize Recall errors of noun resolution into e.g. name-noun or name-name pairs of antecedents and anaphors and compare system responses based on the distributions of errors regarding these different pairs. Also, the approach quantifies the number of links that one system can find in comparison to another system, and it counts e.g. how many Precision errors are shared among the systems, which paves the road to identifying mentions that proof to be especially difficult to resolve and therefore deserve special attention.

Considering the publications cited above, one might argue that systematic error analysis for coreference resolution has gained popularity in the last years. We argue that this is an important development, since error analysis complements system rankings produced by shared tasks. Analysis of errors enables a more detailed view on the system outputs and is able to unveil strengths and weaknesses of particular systems and to identify common errors, which in turn leads to new research directions.

3.4 Evaluation of pronoun resolution

In the previous sections of this chapter, we have discussed evaluation of coreference sets and coreference partitions. In this section, we focus on how pronoun resolution is evaluated, both within coreference chains and as a separate task in a pair-wise fashion.

3.4.1 Ratio-based evaluation

In the pioneering era of research in automated approaches to pronoun resolution, coreference set-level information was non-existent in corpora as an annotation layer. Corpora and data sets first had to be created for the evaluation of pronoun resolution approaches. These annotations generally provided pairs that consisted of a pronoun and its antecedent within a local context. Evaluation then quantified how often the antecedent of such a pronoun was identified correctly by a system. Hobbs (1978), who presented one of the first automatable algorithms for third person pronoun resolution, calculated the success of his approach by the percentage of correctly resolved pronouns in the following way:

$$Success = \frac{|successfully\ resolved\ pronouns|}{|all\ attempted\ resolutions|}$$

Lappin and Leass (1994), another seminal approach to third person pronoun resolution, adapted this measure, followed by many others. The downside of this ratio is that it does not account for pronouns that the system does not attempt to resolve. Also, pronouns that the system resolves but which are not in the gold standard are also discarded.

3.4.2 Introduction of Recall and Precision

An important extension to pronoun resolution evaluation was the introduction of Recall and Precision (and their harmonic mean, the F-score). Recall analyses a system output from the gold standard perspective and quantifies how many of the manually annotated pronouns are resolved correctly. Precision takes the view of the system output and counts how many of all the resolved pronouns are correct. This division introduced the problem of anaphoricity detection (i.e. whether a pronoun should be resolved at all), since typically not all pronouns resolved by a system are in the gold standard. Vice versa, not all manually annotated pronouns are resolved by a system.

Aone and Bennett (1995) defined Recall and Precision in the following manner:

$$Recall = \frac{|correct\ resolutions|}{|pronouns\ identified\ by\ the\ system|}, Precision = \frac{|correct\ resolutions|}{|attempted\ resolutions|}$$

which implies that their system did not attempt to resolve all pronouns that it identified. The denominator in the Recall equation thus does not necessarily extend over all manually annotated pronouns in the gold standard. Baldwin (1997) used the same Precision measure but extended Recall over all gold pronouns:

$$Recall = \frac{|correctly\ resolved\ anaphors|}{|anaphors\ in\ the\ gold\ standard|}$$

Mitkov (2001) critically reflected the evaluation practices in pronoun resolution. He proposed a more fine-grained analyses of system behaviour. His main proposition was to divide evaluation in scoring anaphora resolution *algorithms* and scoring anaphora resolution *systems*. When evaluating the former, evaluation should be freed from noise such as preprocessing errors. The task then is to determine which resolution algorithm best solves the problem of pronoun resolution in a theoretical, idealistic setting. The later aims at determining which of a set of anaphora resolution systems performs best in a real-world setting, i.e. including all subtasks of pronoun resolution pipelines, such as PoS tagging and syntactic parsing etc. More recent shared tasks on coreference resolution (Recasens et al., 2010, Pradhan et al., 2011, 2012) tried to address the comparison of resolution strategies and full systems by running a gold mention and gold boundary setting where systems only resolve the gold mentions provided by the gold standard or are given perfect markable boundaries, respectively.

Mitkov (2001) argued for the use of *success rate* which is calculated by:

$$Success\ rate = \frac{|correctly\ resolved\ anaphors|}{|all\ anaphors\ in\ the\ gold\ standard|}$$

i.e. it is identical to the Recall definition in Baldwin (1997). Mitkov (2001) also introduced the *critical success rate* which only evaluates instances of anaphors that are left with multiple antecedent candidates after gender-based filtering of all potential candidates. That is, resolution strategies were only rewarded for making the right choice when they actually had to make one.

Additionally, Mitkov made a strong point for evaluating resolution strategies against simple baseline, i.e. choosing the most recent candidate or the most recent subject candidate as the antecedent to assess the improvements made by a more elaborate resolution approach.

Another important criticism of pronoun resolution evaluation, to which our view based on downstream applications neatly aligns, was presented by Stuckardt (2001) and later Müller (2008). Their main claim was that pronouns that are only resolved to other pronouns, i.e. do not (transitively) link to a correct nominal antecedent, should not be rewarded in evaluation, because they provide no functional benefit for higher-level applications. For example, given that two systems produce the following responses for the given key:

- Key: $[Noun_x - Pronoun1 - Pronoun2 - Pronoun3]$
- Sys1: $[Pronoun1 - Pronoun2 - Pronoun3]$ (i.e. omitting the link to the noun)
- Sys2: $[Noun_y - Pronoun1 - Pronoun2 - Pronoun3]$ (i.e. linking the pronouns to the wrong noun)

Stuckardt and Müller argued that neither system should be rewarded for correctly resolving Pronouns 2 and 3 to Pronoun 1 because they do not transitively link to a correct nominal antecedent. In Tuggener (2014), we argued that these constraints are not necessarily applicable to all kind of downstream applications, but that the required type of antecedent depends on the type of the intended downstream application.

Müller (2008) then defined Recall and Precision identical to Aone and Bennett (1995), with the difference that "correctly resolved" and "resolvable" imply that pronouns directly or transitively link to non-pronominal antecedents. We proposed the *ARCS inferred antecedent* metric in order to achieve what Müller encompassed and extended it with the wrong linkage (WL) class, which enables a more fine-grained analysis of the false positive class. Stuckardt introduced an even more fine-grained distinction of errors w.r.t. a pair of a pronoun P and its non-pronominal antecedent A, given in table 3.9, accompanied by the ARCS interpretation of the pair classes.

Class	**Description**	**ARCS**
o_{++}	P and A belong to the same key equivalence class	TP
o_{+-}	P and A belong to different key equivalence classes	WL
$o_{+?}$	P, but not A, corresponds to a key occurrence	WL
$o_{+\_}$	P corresponds to a key occurrence, no anchor A determined	FN
$o_{?+}$	P does not correspond to a key occurrence	FP
$o_{?\_}$	P does not correspond to a key occurrence, no anchor A determined	TN

Table 3.9: Pairs of pronouns (P) and antecedents (A) and their classification by Stuckardt (2001).

Stuckardt then calculated Precision and Recall as:

$$Precision = \frac{|o_{++}|}{|o_{++}| + |o_{+-}| + |o_{+?}|}, Recall = \frac{|o_{++}|}{|o_{++}| + |o_{+-}| + |o_{+?}| + |o_{+\_}|}$$

In comparison, in the case of a resolved gold mention pronoun which is linked to an incorrect antecedent (WL), we do not distinguish between antecedents which are gold mention or spurious system mentions (o_{+-} vs. $o_{+?}$), because we do not see how this distinction benefits higher-level applications. Other than that, the Recall calculation for the ARCS inferred antecedents metric, which also requires non-pronominal antecedents for pronouns, is close to Stuckardt's definition[12]. However, for Precision, Stuckardt does not include the $o_{?+}$ cases (FP), i.e. spurious system pronouns. We argue that systems should be penalized for returning nominal antecedents for spurious pronouns, and therefore ARCS includes the FP class in the Precision denominator, which is equivalent to including $o_{?+}$ in the Precision denominator in Stuckardt's definition.

[12] However, the ARCS metrics feature additional rules for classifying gold and system mentions w.r.t. whether they are the first mentions in a coreference chain (cf. algorithm 2 and Tuggener (2014)).

3.4.3 ARCS extensions for pronoun resolution evaluation

When we introduced the ARCS framework (Tuggener, 2014), we showed how it can be used to measure performance on any mention level. Therefore, evaluating pronoun resolution with ARCS is straight-forward, as the mention level accessed is the PoS level and we simply look at the performance of the respective PoS tags to asses pronoun resolution performance of a coreference resolution system, as exemplified in table 3.7. We can also evaluate on the lemmas of the given PoS tags to achieve an even more fine-grained evaluation, as presented in Tuggener and Klenner (2014). Here, we present two additions to the ARCS evaluation framework which are applicable to pronoun resolution and which incorporate ideas presented in related work that we discussed in the previous section.

To achieve what (Mitkov, 2001) deemed the comparison of resolution algorithms, we calculate a score with the mention classification inventory defined in the ARCS framework. We define *ARCS Accuracy* as the ratio of correctly resolved gold mentions divided by all resolved gold mentions, i.e.:

$$ARCS\ accuracy = \frac{TP}{TP + WL} \tag{3.2}$$

Doing so, we cancel out the anaphoricity or mention detection problem and only evaluate on the subset of gold mentions that a system actually resolved. However, systems cannot be directly compared to each other using this metric, because the sum of true positives and wrong linkages is usually not the same for every systems. Therefore, evaluation does not necessarily extend over the same set of mentions for all systems and does not allow a direct comparison. The metric simply indicates how well a single system performs considering the gold mentions it resolves.

The Accuracy metric is still prone to involve noise from preprocessing, because even if the system has correctly determined a pronoun to be anaphoric, it might fail to produce the correct antecedent NP as a candidate due to issues in markable extraction or preprocessing. Therefore, when evaluating classifiers that select an antecedent among a set of candidates, we propose to only evaluate cases where the classifier has access to the correct antecedent. Otherwise, tracking the performance changes of a classifier, e.g. while feature engineering, is interfered by a constant count of cases where it does not have the possibility to make the right choice. In other words, the sole purpose of an antecedent selection strategy is to pick the correct antecedent among a set of candidates. The selection strategy (i.e. a classifier) will never be able to correct errors in preprocessing. Therefore, it makes little sense to punish it for these errors in an evaluation that specifically aims at measuring antecedent classification. To address this, we propose, in Mitkov's spirit, the following simple ratio to assess the performance of a selectional strategy:

$$\begin{aligned} ARCS\ success\ rate &= \frac{|correctly\ resolved\ anaphors|}{|resolvable\ anaphors|} \\ &= \frac{TP}{TP + WL} \text{ where } correct\_antecedent \in candidates \end{aligned} \tag{3.3}$$

"Resolvable" in our case means that the anaphor is annotated as anaphoric in the gold standard and that the classifier has the correct antecedent among the candidates. Clearly, one drawback of this measure is that it cannot be applied to system response files, since TP and WL need to be counted only where $correct\_antecedent \in candidates$, and we cannot tell from the final system response to which mentions this condition applied. Also, like accuracy, the metric cannot be used to compare performance of different systems for the reasons given above. Despite these drawbacks, this measure is still an effective and most accurate method for evaluating a resolution strategy's performance on a clean basis, i.e. without noise from preprocessing. We make use of this metric in section 5.4.2.

3.5 Chapter summary

This chapter discussed the empirical evaluation of coreference and pronoun resolution. We adapted the view of downstream application which benefit from coreference resolution as a preprocessing component. From this view, we argued that the common coreference evaluation is suboptimal in terms of interpretability, informativeness, and differentiability. We presented an alternative evaluation framework, ARCS, to address these issues. The framework adapts w.r.t. the potential requirements of higher-level applications by implementing different requirements regarding the definition of the correctness of the antecedent. We additionally introduced two extensions to ARCS to evaluate pronoun resolution in settings freed from preprocessing noise. With the extensions, ARCS presents a framework that enables a thorough evaluation of systems and which we will use for evaluating our approaches to German pronoun resolution in chapter 5.

We showed that our framework is suitable for quantifying the differences of system outputs and compared four systems for coreference resolution for English. The F-score differences provided by the evaluation based on the common performance metrics suggested that the analysed system outputs do not differ strongly. By contrast, we found that the analysed systems processed up to almost 40% of the mentions differently. Furthermore, we demonstrated how the framework is applied to investigate what drives the difference between two systems, which, in turn, provides insight into the reasons behind the different performances in F-score in the common evaluation. An oracle-based combination of the system responses showed that they are often complementary, indicating that specific weaknesses of one system can be overcome by the specific strengths of another. It remains the task of future work to assess whether such a system combination can be successfully realized in an automated setting.

Chapter 4

Related work on coreference and pronoun resolution for German

In this chapter, we chronologically review related work on coreference and pronoun resolution for German. We examine related work along three axes, i) the applied discourse model[1], ii) the filters that license potentially coreferring markables, and iii) the feature set used to classify these instances of potentially coreferring markables. We also keep track of what the approaches report on the specifics of German w.r.t. the adaption of methods primarily developed for English coreference and pronoun resolution. As in our discussion of discourse processing models in chapter 2, it is difficult to identify a system that performs best overall, since the approaches apply different evaluation protocols and use different corpora as test sets. We thus report F-scores and compare systems where possible.

4.1 Hartrumpf (2001)'s CORUDIS

Hartrumpf (2001) presented a hybrid approach for German coreference resolution called CORUDIS (COreference RUles with DIsambiguation Statistics) which stands out regarding its architecture. A set of manually crafted rules licenced coreference links between markables. The rules captured e.g. possible coreference of a pronoun and a noun antecedent given morphological (in)compatibility or potential coreference between to noun markables based on their semantic class properties.

Foreseeing the entity-mention model later employed for English in Luo et al. (2004)[2], the approach incrementally built partial coreference partitions. Given a markable, a

[1] The strategy used to link the markables; cf. chapter 2.

[2] Cf. section 2.2.3

separate partition was created for all rule-licensed antecedent candidates. The creation of partitions was prevented when distance constraints were not met (i.e. an antecedent candidate was too far away from the markable at hand) and whenever coreference sets emerged that featured mentions with incompatible semantic class properties. To further narrow down the emerging partitions, beam search was applied to only keep those partitions with promising scores. The score of a new possible partition was calculated by summing the pair-wise probabilities of linking the markable with each to the antecedent mentions in a partially established coreference chain. Also, a weight for the coreference rule that licenced the formation of the partition was included. When reaching the end of a document, the highest weighted partition served as the system response.

Beside common features such as morphosyntactic properties (gender, number, part-of-speech), Hartrumpf queried a set of features based on a syntactic-semantic parser. The parser incrementally processed words and their syntactic-semantic representation in a HPSG-like fashion. Features for coreference captured different semantic aspects of antecedent candidates and anaphors.

Hartrumpf reported impressive results of 66% F-score in a MUC-like evaluation setting on a corpus derived from the German newpaper 'Süddeutsche Tageszeitung'. However, he did not give performance results for pronouns.

Hartrumpf's approach featured an entity-mention-like perspective on coreference and a global optimization function, which was arguably ahead of its time.

4.2 Strube et al. (2002)'s adaption of the popular approach by Soon et al. (2001)

Strube et al. (2002) investigated the adaption of the popular mention-pair model for coreference resolution (Soon et al., 2001)[3] to German. Strube et al. used features commonly applied in coreference resolution for English (Cardie et al., 1999, Soon et al., 2001, inter alia) and manually added semantic class labels (*human, concrete objects, abstract objects*) to NPs in their corpus (the Heidelberg Text Corpus[4]) which were used as features. Additionally, they introduced the minimum edit distance (MED) feature for string matching nominal markables. Similar to the Levenshtein distance, MED quantifies the type of string manipulations (substitutions, insertions, deletions) required to transform one string into another and thus provides a string similarity metric. Strube et al. used

[3] Cf. section 2.1

[4] Cf. section 5.1.1

MED to measure string similarity of potentially coreferring nominal markables, which proved to be helpful in increasing Recall.[5]

	Document level features	
1.	doc id	document number (1 . . . 250)
	NP-level features	
2.	ante gram func	grammatical function of antecedent (subject, object, other)
3.	ante npform	form of antecedent (definite NP, indefinite NP, personal pronoun, demonstrative pronoun, possessive pronoun, proper name)
4.	ante agree	agreement in person, gender, number
5.	ante semanticclass	semantic class of antecedent (human, concrete object, abstract object)
6.	ana gram func	grammatical function of anaphor (subject, object, other)
7.	ana npform	form of anaphor (definite NP, indefinite NP, personal pronoun, demonstrative pronoun, possessive pronoun, proper name)
8.	ana agree	agreement in person, gender, number
9.	ana semanticclass	semantic class of anaphor (human, concrete object, abstract object)
	Coreference-level features	
10.	wdist	distance between anaphor and antecedent in words (1 . . . n)
11.	ddist	distance between anaphor and antecedent in sentences (0, 1, >1)
12.	mdist	distance between anaphor and antecedent in markables (1 . . . n)
13.	syn par	anaphor and antecedent have the same grammatical function (yes, no)
14.	string ident	anaphor and antecedent consist of identical strings (yes, no)
15.	substring match	one string contains the other (yes, no)
	New coreference-level features	
16.	ante med	minimum edit distance to anaphor
17.	ana med	minimum edit distance to antecedent

Table 4.1: Feature set of Strube et al. (2002)'s approach to German coreference resolution in the HTC corpus.

Table 4.1 presents Strube et al.'s final feature set. We list these features explicitly because they provide a good overview of the kind of features coreference resolution systems use to date at their core. NP-level features lists features that describe either properties of the antecedent or the anaphor. The NP-level features capture syntactic salience (2, 6), surface forms (3, 7), morphosyntactic compatibility (8), and semantic class membership (5, 9). (New) coreference-level features lists features that concern the relation of antecedent and anaphor. They quantify the distance between antecedent and anaphor (10-12), capture parallelism of the grammatical roles (13), and measure surface string similarity (14-17).

[5] We assume that the features were especially helpful because Strube et al.'s approach did not use lemmatization of the NPs before string matching, but used lexemes. We infer this by looking at Strube et al.'s examples in their table 5. Since German features inflection, exact string matching of lexemes misses many coreferent definite NP pairs because of mismatches caused by inflection. The MED feature helps alleviate such mismatches by allowing a certain degree of fuzzy matching, which approximates string matching based on lemmas.

The approach of Strube et al. featured a battery of filters that prevented the creation of pair instances if

- the anaphor was an indefinite NP
- antecedent and anaphor shared the same syntactic head
- antecedent and anaphor had different values in their semantic class attributes (based on the manually added class labels; only applicable if antecedent and anaphor were not pronouns)
- either antecedent or anaphor was not a third person NP or pronoun
- antecedent an anaphor do not match in their morphological properties (only applicable if the anaphor was a pronoun)

The last filter is of major important for German pronoun resolution, as the number of candidates is commonly large. However, Strube et al. did not elaborate how they treated underspecification of German pronouns in the process of matching morphological properties. Also, they did not report how they assessed morphological properties of possessive pronouns.[6] Applying these filters, Strube et al. reported an overall reduction of negative pair instances by 50%.[7]

Strube et al. reported performance for individual types of anaphors, i.e. definite NPs, named entities, and demonstrative, possessive, and personal pronouns. However, the paper does not explicitly state what evaluation approach was used. We assume that the authors reported the evaluation output of their decision tree classifier which classified positive and negative instances of potential antecedent-anaphor pairs. Relevant for coreference is the performance on the positive instances, since they determine which coreference relations the system establishes. Strube et al.'s results (on what we presume to be the positive instances) are given in table 4.2.

The table shows that the approach achieved solid performance, with demonstrative pronouns (PDS) being the exception. Demonstratives are problematic to resolve, since it is difficult to determine their anaphoricity status. That is, demonstratives can refer to whole clauses or verbs, which are not annotated in the corpora. Resolving them to nominal antecedents thus often yields incorrect resolutions. We further note that the

[6] In most corpora, possessive pronouns are marked with the morphological properties of their head. For example, in the NP "seine Frau (his wife)", "seine" will be annotated with the number and gender of the head "Frau", i.e. singular and feminine. However, "seine" cannot refer to feminine antecedents, but only to masculine of neuter ones. Therefore, one needs to manually address morphological properties of possessive pronouns and account for the underspecification in the matching process.

[7] Cf. section 2.1.1.2 for a discussion of the problem of imbalanced training data in the mention-pair model.

	Prec.	Rec.	F1
defNP	69.26%	22.47%	33.94%
NE	90.77%	65.68%	76.22%
PDS	25.00%	11.11%	15.38%
PPER	85.81%	77.78%	81.60%
PPOS	82.11%	87.31%	84.63%
all	84.96%	56.65%	67.98%

Table 4.2: Anaphora type specific evaluation of Strube et al. (2002)'s approach to German coreference resolution in the HTC corpus.

performance for pronouns is higher than that reported in Hinrichs et al. (2005) (discussed below), although we cannot directly compare the scores since different corpora were used. Furthermore, the F-score is similar to that reported by Hartrumpf (2001). However, Hartrumpf (2001) reported MUC F-scores, i.e. performed evaluation of full coreference chain, while Strube et al. evaluated pair instances and it is not clear how performance on pair instances carries over to coreference chain-level performance (Ng and Cardie, 2002b, e.g.). Also, Hartrumpf and Hinrichs et al. evaluated on newspaper texts, while Strube et al. used texts evolving around historical aspects of the city Heidelberg.

4.3 Schiehlen (2004)'s exploration of algorithms and features for German pronoun resolution

Schiehlen (2004) explored several features used for pronoun resolution in the English literature for the resolution of German pronouns. He applied the features as either hard constraints for antecedent filtering or as soft constraints for weighting antecedent candidates for German pronoun resolution.

He quantified the impact on the availability of the correct antecedent and the average number of antecedents per pronoun when applying several features as hard constraints. Doing so, he was able to give i) upper bounds for applying different filters and ii) an estimate of the resulting difficulty of resolving the pronouns (indicated by the remaining average number of candidates - the lower, the easier).

He found that removing antecedent candidates based on morphological disagreement cuts the number of candidates by more than half and only marginally lowers the upper bound, which aligns with the findings of Strube et al. (2002)[8] and substantiates the necessity of the filter especially for German pronoun resolution. Schiehlen further examined how other features, such as parallelism of grammatical roles of antecedent and pronoun, distance, and binding constraints, affect upper bounds when used as filters. However,

[8] Cf. the previous section 4.2

none of them showed to be as efficient as morphological agreement in removing irrelevant candidates and keeping a high upper bound.

Schiehlen then compared different rule- and machine learning-based approaches from the English literature to combine the features as weights to rank antecedent candidates. He further evaluated using gold annotation for parsing vs. relying on automatically parsed texts and reported the observed performance drops.

He found that the popular approach by Soon et al. (2001) relying on decision trees fared poorly when applied to German pronoun resolution. He achieved the best results both for automated and gold parses using a decision tree variant with an adapted feature set. On a test set derived from the NEGRA corpus (Skut et al., 1997), Schiehlen achieved an overall F-score of 71.10% on gold parses and an F-score of 51.70% on automated parses. It is worth noticing that the adaption of Lappin and Leass (1994)'s rule-based approach[9] performed second best for the automatically parsed variants, outperforming other machine learning approaches such as the Soon et al. (2001) approach.

The work of Schiehlen provided important insights into filtering techniques and the impact of features for German pronoun resolution. He demonstrated that over 97% of the pronouns have an antecedent within a window of the three preceding sentences. Thus, antecedent candidate generation for pronouns can safely be constricted to this window. He further found that cataphora is a rare phenomenon, i.e. only 1.6% of the pronouns were cataphoric, indicating that cataphora can be discarded without much performance loss. This facilitates the pronouns resolution process by reducing the number of candidates that have to be considered, since only antecedent candidates are accessed and potential postcedents (i.e. NPs following the pronoun) are dismissed.

4.4 RAP-G and the TiMBL variant by Hinrichs et al. (2005)

Lappin and Leass (1994)'s Resolution of Anaphora Procedure (RAP) was one of the first automated approaches to English pronoun resolution that used salience features with manually assigned weights to rank antecedent candidates. Like Schiehlen (2004), Hinrichs et al. (2005) re-implemented this approach for German pronouns, but put effort into adapting it to German pronoun resolution.

Lappin and Leass's approach mainly relies on a hierarchical ordering of grammatical roles the antecedent candidates occur with. For example, candidates bearing the role *subject*

[9] Cf. the next section.

are ranked higher than candidates with the role *object* etc. Hinrichs et al. revamped the inventory of grammatical roles for German and empirically fine-tuned the assigned weights. Interestingly, Hinrichs et al. found that the subject emphasis worked optimally for German if set more than twice as high as in the original English approach. Also, they reported that lowering the penalty for short distance cataphora for German improved their results. Table 4.3 shows the adapted feature set and weights.

Feature	**Weight**
Subject emphasis	170
Accusative object emphasis	70
Dative object emphasis	50
Genitive object emphasis	50
Head noun emphasis	80
Short distance cataphora penalty	-80
Long distance cataphora penalty	-175
Parallelism reward	35
Current sentence reward	20

Table 4.3: Features and their weights in RAP-G, Hinrichs et al. (2005)'s adaption of Lappin and Leass (1994)'s approach to English third person pronoun resolution.

The salience of an antecedent candidate is calculated by summing the applicable feature weights. The weight is lowered by a salience decaying function to reflect the importance of distance between a pronoun and its antecedent. The salience value sv' for an antecedent candidate given a pronoun is calculated based on the prior salience sd by $\frac{sv}{2^{sd}}$ where sd denotes the sentence distance between the candidate and the pronoun.

Much like in the entity-mention model[10], the established antecedent-pronoun pair is appended to the discourse entity denoted by the antecedent. This discourse referent has a salience value equal to the sum of all its mentions, which reflects that discourse referents occurring multiple times in discourse have a higher salience.

Hinrichs et al. (2005) evaluated their approach on the 5540 third person pronouns (possessive, reflexive, and personal) in the TüBa-D/Z corpus. A pronoun was deemed correctly resolved when it selected an antecedent that was in the same coreference chain as the pronoun. Furthermore, Hinrichs et al. allowed for partial mention boundary matching as long as the head noun of the NP in the gold standard was contained in the found antecedent. Doing so, they achieved an F-score of 76.56% with almost identical Recall and Precision scores.

Hinrichs et al. (2005) compared their re-implementation of RAP to a variant using a memory-based kNN classifier (TiMBL). First, potential antecedent candidates were filtered based on sentence distance (not more than three sentences away; in the same

[10] Cf. section 2.2.3

sentence if the pronoun was a reflexive), binding constraints, and morphological agreement. The features of the antecedent candidates used in RAP-G were mapped to vectors were labeled as positive or negative instances for the memory-based classifier. This approach yielded an F-score of 70.4%, thus underperformed compared to the rule-based RAP-G.

Hinrichs et al. (2005, 2007) further explored the TiMBL variant. They focused on the syntax-based features and included a notion of discourse history. They modelled the history of grammatical roles that an antecedent entity had occurred with before appearing in the context of a given pronoun. The idea was to capture the global salience of antecedent candidates w.r.t. the pronoun at hand. One version implemented this idea by encoding how often an antecedent entity occurred with a grammatical role (i.e. has occurred two times as subject and one time as object etc.), the other version captured the distance to the last mention of the entity with a given grammatical role (i.e. has occurred as object one sentence before and as subject two sentences before). Hinrichs et al. showed that both these approaches significantly outperformed the variant which did not model global salience of antecedent candidates. They achieved results of 74.3% F-score on the TüBa-D/Z using ten-fold cross-validation compared to the version which was uninformed of the discourse history of the antecedent entities, which had yielded 70.4% F-score. The features based on the discourse history of the antecedent candidates showed that keeping track of earlier mentions of antecedent entities is important for pronoun resolution, which aligns with findings of Yang et al. (2004b) on English data.[11] This provided a significant improvement of performance, however still below the 76.56 % F-score mark of the RAP-G approach.

Finally, Hinrichs et al. (2005) implemented a heuristic to resolve pronouns for which the TiMBL classifier did not find any positive pairs of antecedent and pronoun, i.e. which the classifier left unresolved. For such unresolved pronouns, Hinrichs et al. selected the closest compatible subject-bearing candidate as antecedent. This improved Recall by almost 10 percentage points at the cost of roughly 4 points in Precision and yielded an F-score of 77%.

Three more notable publications evolved around the RAP-G approach and the hybrid TiMBL variant. Wunsch (2006) investigated the features and their manually set weights in RAP-G. Wunsch manually increased or decreased the weights for each feature and tracked the changes in performance. He found that positional features such as distance play a less important role in German than in English pronoun resolution. He reported that the syntax-based features, especially the subject emphasis, is more substantial for German pronoun resolution than for the English variant. Wunsch hypothesized that

[11] Cf. section 2.2.3

this difference stems from the relatively free word order in German which is not given to the same degree in English.

Wunsch et al. (2009) investigated the effect of instance sampling for the TiMBL hybrid. As discussed in section 2.1.1.2, the mention-pair model in the spirit of Soon et al. (2001) produces a high number of negative instances due to its pair generation mechanics, and the TiMBL hybrid featured a similar architecture. This yield classifiers with a strong bias towards negative classification, which in turn often leads to none of the candidate pairs being classified as positive, even for clearly anaphoric markables such as third person pronouns like *er*. Using random sampling of the negative instances, Wunsch et al. lowered the data imbalance and raised performance of their decision tree-based approach from 56.10% to 61.10% F-score.

Finally, the thesis of Wunsch (2010) wrapped up the experiments evolving around the adaption of RAP to German and the TiMBL variant.

4.5 Klenner and Ailloud (2009)'s model of enforcing global coreference constraints using ILP

As discussed in section 2.1.1, the mention-pair model suffers from a narrow perspective on discourse when processing coreference. That is, decisions about pairs of antecedent candidates and anaphors are kept local and are not propagated to further decisions. When the pairs classified positive are merged to from coreference chains, inconsistencies can arise.

To propagate local classification decisions to the global merge step and enforce consistency, Klenner and Ailloud (2009) proposed an approach based on Integer Linear Programming (ILP), a framework for finding optimal local decisions given global constraints. In this approach, the classifications and their weights of local antecedent-anaphor pairs[12] served as the input for the ILP constraints. These then enforced the transitivity and exclusiveness properties of coreference during the pair merging step. The transitivity property states that if mention A and mention B are coreferent, and mention B and C are coreferent, then mention A and C have to be coreferent as well. Exclusiveness operates analogously, i.e. if mention A and B are coreferent, and mention B and C are exclusive, then mention A and C are exclusive as well.

Klenner and Ailloud used mostly the same feature set as Strube et al. (2002), with limited string matching features (i.e. only one feature that captured whether the syntactic heads

[12] This approach also used TiMBL as a classifier. The weight of a TiMBL decision was calculated as the ratio of the positive and negative neighbors of the test instance.

of an antecedent candidate and an anaphor matched). Klenner and Ailloud showed that their approach using the global ILP constraints during pair merging outperformed a simple transitive merge of the corresponding positive pairs. The authors obtained an F-score of 64.27% measured in CEAFM (Luo, 2005) with the ILP model, compared to 62.83% using the baseline merge operation.

4.6 Broscheit et al. (2010)'s adaption of BART to German

Broscheit et al. (2010) ported the English coreference system BART, a mention-pair system along the lines of Soon et al. (2001), to German using the feature set introduced in Klenner and Ailloud (2009). Broscheit et al. explored five additional features:

- 1/2 PERSON: for each antecedent and anaphor in turn, TRUE if it is first or second person, FALSE otherwise.
- SPEECH: for each antecedent and anaphor in turn, TRUE if it is inside quoted speech, FALSE otherwise.
- NODE DIST: the number of clause nodes and prepositional phrase nodes along the path between anaphor and antecedent in the parse tree.
- PARTIAL MATCH: TRUE if the head of anaphor is contained in the head of antecedent or vice versa, FALSE otherwise.
- GERMANET RELATEDNESS: the semantic relatedness between antecedent and anaphor, as found in GermaNet.

Of the added features, especially the PARTIAL MATCH feature improved performance. Broscheit et al. also compared a Decision Tree and a Maximum Entropy classifier and found that the latter outperformed the former. Additional performance increase was obtained when separate Maximum Entropy classifiers were trained for pronominal and non-pronominal anaphors.

Unfortunately, the authors only reported performance in the true mention setting. That is, the system only had to establish coreference relations between the coreferent NPs, and did not have to decide which markables it should attempt to resolve. Klenner and Ailloud (2009), for example, had demonstrated the difference between evaluating on gold mentions vs. evaluating on all markables. Broscheit et al. (2010) compared their system to Klenner and Ailloud (2009)'S ILP extended approach and achieved an F-score of 65.00%. Klenner and Ailloud had achieved an F-Score of 71.50% on the gold mentions.

4.7 The SemEval 2010 shared task (Recasens et al., 2010)

The SemEval 2010 shared task on coreference resolution in multiple languages (Recasens et al., 2010)[13] featured German as a language. A subset of the TüBa-D/Z was taken as data and four out of the participating six systems submitted responses for the TüBa-D/Z test set. A major problem with the evaluation in the Semeval 2010 shared task was that it included and scored singletons (i.e. NPs not in coreference chains) in the coreference annotation. The reason for doing so was the suggestion that systems should be rewarded for not resolving singletons. However, as shown by e.g. Kübler and Zhekova (2011), singletons tend to artificially boost coreference resolution system scores and therefore give a biased impression of the actual system performance. This becomes obvious when the results of the SemEval shared task are analysed based on the MUC metric, which only scores coreference links and is, therefore, not affected by singletons.

One system that achieved competitive MUC F-scores (i.e. above 40%) for German was the BART system which we have described in the previous section. The other system was SUCRE (Kobdani and Schütze, 2010). It also implemented a mention-pair model and pair generation mechanics derived from Soon et al. (2001). Noticeably, the system featured elaborate string matching features to determine whether two markables should corefer. This allowed it to recognize e.g. that the two markables *the university student from Germany* and *the university student from France* should not be linked, although their syntactic head and a large portion of the other words contained in the NPs match. Apart from these features, the system made use of features commonly used in coreference resolution and relied on a Decision Tree classifier to classify potential pairs of antecedents and anaphors. Regarding the MUC F-scores, SUCRE outperformed the BART system in the gold setting (where perfect preprocessing, such as dependency parsing, was provided[14]) with 58.40% vs. 51.1%, but was outperformed by BART in the regular setting (where preprocessing was done automatically) with 45.50% vs. 40.90%. The significant performance drop for SUCRE indicates that the system had problems identifying the markables in automatically preprocessed documents.

4.8 Versley (2010)'s investigation of semantics for noun and name coreference

Versley (2010) focused on coreference of noun and name mentions. His special interest was the resolution of definite noun anaphors (NPs which have a definite article) which

[13] `http://stel.ub.edu/semeval2010-coref/`

[14] Cf. section 1.5.4

cannot be resolved to an antecedent using string matching techniques as e.g. *[Monsanto - the company]*. He called these cases coreferent bridging. However, he also investigated coreference between non-pronominal mention that can be identified through string matching.

In his first series of experiments, Versley excluded the problem of anaphoricity detection and evaluated his approaches only on definite NPs that are annotated as coreferent in the gold standard. Versley used the first 125 documents in the TüBa-D/Z version 5 as a test set. He found that roughly 50% of the definite coreferent NPs can be resolved using head string matching, as in e.g. *[A company - the company]*.

For the remaining 50% of the definite NPs, Versley explored a batch of features which captured the relatedness of definite anaphoric NPs and their antecedents. He found that a hyperonymy look-up in GermaNet achieved the best Precision (67%) for the non-matching anaphors, and the best F-score (68%) for all definite anaphoric NPs. Apart from GermaNet, Versley applied an impressive amount of diverse features based on different distributional similarity models and pattern-based approaches to derive noun relatedness measures from corpora. Combining these features and applying a batch of filters, Versley achieved and F-score of 73% for all definite anaphoric NPs in his test set.

For the second series of experiments, Versley included the anaphoricity detection problem of deciding whether a definite NP should be resolved or not. In this more realistic setting, he found that trying to resolve non-matching definite NP anaphors is more difficult and less beneficial for definite NPs resolution.

4.9 The incremental entity-mention model by Klenner and Tuggener (2010)

Beside Hartrumpf (2001), Klenner and Tuggener (2010) was, to the best of our knowledge, the only approach to German coreference resolution which departed from the mention-pair paradigm. Since this system and the later publications surrounding it (Klenner and Tuggener, 2011a,b, Tuggener and Klenner, 2014) serve as a basis for this thesis, we will not discuss it in more detail at this point.

4.10 Rösiger and Riester (2015)'s adaption of HOTCoref to German

Recently, Rösiger and Riester (2015) presented an adaption of the HOTCoref system (Björkelund and Kuhn, 2014) to German. HOTCoref itself is currently the best-performing combined system for multilingual coreference resolution (i.e. for English, Arabic, and Chinese). HOTCoref promotes the idea to model coreference sets as trees, following Fernandes et al. (2012), which in some cases yields more plausible antecedents for learning than the pair generation mechanics presented in Soon et al. (2001), i.e. the mention-pair model. Since coreference sets are modeled as trees, it is possible to apply structure prediction algorithms, such as structured perceptron. Rösiger and Riester's adaption of HOTCoref to German was geared towards exploring the use of prosodic features in coreference resolution based on the DIRNDL corpus (Björkelund and Kuhn, 2014). Therefore, they did not focus on improving the baseline system performance on the TüBa-D/Z corpus or pay particular attention to pronoun resolution. Nevertheless, Rösiger and Riester showed that their adaption achieved competitive results on the TüBa-D/Z corpus version 9 with an average F-score of 53.63%, as well as on the SemEval shared task data with average F-scores of 60.35% for the evaluation including singletons and 48.61% in the run without singletons.

4.11 Chapter summary

In this chapter, we provided an overview of related work on German pronoun and coreference resolution. We found that most approaches straight-forwardly adapt methods from approaches to coreference resolution for English and do no account explicitly for the problem of underspecification of certain German pronouns. That is, none of the related work accounts for the fact that German pronouns can refer to both animate and inanimate entities and that certain German pronouns are morphologically underspecified.

Only Hartrumpf (2001) and Strube et al. (2002) included semantic features which denoted the entity type of an antecedent candidate and which implicitly modeled animacy. However, in the approach of Strube et al., these features were added manually and were therefore not applicable to an end-to-end coreference system. Also, it is not clear how animacy and semantic class membership of antecedent candidates that are pronouns is handled in these approaches. In our incremental entity-mention model, which we will discuss in the next chapter, we encode animacy and named entity class as features to capture prior probabilities of linking pronouns to animate or inanimate entities and to different named entity types. Also, our entity-mention model is able to label resolved pronouns with these semantic class properties by projecting the properties of the selected antecedents onto them. Therefore, pronominal antecedents then carry the semantic load of the entity they denote and are no longer semantically empty.

We note further that most of the related work employs a mention-pair model. The rule-based RAP-G approach by Hinrichs et al. (2005) featured an equivalency class representation which kept track of previous decisions. However, this class was not used to propagate entity-level features or to ensure consistency within coreference chains. Thus, most of the related approaches are susceptible to the commonly known weaknesses of the mention-pair model. A clear exception is the CORUDIS system (Hartrumpf, 2001) which employed an incremental architecture that enabled it to enforce (at least semantic) consistency in the coreference chains.

One technique shared by almost all related approaches is that of filtering morphologically incompatible antecedent candidates. We argued that German pronoun resolution is difficult since animacy is not a hard constraint for filtering candidates. Therefore, the number of candidates that need be considered is tentatively large. Filtering based on morphological constraints (i.e. gender and number agreement) is thus a reliable technique to reduce the number of potential candidates and has been shown to effectively do so, despite the underspecification of certain pronouns which increases the number of candidates that have to be considered even more. Our incremental entity-mention model also adapts this kind of filtering. One advantage of our model over related work is that resolved pronouns are disambiguated with the properties of the selected antecedent. Thus, morphological compatibility is established based on the entity-level morphological properties, which alleviates the problem of local underspecification of e.g. pronominal antecedents. Once disambiguated, such pronouns can only act as antecedents for specific subsequent pronouns.

Chapter 5

Empirical validation of our entity-mention model

In this chapter, we empirically evaluate the models introduced in chapter 2. The evaluation serves two main purposes:

1. Evaluate empirically the claims of the theoretical advantages of the entity-mention model compared to the mention-pair model made in section 2.3 w.r.t German pronoun resolution.

2. Compare different antecedent selection strategies outlined in chapter 2, namely the best-first heuristic, the twin candidate model, and mention ranking. We implement three machine learning frameworks for antecedent selection which correspond to the three strategies.

We first overview the data and then discuss relevant implementation details regarding the models. Finally, we empirically explore pronoun resolution performance of our approaches in this chapter. Throughout our discussion here, we focus on third person pronouns, because it is the area where we believe our approach makes a major contribution. All our models use the same state-of-the-art strategy to resolve noun markables, providing full coreference resolution for German. We outline the details of our approach for coreference resolution between nominal markables in section A.2. Details on the resolution of first person pronouns to nominal antecedents are given in section A.1.

5.1 Data and preprocessing: The TüBa-D/Z corpus

We make use of the TüBa-D/Z treebank (Telljohann et al., 2004) version 9.1 for training and testing our system. The TüBa-D/Z consists of articles gathered from the German newspaper 'die Tageszeitung' (taz).[1] TüBa-D/Z version 9.1 features 3644 articles (95'595 sentences; 1'787'801 tokens). The corpus provides gold annotation layers for morphology, syntax, named entity types, and coreference, among others and has been used for German coreference resolution in related work extensively (Hinrichs et al., 2005, 2007, Wunsch et al., 2009, Wunsch, 2010, Versley, 2010, Màrquez et al., 2012, our work).

Versley (2006) investigated interannotator agreement for coreference on 60 documents and found an agreement MUC F-score of 85% after resolving mention boundary issues.[2] Versley (2006) concluded that this was slightly higher than the figures for a comparable corpus for English coreference resolution (the MUC-6 corpus). The TüBa-D/Z has undergone several refinements since. Still, one can never expect perfect gold annotation, and thus the 85% MUC score can be viewed as an upper bound for automated coreference resolution.

Figure 5.1 provides a quantitative overview of the third person pronouns in the TüBa-D/Z 9.1 that are relevant to our investigation. The figure shows both the count of anaphoric (upper black bar) and non-anaphoric (lower gray bar) pronouns per PoS type as given by the gold standard annotation.[3] Since our approach makes the naive assumption that all considered pronouns are anaphoric, this figure gives an estimate on how much false positives we will encounter, which is bound to lower performance of our approach in terms of Precision.

The non-anaphoricity of third person pronouns in the gold annotation has several sources. We manually investigated some of these pronoun instances and found that many of them simply lack annotation in the gold standard. This occurred particularly often for relative pronouns (PRELS, PRELAT). Demonstrative pronouns (PDS) frequently refer to whole clauses and therefore do not have a NP antecedent. In such cases, a demonstrative pronoun is not annotated with coreference. Finally, there are cases where pronouns are indeed anaphoric and do have a linguistic antecedent, but the antecedent does not denote an actual extralinguistic entity. Such cases of bound anaphora are not always annotated in the corpus.[4] This accounts for the lacking annotation regarding some personal (PPER) and possessive (PPOSAT) pronouns.

[1] `http://www.taz.de/`

[2] Given two annotators, the agreement MUC F-score was calculated by taking one annotator's annotation as the key and the other annotator's annotation as the response. Then, key and response were inverted to get the second F-score. The average F-score then served as the agreement score.

[3] Note that we exclude pleonastic uses of *es (it)*, see below.

[4] Cf. section 1.4.1.3

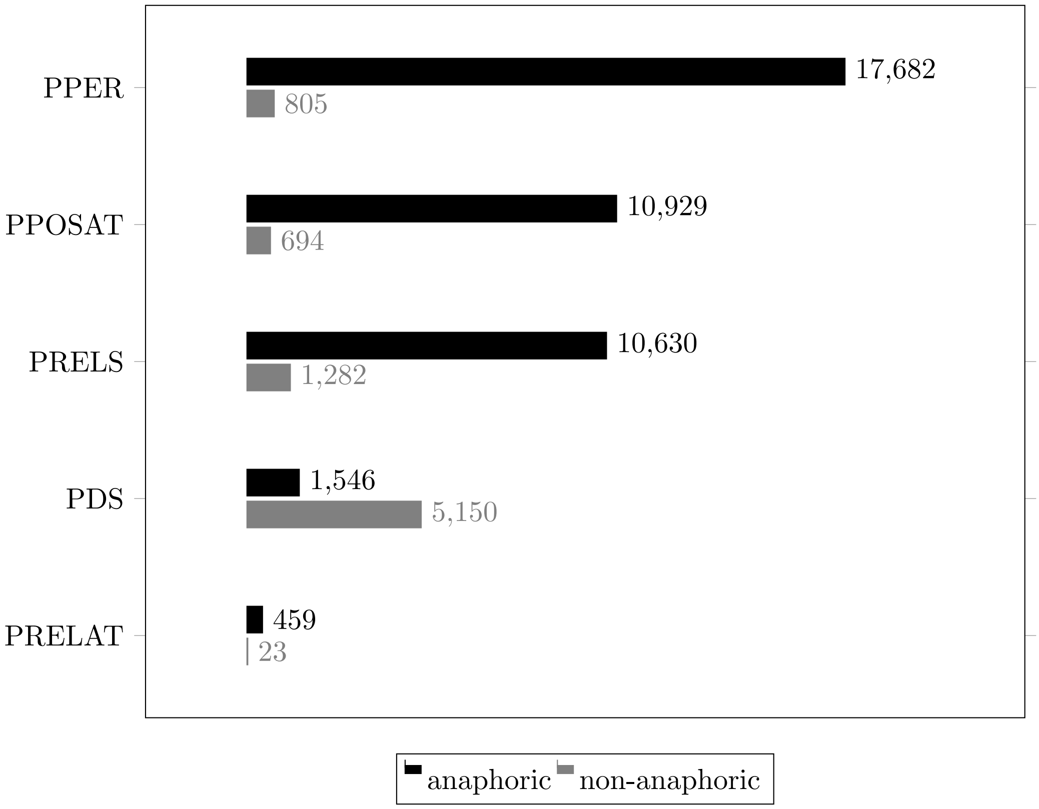

FIGURE 5.1: Distribution of 3rd person pronouns (excluding *es*) in the TüBa-D/Z version 9.1

We perform a 20%-20%-60% split of the TüBa-D/Z to obtain the test, development, and training set, respectively. The test set consists of the first 690 documents, the development set of the following 690, and the training set of the remaining 2264 documents. For comparison, Wunsch (2010) performed ten-fold cross validation on an earlier version of the TüBa-D/Z which contained 27'125 sentences (473'747 tokens), i.e. which was about 30% the size of the current one. Versley (2010) used the first 125 TüBa-D/Z articles of an earlier version as a test set in his experiments on noun and name coreference. The SemEval shared task on coreference resolution for multiple languages (Màrquez et al., 2012) also featured a subset of an earlier TüBa-D/Z version. The test set contained 136 documents. We thus deem our test set to be of reasonable size.

5.1.1 Other German corpora featuring coreference annotation

There are three other German corpora we are aware of which feature coreference annotation. The Potsdam Commentary Corpus (Stede, 2004) provides an annotated data set comprised of 176 articles from the newspaper 'Märkische Allgemeine Zeitung'. Besides coreference, the corpus features annotation of discourse structure, as well as connectives and their arguments. The coreference annotation is available in a CoNLL-style format. We evaluate our models on this corpus in section 5.4.4.

The DIRNDL corpus (Björkelund et al., 2014) contains 618 documents gathered from German radio broadcasts and is also coreferentially annotated and available in a CoNLL format. However, possessive and relative pronouns are not annotated with coreference, and we only counted 218 instances of annotated personal pronouns (excluding *es*), which is a low count compared to the TüBa-D/Z corpus.

Finally, the Heidelberg Text Corpus (HTC) explored in Strube et al. (2002) and Kouchnir (2004) consists of 242 short texts about sights, historic events, and persons in Heidelberg. Unfortunately, the corpus is only available in the MMAX format (Müller and Strube, 2001).[5] Without the CoNLL format, the official coreference scorer and our ARCS scorer are not applicable and thus we are not able to evaluate our system on this corpus.

5.1.2 Markable extraction

In this section, we describe how we extract the markables, i.e. the noun phrases and pronouns that we consider for coreference resolution. Recall that we view coreference resolution through the scope of downstream applications in Computational Linguistics and NLP. In this perspective, we identify those types of markables whose resolution we deem useful for such subsequent applications.

5.1.2.1 Part-of-speech-based identification of markables

We define the type of markables we aim to extract by their part of speech (PoS) tags. The TüBa-D/Z employs the Stuttgart/Tübingen Tag Set (Schiller et al., 1999, STTS) for the annotation of part of speech of tokens. Among the STTS, we select the following subset shown in table 5.1 which features coreference annotation in the corpora investigated.

The TüBa-D/Z contains coreference annotation for additional pronouns, but these feature few instances in the data compared to the ones listed above. Additionally, related work on German pronoun resolution has mainly focused on personal and possessive pronouns (Strube et al., 2002, Schiehlen, 2004, Hinrichs et al., 2005, Wunsch, 2010, inter alia).

[5] Initial efforts to convert the MMAX format to CoNLL proved to be cumbersome and were not fruitful due to idiosyncrasies in the MMAX format. Due to time constraints, we were not able to pursue the effort further, unfortunately.

PoS	**Description**	**Example DE**	**EN**
Nouns			
NN	Common noun	*Anwalt*	*(attorney)*
NE	Named entity	*Danzig*	*(Danzig)*
Pronouns			
PPER	Personal pronoun	*sie*	*(she / they)*
PPOSAT	Possessive pronoun	*ihr*	*(her / their)*
PRELS	Relative pronoun	*Leute, die*	*(people who)*
PRELAT	Attributing rel. pronoun	*Menschen, deren*	*(people whose)*
PDS	Demonstrative pronoun	*dies*	*(this)*

TABLE 5.1: PoS tags of head tokens of NPs and pronouns considered for coreference resolution in our system.

Regarding reflexive pronouns (PRF: *sich, mich, dich, euch*), the TüBa-D/Z coreference annotation guideline (Naumann, 2007) employs two categories. The first category consists of uses of the reflexive pronoun *sich* as part of an inherently reflexive verb, e.g. *sich ereignen (to happen itself*)*. The reflexive in these cases is bound to the verb, and the verb cannot be used correctly without it. Here, the reflexive pronoun arguably does not refer to an antecedent and is therefore not annotated as anaphoric. The second category covers the use of reflexives in combination with verbs which do not depend on them, e.g. *sich waschen (to wash oneself)* vs. *etwas waschen (to wash something)*. These cases of *sich* are annotated as anaphoric.

The TüBa-D/Z 9.1 contains 10689 instances of the reflexive *sich*, of which 9284 are annotated as being non-anaphoric, i.e. as part of an inherently reflexive verb. On the one hand, this yields a baseline of 86.86% accuracy of deciding not to resolve *sich*. On the other hand, when the reflexive is annotated as anaphoric, it refers to the grammatical subject of the governing verb in virtually all cases. That is, there is little or no ambiguity involved and, if desired, *sich* can safely be resolved to the aforementioned subject. Here, again, we take the view of higher-level NLP applications and argue that the annotation guideline is too fine-grained. Therefore, we exclude reflexives from our approach. However, it is noteworthy that when we use the common coreference evaluation metrics, which are ignorant of mention PoS types, we lower the upper bound for Recall of our experiments.

Apart from excluding reflexives and certain rare pronouns, we make an additional and arguably big cut: the exclusion of the notorious (because potentially pleonastic) 3rd person pronoun *es (it)*. Related work on English and German pronouns has introduced several approaches to determine the (non-)anaphoricity of *it* and *es*, ranging from simple filters (lists of verbs which subcategorize a pleonastic *it* as the grammatical subject, e.g. *it rains* (Lappin and Leass, 1994)) to elaborate machine learning approaches (Bergsma et al., 2008b).

The TüBa-D/Z 9.1 contains 8891 instances of *es* of which only 706 (7.94%) are anaphoric. This yields a baseline accuracy of 92.06% for not resolving *es*.[6] Furthermore, our initial approaches to address the problem were disappointing. We have found that identifying pleonastic uses of *es* based on heuristics achieves relatively solid Precision accompanied by poor Recall for the anaphoric uses of *es*. That is, identifying anaphoric use of *es* has proven to be more difficult, mainly because of the large imbalance in the training data. Thus, finding indicators for the anaphoric use of *es* has been proven to be futile in our initial approaches.

Given the reasoning above, we decided not to resolve *es* in our experiments, like Klenner and Ailloud (2008). For comparison, (Wunsch, 2010, p. 151) stated that he relied on the gold annotation to decide whether to resolve *es*, i.e. he excluded the anaphoricity detection problem. Schiehlen (2004) also relied on the gold standard for anaphoricity detection of third person pronouns. That is, he did not restrict the anaphoricity check to *es*, but all third person pronouns.

Beside the restrictions listed above, we process all NPs and pronouns of the PoS subset listed in table 5.1. More specifically, we consider all encountered pronouns to be anaphoric, and if we find compatible antecedent candidates (which meet several filter criteria), we resolve them.

5.1.2.2 Identification of markable boundaries

As outlined in section 1.5.3, correctly identifying markable boundaries (i.e. the span of tokens that denote an NP) is of crucial importance, since evaluation only deems system mentions to be resolved correctly if their boundaries perfectly align with those of the gold mentions. An easy way of ensuring correct mention boundaries would be to only extract gold mentions from the data. However, we aim to develop an end-to-end coreference resolution approach which can be applied on raw text where gold mention boundaries are not known, which is a more realistic setting.

We apply the following heuristics to identify markable boundaries. Traversing the CoNLL-style format of the TüBa-D/Z (e.g. table 1.1), we look at the PoS tag of each token. If we encounter a pronoun PoS tag from the set listed in table 5.1, we assume that the pronoun features a single-word markable extension. That is, we extract the pronoun as a markable denoted by a singe token.

For nouns and named entity tokens, we find the maximal NP projection by gathering all tokens that (recursively) link to the noun or name token, as indicated by the dependency

[6] By comparison, in the state-of-the-art corpus for coreference resolution for English, OntoNotes (Pradhan et al., 2013), 771 out of 1318 (58.5%) of the occurrences of the pronoun *it* are anaphoric.

parse. We then cut relative clauses and punctuation etc. at the end of the projection with heuristics developed over the development set.

To cope with mention boundary problems, we check whether all gold mentions are represented as markables once the end of a document is reached. If a gold mention lacks a corresponding markable, we traverse the markables and heuristically determine[7] the closest matching one and adjust its boundaries to those of the gold mention.

By comparison, Hinrichs et al. (2005) allowed system mentions to contain the gold mentions and vice versa in order to be counted as correct in their evaluation. Wunsch (2010), p. 123, counted system antecedents as correct if they featured the same head token as the corresponding gold antecedent. Versley (2010), p. 213, collected all projections of an NP's head and counted system antecedents as correct if any of the corresponding projections denoted the gold antecedent.

Once markable spans are identified, we create feature vectors that describe the markables. Table 5.2 shows the initial vector representation we create for the possessive pronoun and its antecedent in our running example, i.e. [*ihren*] and [*Arbeiterwohlfahrt*] (abbreviated as *AWO*). This representation, with several modifications and additions, constitutes the input for our coreference models.

markable ID	**sent. number**	**start token**	**end token**	**PoS**	**person**	**gender**	**number**	**gram. role**	**animacy**	**gov. token ID**	**gov. verb**	**NE type**	**connector**	**lemma**
11	4	4	6	NN	3	FEM	SG	SUBJ	-	13	entlassen	ORG	-	AWO
12	4	7	7	PPOSAT	3	*	*	DET	*	13	entlassen	-	-	ihr

Table 5.2: Example of markable instances represented as feature vectors. '*' denotes underspecified feature values.

All features shown in table 5.2 are directly extracted from the TüBa-D/Z data, except the animacy feature and the gender of person entities. To determine the gender of person entities, e.g. "Bill Clinton", we use a list of first names divided into male and female names which was gathered from the Internet. We determine animacy of common noun markables by a look up in GermaNet (Hamp and Feldweg, 1997), a German word net. We deem a markable animate if its head noun is a hyponym of the synset *Mensch (human)*.

Having outlined the data and our markable extraction procedure, we next discuss the implementation of the coreference models.

[7] For example by removing or adding PP attachments, adverbs, or parentheses.

5.2 Model implementations

In this section, we specify how we implement the incremental entity-mention and the mention-pair models which we evaluate later on in this chapter. In chapter 2, we have outlined the conceptual differences of the two models and argued for the advantages of our incremental entity-mention model. Here, we focus on how certain functions of the algorithms are realized in their implementations. This includes e.g. choosing appropriate filter settings for certain parameters, such as sentence distance between two potentially coreferring markables etc.

5.2.1 Outline of the algorithms

In section 2.1, we have outlined the standard algorithms for training and testing in the Soon et al. (2001) implementation of the mention-pair model. In the previous chapter, we have discussed how German coreference resolution approaches adapted this model (Strube et al., 2002, Hinrichs et al., 2005, 2007, Klenner and Ailloud, 2009, Wunsch, 2010). Since related work used different evaluation protocols and different test sets, we implement a related work baseline which is based on the mention-pair model. This enables us to more directly compare our incremental entity-mention model to previous work than contrasting F-scores obtained in different evaluation settings.

Table 5.3 juxtapositions the two algorithms we investigate in our experiments. Note that we depart slightly from the original mention-pair algorithms presented in Soon et al. (2001) and use an adaption in line of Hinrichs et al. (2005), Klenner and Ailloud (2009), Wunsch (2010). That is, for personal and possessive pronouns, we query the preceding three sentences for antecedent candidates, both during the creation of training instances and during automatic resolution of the pronouns. During testing, we choose the highest weighted candidate as antecedent within the three sentence window, i.e. we adapt a best-first heuristic.[8]

Comparing the two algorithms, we see that the mention-pair model lacks three main features compared to our incremental entity-mention model:

- The incremental formation of the coreference partition (lines 11-13), which alleviates the need for a final clustering step
- The division of antecedent candidates into those stemming from the coreference partition and those coming from the buffer list (lines 2-7)
- The disambiguation of resolved anaphors (line 10)

Algorithm: Mention-pair model

Input: Markables
Output: Coreference partition

```
for $m_i \in Markables$ do
    for $m_j \in BufferList$ do
        if $compatible(m_j, m_i)$ then
            $Candidates \oplus m_j$
    $ante \leftarrow get\_best(Candidates)$
    if $ante \neq \varnothing$ then
        $Pairs \oplus \{ante \oplus m_i\}$
    $BufferList \oplus m_i$
$CorefPartition \leftarrow trans\_merge(Pairs)$
return $CorefPartition$
```

Algorithm: Entity-mention model

Input: Markables
Output: Coreference partition

```
for $m_i \in Markables$ do
    for $e_k \in CorefPartition$ do
        if $compatible(e_{k_n}, m_i)$ then
            $Candidates \oplus e_{k_n}$
    for $m_j \in BufferList$ do
        if $compatible(m_j, m_i)$ then
            $Candidates \oplus m_j$
    $ante \leftarrow get\_best(Candidates)$
    if $ante \neq \varnothing$ then
        $ante, m_i \leftarrow disambiguate(ante, m_i)$
        if $\exists e_k \in CorefPartition : ante \in e_k$ then $e_k \oplus m_i$
        else
            $CorefPartition \oplus \{ante \oplus m_i\}$
            $BufferList \ominus ante$
    else
        $BufferList \oplus m_i$
return $CorefPartition$
```

Table 5.3: Mention-pair vs. entity-mention algorithms used in our experiments.

We have outlined in section 2.3 that the function *disambiguate*$(\cdot,\cdot)$ propagates all semantic and morphological information from the antecedent onto the anaphor, effectively disambiguating its morphological properties in the case of underspecification. Additionally, we project the grammatical role of an antecedent to a possessive pronoun in the case that both are in the same sentence. The salience of an entity depends to a large degree on the grammatical role of its last mention in our entity-mention model. In the case that an entity occurs e.g. first as a subject (high salience) and then as a possessive pronoun (lower salience) in a sentence, we want to keep the high salience evoked by the subject mention. In other words, we do not want the salience of the entity to degrade in a sentence only because there is also a possessive pronoun mention of that entity. We found in Tuggener and Klenner (2014) that this technique improves performance overall.

5.2.2 Morphological agreement and distance constraints

One main function in both algorithms is *compatible*$(\cdot,\cdot)$. This function determines morphological compatibility of a pronoun and a potential antecedent candidate. As we have seen in the previous chapter, filtering based on morphological agreement is a crucial step to reduce the number of potential antecedent candidates for German pronouns, which

[8] Cf. section 2.1 for an explanation of the best-first vs. the closest-first heuristic.

removes up to 50% of the incorrect potential candidates. However, underspecification of certain German pronouns complicates this step, as these underspecified pronouns allow for specific subsets of all possible morphological attributes. That is, these pronouns are not fully underspecified, which would license any candidate antecedent. Therefore, testing morphological compatibility is not a simple unification process. To account for the specific possible combination of morphological properties, we implement a PoS- and lemma-based filtering scheme for matching number and gender properties, similar to Wunsch (2010). Note that all antecedent candidates and pronouns have to match regarding their person feature.[9] Also, a personal pronoun cannot link to an antecedent governed by the same verb as the pronoun. For example, in the sentence "Peter likes him", "him" cannot refer to "Peter" due to binding constraints. Our filtering scheme works as follows:

- **Lemma-based filtering**: We first ensure exclusiveness of gender incompatible pronouns by lemma-based matching. That is, pairs are discarded if the pronoun lemma is either *sie* (*she/they*) or *ihr* (*her/their*) and the antecedent candidate is singular, masculine, or neuter. Conversely, if the pronoun lemma is either *er*(*he*) or *sein* (*his/its*) and the pronoun lemma of the antecedent candidate is *sie* (*she/they*) or *ihr* (*her/their*), the pair is discarded.
- **Gender and number matching for non-possessive pronouns**: Pairs of personal, relative, and demonstrative pronouns and potential antecedent candidates are licensed if they match in number and gender, i.e. if they share the same respective values. This constraint is relaxed in the following manner. If the pronoun or the antecedent candidate is underspecified in both number and gender, the pair is licensed. If either the pronoun or the antecedent is underspecified in number, but gender matches, the pair is licensed. Conversely, if either the pronoun or the antecedent is underspecified in gender but number matches, the pair is licensed.
- **Possessive pronouns**: For personal pronouns, we license antecedent candidates based on the lemma of the possessive pronoun. That is, for *sein* (*his/its*), the candidate has to be masculine or neuter, but not feminine and singular. For *ihr* (*her/their*), the candidate has to be either plural, or singular and feminine.

Both the mention-pair and our entity-mention model use the same filtering scheme. The difference between the two models w.r.t. to morphological filtering becomes evident when a pronoun which is per se underspecified serves as an antecedent candidate for another pronoun. For example, assume we resolve an instance of the personal pronoun

[9] Cf. section A.1 for the heuristics we apply to resolve first person pronouns to their third person antecedents.

$[sie]^{Sg.}$ (*she*) and the antecedent candidate is a possessive pronoun $[ihr]^*(her/their)$. Let us assume we have already resolved $[ihr]^*$ to $[Frauen]^{Pl.}$ in both models. The mention-pair model would generate the pair $[ihr]^* - [sie]^{Sg.}$, although they are exclusive. By contrast, the entity-mention model would have projected the morphological properties to the possessive pronoun (line 10 in the right algorithm in table 5.3) and thus would not create the pair $[ihr]^{Pl.} - [sie]^{Sg.}$.

As stated in the previous section, we allow for a window of three preceding sentences to look for antecedent candidates for personal and possessive pronouns. Obviously, relative pronouns can only bind to antecedents in the same sentence. Demonstrative pronouns present a difficult class, since our assumption that all pronouns are anaphoric here leads to many false positives, as can be seen in figure 5.1. The difficulty of resolving demonstrative pronouns is also documented in Schiehlen (2004) and Strube et al. (2002). Schiehlen reported an overall pronoun F-score of 65.4%, but demonstratives only achieved an 16.6% F-score. Similarly, Strube et al. reported an F-score of 82.79% for personal pronouns, but only 15.38% F-score for demonstratives. To cope with the anaphoricity detection problem, we limit the search for antecedent candidates to the current and previous sentence. This restriction can be rooted in linguistic theory on entity accessibility in short-term memory. Theories on givenness (Ariel, 1988, Gundel et al., 1993, inter alia) state that demonstratives can be used to refer to entities that are activated in the hearer's short-term memory, but are not necessarily in the discourse focus at the time of the occurrence of the demonstrative. That is, demonstratives are generally used to refer to entities which have been mentioned very recently but are not currently the most salient ones. We thus argue that limiting sentence distance more strictly for antecedent candidates for demonstrative pronouns is a reasonable approach to tackle the anaphoricity problem. That is, if no candidates are found in the current or previous sentence, we do not resolve a demonstrative pronoun.

5.3 Pronoun Resolution

In this section, we evaluate the performance of our incremental entity-mention model and compare it to various baselines. We first assess the difficulty of the resolution problem based on the average count of antecedent candidates per pronoun type. We then discuss the issue of the imbalance of positive and negative examples in the training data and present a feature weighting scheme which accounts for the imbalance. We evaluate this weighting scheme by comparing it to more elaborate machine learning frameworks. We overview feature sets used in our approach and in related work and compare them empirically. Finally, we perform an error analysis and discuss pronoun instances that are difficult to resolve.

5.3.1 Estimation of the difficulty of the antecedent selection

We have argued that pronoun resolution in German is more difficult than e.g. in English, since i) certain German pronouns are morphologically underspecified, and ii) German pronouns can refer both to animate and inanimate entities. Therefore, the number of candidates that have to be considered for a given pronoun is potentially higher than in English. Recall, however, that both filtering based on morphosyntactic constrains and distance, and the pair generation mechanics in the entity-mention model reduce the initial set of candidates.

To assess the difficulty of the pronoun resolution task for German, we count how many candidates there are on average per pronoun type. We compare the numbers for the mention-pair and the incremental entity-mention model, thereby quantifying the reduction of candidates when moving to an entity-mention architecture.

We discussed in section 3.4 that Mitkov (2001) defined the *critical success rate* to measure pronoun performance in only those cases where there are competing antecedent candidates (i.e. more than one) for a pronoun. In this spirit, we further divide the analysis into cases where there are multiple candidates and all cases, which includes cases with only one candidate. We plot the average number of candidates in figures 5.2 and 5.3. The numbers are obtained from our test set, i.e. on the first 690 documents of the TüBa-D/Z v9.1. The plots indicate the mean and standard deviation of the numbers of candidates per pronoun type.

The left plots in figures 5.2 and 5.3 show the average number of antecedent candidates for pronoun instances with multiple candidates. The plots to the right show the numbers for all pronouns, i.e. including those with only one candidate. For example, in the mention-pair model (figure 5.2), we generate 6.13 candidates on average for personal pronouns (PPER) when there are multiple candidates (left plot). By contrast, we only create 4.89 candidates on average for personal pronouns in the entity-mention model (figure 5.3) when there are several candidates (left plot). Furthermore, it occurs more frequently that there are multiple candidates in the mention-pair model (i.e. 2649 times; indicated below PPER) than in the entity-mention model (2261 times). We can interpret this in the following way: The problem of selecting the correct antecedent in the mention-pair model is more difficult than in the entity-mention model, because there are on average more candidates to choose from, and it is more often the case that there are multiple candidates to choose from.

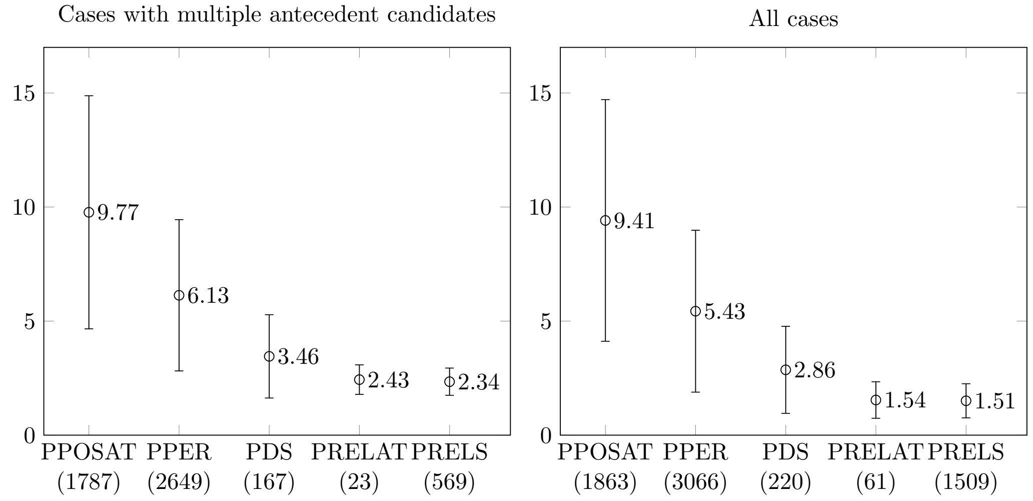

FIGURE 5.2: Average number of antecedent candidates per pronoun type in the mention-pair model.

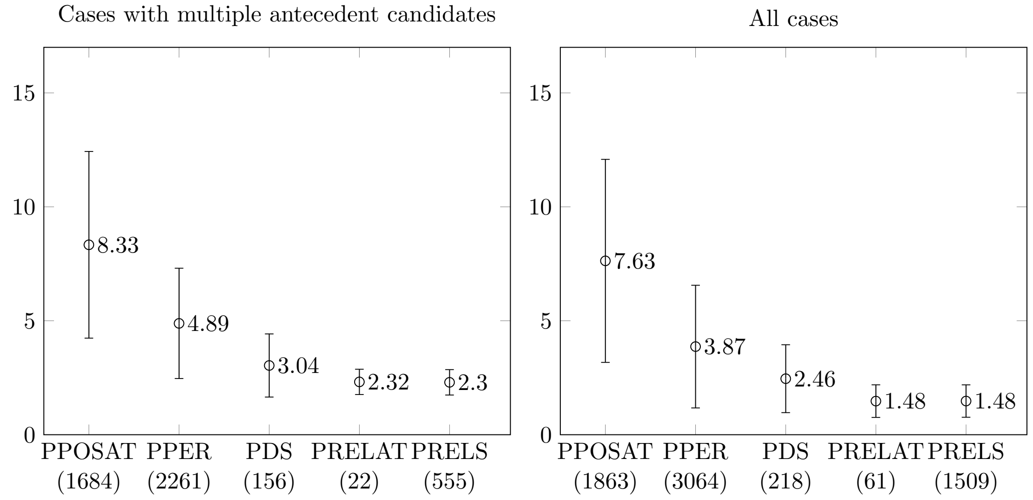

FIGURE 5.3: Average number of antecedent candidates per pronoun type in the incremental entity-mention model.

Note that the plots only show the count of pronoun instances that actually have compatible antecedent candidates. Looking at the plots to the right in both figures, we see that the entity-mention model, compared to the mention-pair model, only lacks compatible candidates for 2 instances of personal pronouns (3066 vs. 3064) and demonstrative pronouns (220 vs. 218).

Furthermore, the means in the right plots are lower, since the single candidate cases are included, lowering the average number. However, they are only marginally lower, indicating that for the most pronoun instances there are multiple antecedents to choose from. In the mention-pair model, for the personal pronouns (PPER), we create multiple candidates in 86% of the cases (2649/3066) and for possessive pronouns (PPOSAT) in 96% of the cases (1787/1864). In comparison, in the entity-mention model, we generate multiple candidates in 74% of the cases (2261/3064) for personal pronouns, and for possessive pronouns there are multiple candidates in 90% (1684/1863) of the cases.

The relative pronouns (PRELS and PRELAT) and the demonstratives (PDS) show different effects compared to the personal and possessive pronouns. When all cases are considered, the mean of the relative pronouns is below 2, and when only cases with multiple candidates are considered their mean is near 2, the minimum number of candidates in this view. In the entity-mention model, we need to choose a candidate only in 37% of the cases (555/1509; PRELS), or in 36% (22/61; PRELAT), respectively. The numbers are similarly low for the mention-pair model. Obviously, relative and demonstrative pronouns have few antecedent candidates, because we only search the current sentence for antecedents given relative pronouns and extend the window to the previous sentence for demonstrative pronouns.

The highest number of candidates is generated for the possessive pronouns (PPOSAT). Like for the personal pronouns (PPER), we search a window of three sentences to the left of possessive pronouns for compatible candidates. However, since the most frequent possessive pronouns (*sein (his/its), ihr (her/their)*) are underspecified, we generate on average twice as many candidates (mean 7.63 for all cases; entity-mention model) for the possessive pronouns as for the personal pronouns (mean 3.87 for all cases; entity-mention model) within the three sentence window.

Finally, we note that the standard deviations are rather high for all pronoun types and models, indicating that the number of candidates varies quite strongly w.r.t. the mean. To the best of our knowledge, we are the first to make visible the heterogeneous distribution of the candidate counts. This potentially opens up interesting opportunities for future research. Inspecting cases with a high or low count of candidates and devising different processing strategies for them might be a fruitful endeavour. In our current approach, we do not distinguish between cases with many or few candidates.

Overall, we saw that the entity-mention model reduces the number of candidates and the pronoun instances that have multiple candidates. To exemplify how the entity-mention model reduces candidates consider the sentence in figure 5.4.

To resolve the relative pronoun "der$_2$", the mention-pair model would pair the pronoun with the three morphologically compatible markables "Mann$_1$", "seinem$_1$", and "Hund$_2$", thus creating three pairs (arcs with question marks above the example sentence). In the incremental entity-mention model, we would have resolved "seinem$_1$" to "Mann$_1$" (indicated by the dashed green arc below the example sentence), before encountering the relative pronoun "der$_2$". Therefore, the model would only pair "der$_2$" with "seinem$_1$", which is the most recent and thus only accessible mention of the "Mann$_1$"

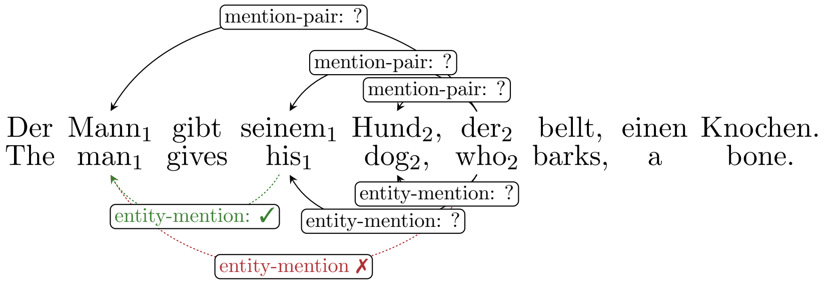

Figure 5.4: Differences in the pair generation mechanics between the mention-pair and entity-mention model. Identical numbers in the subscripts denote coreference between words. Question marks (?) indicate proposed pairs, check marks (✓) pairs classified as positive, and crosses (✗) pairs not proposed by the pair generation mechanics.

entity, but not with the "Mann$_1$" mention (indicated by the dashed red arc) and "der$_2$" with "Hund$_2$".

Obviously, the reduction of the number of candidates in the entity-mention model compared to the mention-pair model is only beneficial when the correct antecedents are not removed. To assess whether and how frequently the entity-mention discards the correct antecedent, we count how often the correct antecedent is accessible in both models. We do so for all pronoun instances, i.e. not only for those that have multiple candidates, in order to directly compare the models over the same set of pronouns. Table 5.4 shows the results.

	PPER	**PPOSAT**	**PRELS**	**PDS**	**PRELAT**	**ALL**
	(2921)	**(1707)**	**(1402)**	**(167)**	**(59)**	**(6256)**
M-P	87.50	94.26	89.44	84.43	84.75	89.67
E-M	85.93	93.32	89.37	82.63	84.75	88.62

Table 5.4: Accessibility of the correct antecedent in the mention-pair (**M-P**) and entity-mention (**E-M**) models (percentages) in the test set. The numbers under the PoS tags denote absolute counts.

The head column in table 5.4 indicates the PoS tags of the pronouns, and the number of anaphoric gold mentions underneath, i.e. pronouns that are annotated as anaphoric in the gold standard.[10] The table cells show the percentage of instances where the correct antecedent is available in the respective models. Overall, we see that the entity-mention model removes the correct antecedent in only 1% of all pronoun instances compared to the mention-pair model.[11] Also, we see that availability of the correct antecedent is limited to a significant degree in both models. Missing antecedents are caused by

[10] Note that we exclude cataphors, since our system does not handle them.

[11] Note that there is a difference in the instance counts of the pronoun types compared to the plots in figure 5.3. For example, personal pronouns have a count of 3064 in the plot, while in table 5.4 there are only 2921 instances. The reason for the lower number in the table is that we only considered the gold mention pronouns for measuring the antecedent availability. However, to calculate the average antecedent count, we used all pronoun instances the systems resolve.

problems in the markable extraction step and distance filters. The markable extraction step sometimes misses the correct NP boundary of the correct antecedent, and therefore it cannot be identified among the markables. For personal and possessive pronouns, we filter candidates more than three sentences away, which sometimes excludes the correct antecedent from the candidate set.

To compare our average count of antecedent candidates to related work, i.e. Wunsch (2010)[12], we calculate the average over all pronouns. Wunsch (2010), p. 194, reported an average ratio of positive and negative instances of 1:4.29 over all pronouns. That is, on average, 4.29 negative instances of an antecedent candidate and a pronoun are created (and one positive instance, obviously). Wunsch investigated personal, possessive, and reflexive pronouns. By contrast, we examine personal, possessive, relative, and demonstrative pronouns. On average, we create 4.31 antecedent candidates per pronoun. For personal pronouns, Wunsch reports a ratio of 1:3.27 and for possessive pronouns a ratio of 1:6.35. Our average antecedent counts for personal pronouns is 3.87, and 7.63 for possessive pronouns. Thus, we can infer that our strategies for filtering antecedent candidates perform similarly to that of Wunsch. Also, the difficulty of the problem is roughly the same in Wunsch's approach. To address the imbalance problem, Wunsch (2010) explored instance sampling, i.e. down-sampling the negative instances, to achieve a ratio of 1:2, which gave best results in his experiments. By contrast, we include the class imbalance in our feature weighting scheme, which we present in section 5.3.2.

Our analysis complements the intuition that personal and possessive pronouns are more difficult to resolve than relative pronouns. Relative pronouns always have their antecedents within the same sentence. Therefore, we need to consider fewer candidates and are less likely to make a mistake when selecting an antecedent. However, we need to take into account that we have not made sure that the correct antecedent is among the candidates in the analysis of the average antecedent candidate counts. That is, an additional problem arises when the correct antecedent is not present; a problem we cannot tackle with feature engineering. Therefore, the problem of selecting the correct candidate is further complicated by the (non-)availability of the correct antecedent. To tackle this problem, we have proposed the ARCS success rate, i.e. the ratio of correctly resolved anaphors and resolvable anaphors in section 3.4.3. This measure helps us to weed out the cases irrelevant for our investigation of classifier performance. Before exploring performance of our classifiers with this measure, we outline a feature weighting scheme and the feature sets relevant to our experiments.

[12] Schiehlen (2004) also reported average antecedent candidate counts, but only for individual filtering techniques, i.e. he gave no cumulative averages.

5.3.2 A flexible and simple scheme for feature weighting

In this section, we introduce a simple scheme for calculating feature weights and ranking antecedent candidates which combines two ideas we developed in previous work.

In Klenner and Tuggener (2010), we presented a simple maximum likelihood-based measure to rank antecedent candidates according to their grammatical functions. The salience of a grammatical function gf was calculated by the number of gold mentions bearing that function divided by the total number of gold mentions, $\frac{|gold\ mentions\ with\ gf|}{|gold\ mentions|}$, which yields a preference ranking of grammatical functions. The candidate with the highest ranking grammatical function was then selected as antecedent. If several candidates shared the same function, the most recent one was selected. We found that this simple candidate ranking approach performed surprisingly well compared to a kNN classifier with an extensive feature set.

Instead of deriving a ranking of a single feature, e.g. grammatical functions, we interpret the count ratio as a probabilistic weight and use this approach to weight all features. The candidate with the highest weight product over all features then is selected as the antecedent, much like in Naive Bayes classification. A similar strategy was introduced in Ge et al. (1998). Ge et al. calculated conditional probabilities for seeing a pronoun type p given a word w with feature a, i.e. w_a. For example, to obtain a weight for the animacy feature, Ge et al. calculated the following conditional probability:

$$P(p|w_a) = \frac{|w_a\ is\ ante\ of\ p|}{|w_a|} \tag{5.1}$$

This conditional probability can be interpreted as the likelihood of animate words ($a = animate$) to emit a certain pronoun, e.g. *he*. The conditional probability is calculated by dividing the count of animate words that are antecedents of *he* by the count of all animate words.

Our salience measure for grammatical functions can also be interpreted as such a conditional probability:

$$P(gf|gold\ mentions) = \frac{|gold\ mentions\ with\ gf|}{|gold\ mentions|} \tag{5.2}$$

which estimates the likelihood of seeing a grammatical function given the gold mentions, which is closely related to the formulation in Ge et al. (1998). Based on these similarities, we apply the following general formula to weight a value i of a feature x, i.e. x_i, in the domain of pronoun resolution. We calculate how likely it is for a feature value to occur with the correct antecedent, or, more concisely, how likely it is to see the correct antecedent given the feature value. This is given by the following formula:

$$w(x_i) = P(y_{pos}|x_i) \approx \frac{|x_i \cap y_{pos}|}{|x_i|} \tag{5.3}$$

where $|x_i| = |x_i \cap y_{pos}| + |x_i \cap y_{neg}|$, i.e. the overall occurrence count of the feature value, where y_{pos} indicates a correct antecedent and y_{neg} an incorrect antecedent. To score an antecedent candidate a_k for a pronoun p, we simply multiply all applicable feature weights $w(x_i)$, i.e.:

$$score(a_k) = \prod_{\forall x_i \in f(a_k,p)} w(x_i) \tag{5.4}$$

where $f(a_k, p)$ is a feature mapping function that maps all applicable features to an antecedent candidate and a pronoun. We then select the candidate with the highest score as the antecedent.

However, there is a complication. In Tuggener and Klenner (2014), we introduced features that only apply to certain contexts, i.e. certain combinations of antecedent candidates and pronouns. For example, we introduced the named entity class of the antecedent as a feature. This feature only applies if the antecedent at hand actually denotes a named entity (NE). The feature does not apply to common noun entities, because they have no named entity class. In Tuggener and Klenner, we showed how the first-order predicate logic formulas in Markov Logic Networks (MLN) can be used to constrain features to such contexts and that it is beneficial to do so. Therefore, we aim to port the expressiveness of MLN formulas to our simpler weighting scheme.

To do so, we have to address the problem that not all candidates have the same number of applicable features. Using equation 5.3, all weights are in the range $0 < w(x_i) < 1$, since we take a conditional probability as the feature weight. When multiplying the weights, as in equation 5.4, having more features always lowers the score of a candidate with more features than others, i.e. a candidate is always punished for having more features than another candidate.

Consider again the named entity (NE) class feature. Let us assume we want to resolve an instance of the pronoun *er* (*he*) and we see two candidates, *Obama* and *Pfannkuchen* (*pancake*), both of which occur as subjects with weight 0.5. Now, we add the weight for the NE class, which only applies to the *Obama* candidate. *Obama* has the NE class *PER* (person), and we have calculated a high weight for this class, i.e. 0.8. However, since we multiply the weights of a candidate to calculate its score, the overall score for the Obama candidate is lowered by the high feature weight of PER, as depicted below.

	Grammatical function	**Named entity class**	$\prod$
Obama	0.5	0.8	0.4
Pfannkuchen	0.5		0.5

This is the opposite of the desired outcome when introducing specific features for certain candidates.

Alternatively, we could take the sum of the weights to score the candidates. However, having additional features in that case always increases the score of a candidate, but we want to capture the impact of additional features w.r.t. their ability to identify the correct antecedent. For example, the NE class LOC (location) should have a low weight for personal pronouns. However low this weight is in the range $0 < w(x_i) < 1$, it always increases the overall score of a candidate when adding it to the weight sum. Thus, summing the weights is not a viable alternative.

Our solution to handling weights that only apply to certain contexts lies in adjusting the weight range so that 'good' feature values (e.g. the NE class PER for antecedent candidates for personal pronouns) increase the weight product, and 'bad' features (e.g. the NE class LOC for candidates for personal pronouns) lower the product. We achieve this by incorporating a bias factor into formula 5.3 that allows weights to be larger than 1. Ideally, 'good' feature values obtain weights larger than 1 and thereby increase the weight product, while the weights of 'bad' feature values range below 1 and therefore decrease the weight product. Features with no discriminatory power w.r.t. antecedent selection should have weight 1 and thus not affect the weight product. To achieve this, we need to scale up the weights by a constant term, and for this purpose we make use of the class imbalance of the training data.

For simplicity of exposition, let us assume that our incremental entity-mention model generates 5 antecedent candidates per pronoun on average. Recall that there is always only one correct antecedent among the candidates. Thus, the prior probability of a feature occurring with the correct antecedent is the count of correct antecedents divided by all antecedents, i.e. $\frac{|\mathcal{Y}_{pos}|}{|\mathcal{Y}_{pos}|+|\mathcal{Y}_{neg}|} = \frac{1}{5}$ on average.[13] We can use the inverse of this imbalance as the bias factor to shift the weights to the desired range outlined above. To do so, we multiply the inverse of the imbalance with the weight of a feature obtained in equation 5.3:

[13] Note that $\mathcal{Y}_{pos}$ and $\mathcal{Y}_{neg}$ denote all correct and incorrect antecedents, respectively, while y_{pos} and y_{neg} denote correct and incorrect antecedents that occur with feature x_i.

$$w(x_i) = \frac{|x_i \cap y_{pos}|}{|x_i|} * \frac{|\mathcal{Y}_{pos}| + |\mathcal{Y}_{neg}|}{|\mathcal{Y}_{pos}|} \tag{5.5}$$

A non-informative feature now obtains a weight of 1. Consider a feature that occurs with all antecedent candidates in every training instance. The conditional probability of seeing the correct antecedent given the feature as calculated by formula 5.3 equals the prior probability of seeing the correct antecedent, i.e. $\frac{1}{5}$. If we take this probability as the weight and multiply it with the inverse prior probability of seeing the correct antecedent, i.e. $\frac{5}{1}$, the weight becomes 1 and thus does not affect the feature weight product, i.e. the score of a candidate. All features that score a conditional probability above $\frac{1}{5}$ w.r.t. emitting the correct antecedent will obtain weights higher than 1, and all features with a conditional probability below $\frac{1}{5}$ will yield weights below 1 after multiplication with the bias factor. Thus, the inverse of the class imbalance provides us with a bias term that produces the weight range we desire.

Returning to our previous examples, the weights and their products now look as follows. We have introduced the class bias of 5, i.e. all weights are multiplied by 5. Incorporating the named entity class feature now has the desired effect, i.e. increasing the weight of the NE candidate with class *PER*:

	Grammatical function	**Named entity class**	$\prod$
Obama	$0.5 * 5 = \mathbf{2.5}$	$0.8 * 5 = \mathbf{4.0}$	**10**
Pfannkuchen	$0.5 * 5 = \mathbf{2.5}$		**2.5**

Obviously, multiplying weights with a constant like the inverted prior probability does not affect comparison of two candidates that have values for an identical set of features. However, it allows us to include specific features for certain individual candidates, such as named entities, and assign weights that reflect their benefit in identifying the correct antecedent.

Note that Naive Bayes classification also multiplies conditional probabilities as feature weights which yield a weight range of $0 \leq w(x_i) \leq 1$. Thus, it is not possible to have specific feature sets for certain candidates for the reason outlined above (additional weights always decrease the weight product). Also, unlike in Naive Bayes, our feature weights do not represent a probability distribution over the classes, since our weight range is $0 \leq w(x_i) \leq \frac{|\mathcal{Y}_{pos}|+|\mathcal{Y}_{neg}|}{|\mathcal{Y}_{pos}|}$. Furthermore, Naive Bayes classification also uses the prior class probability in the weight product (in multinomial classification, at least). By contrast, we include the inverse of the prior class probability of only one class (the correct antecedents) which we use to uniformly shift the weight range of all features in a binary classification setting (incorrect vs. correct antecedent).

In summary, our weighting scheme first assigns weights to feature values based on the probability that they occur with the correct antecedent. The weights are then scaled in order to obtain a weight range that allows us to incorporate arbitrary weights for individual candidates. We calculate weights in this manner for each pronoun type (personal, possessive, etc.) and use the weighting scheme as a baseline for other machine learning classifiers in the evaluation throughout the following sections.

5.3.3 Feature set for pronoun resolution for German

In chapter 4, we have discussed related work on German coreference and pronoun resolution. We argued that none of the approaches introduced features to address the specifics of German pronouns, namely morphological underspecification of certain pronouns (*sie/ihr, sein, die*) and the circumstance that certain German pronouns (*er/sein, sie/ihr*) can refer to both animate and inanimate entities in comparison to e.g. English, where singular pronouns with gender generally refer to persons (*he/his, she/her*).

We have shown in section 2.3 how the incremental entity-mention architecture disambiguates morphological underspecification. In Tuggener and Klenner (2014), we have explored a feature set that incorporates animacy and named entity classes to address the animacy ambiguity and evaluated the impact of these features. Here, we present an extended feature set which featurizes morphology to address underspecification and introduces feature conjunctions that further improve classifier performance. Table 5.5 gives an overview of the two feature sets used in our experiments.

The standard feature set is aimed at collecting features commonly used in mention-pair-based related work on German pronoun resolution (Strube et al., 2002, Kouchnir, 2004, Hinrichs et al., 2005, Wunsch, 2010).[14]

We discuss the extended set in detail in the upcoming sections. It presents a mixture of features and feature conjunctions based on linguistic theory and features designed while investigating errors the classifiers made on the development set.

We also show example weights calculated by our weighting scheme outlined in the previous section. The weight of a feature is calculated by the ratio of the feature occurring with the correct antecedent divided by the total count the feature occurrence, multiplied by the inverse prior probability of seeing a correct antecedent candidate. That is, if we divide a weight by this inverse prior probability, the result denotes the ratio of the feature occurring with a correct antecedent. A benefit of our weighting scheme is thus

[14] Note that we do not add the PoS tag of the pronoun to the set, since we calculate the feature weights for each pronoun type separately. Also, we do not include string matching features in the set, since they are mainly aimed at capturing coreference between nominal mentions.

Standard feature set

Sentence distance	Sentence distance between antecedent and pronon
Markable distance	Distance in markables between antecedent and pronon
Gram. funct. antecedent	Grammatical function of the antecedent
Gram. funct. parallel	Grammatical function parallelism between the antecedent and the pronoun (yes/no)
PoS antecedent	PoS tag of the antedecent
Def. antecedent	Definiteness of the antecedent (induced from determiner)

Extended feature set

Sentence distance w.r.t. connector	Sentence distance w.r.t. presence of discourse connector
Markable distance	Markable distance, only if antecedent and pronoun are in the same sentence
Candidate index	Index of the antecedent in the candidate list
Gram. funct. antecedent	Grammatical function of the antecedent
Sequence gram. funct.	Sequence of grammatical function of antecedent to pronoun, including sentence distance and PoS of antecedent
Sequence gram. funct. PPOSAT head1	Sequence of grammatical function of ante to possessive pronoun's head, including sentence distance and PoS of ante
Sequence gram. funct. PPOSAT head2	Same as GF seq. poss. head1, only applies when possessive pronoun's head is governed by the same verb as the antecedent
Preposition	Preposition if antecedent is a PP
Clause type	Clause type of the antecedent, determined by the grammatical role of the verb governing the antecedent
Clause type sequence	Sequence of clause type of antecedent to clause type of pronoun, conditioned on whether antecedent and pronoun are in the same sentence
Number	Number of the antecedent
Gender	Gender of the antecedent
Animacy	Animacy of the antecedent conjuncted with its number and gender
NE class	NE class, only if the antecedent is an NE

Entity-mention features

Entity age	If the antecedent is in a coreference chain, distance of entity introductory sentence to document beginning
Discourse status	Whether the antecedent is in a coreference chain or not

Table 5.5: Feature sets used for pronoun resolution

that we can inspect feature weights to straight-forwardly evaluate linguistic intuitions. In this light, we give example weights for several instantiations of features we add to the standard set.

5.3.3.1 Distance-based features

The first modification made to the standard feature set is to relate sentence distance to the presence of a discourse connector (**Sentence distance w.r.t. connector**). We observed that pronouns preceded by a discourse connector such as *weil (because)*, *aber (but)* etc. tend to bind to intra-sentential antecedents. Discourse connectors in sentences contribute to cohesion by logically or rhetorically connecting two clauses of a sentence through a discourse relation like explanation or contrast. Intuitively, pronominalized entities on the right hand side of the connector are therefore likely to refer to antecedents on the left side of the connector to which the right hand side is bound by the connector. To verify this intuition, we inspect the weights calculated for sentence distance values w.r.t. the presence of a connector for personal pronouns. Table 5.6 shows the weights.

Sentence distance	**-connector**	**+connector**
0	2.41	2.42
1	1.42	0.56
2	0.26	0.10
3	0.13	0.05

Table 5.6: Weights for sentence distances between antecedent and personal pronouns with (+) and without (-) the presence of a discourse connector.

We see that both features favor intra-sentential candidates (Val.=0) equally. However, the weight for candidates from the previous sentence (Val.=1) is already below 1 for the connector-conditioned weight (+conn.; 0.56). This diminishes the overall weight product of candidates outside the sentence of the pronoun. Candidates in the preceding sentence for pronouns without a connector still get a weight boost (-conn.; 1.42). Thus, the weight analysis seems to support our intuition about relating sentence distance to the presence of a discourse connector.

The next standard feature we modify is the markable distance between an antecedent and a pronoun (**Markable distance**). We only apply this feature to candidates within the same sentence as the pronoun. We empirically found that allowing the feature to trigger for inter-sentential candidates lowers performance. Allowing the feature for inter-sentential candidates produces many sparse features, which, when triggered, give a high weight to its candidates. This deteriorated performance to a small degree. However, limiting the feature to intra-sentential candidates resolved the issue.

The candidate index feature (**Candidate index**) is a novel feature which captures a yet unexplored notion of markable distance. It enumerates the antecedent candidates for a pronoun from right to left, i.e. the closest candidate has index 0, the second closest 1, etc. The intuition behind this feature is that distance between a pronoun and its candidates should be measured in terms of compatible markables and thus relevant markables, instead of based on all intermediate markables. Doing so, the number of different feature values becomes lower than for the markable-based distance feature, as it does not count intervening, incompatible and therefore irrelevant markables between a pronoun and its candidates. Table 5.7 compares the weights for the candidate index and markable distance feature values for personal (PPER) and possessive pronouns (PPOSAT).

Distance	**Candidate index weight**	**Markable distance weight**
	PPER	
0	2.93	2.79
1	1.08	2.93
2	0.39	2.59
3	0.17	2.20
	PPOSAT	
0	5.06	4.95
1	2.01	4.21
2	0.82	3.06
3	0.35	2.00

Table 5.7: Comparison of the candidate index and markable distance feature weights.

As indicated by table 5.7, the weight for markable distance for personal pronouns (last column) only slowly decreases when distance increases, and the weights do not vary strongly. By contrast, the candidate index weights react more directly to increasing distance, i.e. a candidate with index 2 already receives a weight below 1 which decreases its score. The same applies for possessive pronouns. Interestingly, the closest candidate receives a much higher weight than for personal pronouns (5.06 vs. 2.93). Also, the weights for markable distance decrease more quickly for increasing distance for the possessive pronouns. This suggests that proximity is a more relevant factor for the resolution of possessive pronouns than for personal pronouns. We investigate this point again when we measure performance of the baseline which selects the most recent candidate as antecedent in section 5.4.2.

5.3.3.2 Feature conjunctions derived from the entity grid representation of discourse

The next family of features is derived from the entity grid representation of discourse (Barzilay and Lapata, 2008) which models local coherence in texts. The entity grid representation itself is based on Centering theory (Grosz et al., 1995) which models local coherence by tracking sequences of grammatical functions that entities subsequently occur with in coherent discourse. Table 5.8 depicts the example grid and text given by Barzilay and Lapata (2008), p. 6. Note that the inventory of grammatical functions is reduced in the entity grid representation (S=subject, O=direct object, X=all other, -=entity not present).

1 [The Justice Department]$_S$ is conducting an [anti-trust trial]$_O$ against [Microsoft Corp.]$_X$ with [evidence]$_X$ that [the company]$_S$ is increasingly attempting to crush [competitors]$_O$.

2 [Microsoft]$_O$ is accused of trying to forcefully buy into [markets]$_X$ where [its own products]$_S$ are not competitive enough to unseat [established brands]$_O$.

3 [The case]$_S$ revolves around [evidence]$_O$ of [Microsoft]$_S$ aggressively pressuring [Netscape]$_O$ into merging [browser software]$_O$.

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software
1	S	O	S	X	O	-	-	-	-	-	-
2	-	-	O	-	-	X	S	O	-	-	-
3	-	-	S	O	-	-	-	-	S	O	O

Table 5.8: Entity grid representation of entity occurrences in discourse.

The basic assumption of the entity grid model is that there exist certain regularities in these sequences (i.e. the columns in the right table in table 5.8) which interplay with coherence. Regularities in the sequences can be learned from coherent texts and then be used to score coherence in other texts, such as summaries.

The entity grid can also be interpreted as a representation of the coreference partition of a document, since it depicts all occurrences of the entities in the document (singletons and coreferent ones). Based on this view, we create feature conjunctions containing the grammatical roles of the antecedent candidates and the pronoun. That is, we model how likely are the different transitions of the grammatical functions from the antecedent candidates to the pronouns. Since we resolve pronouns in their local context and do not track all individual mentions of coreferent entities, we create bigram sequences, i.e. we do not query grammatical functions of antecedent entities that have occurred multiple times before the pronoun. However, we do not simplify the grammatical roles as Barzilay and Lapata (2008), but use the roles as encountered in the data.

We make two additions to the feature conjunction (**Sequence gram. funct.**). First, we add the sentence distance between antecedent candidate and pronoun. The example grid in table 5.8 shows that certain transitions only occur in relation to a certain sentence distance, e.g. we only see the transition [X,O] for the “Evidence” entity with an intervening sentence, i.e. the original transition is [X,-,O]. Since the coreference partition does not track non-occurrence of entities, as the entity grid does, this information is lost. In other words, if we learned transitions from the coreference partitions in the training data without relating them to sentence distance, we would extract the pattern [X,O] for the “Evidence” entity. However, the pattern occurs here with an intervening sentence. Therefore, we add sentence distance to the conjunction of the grammatical roles. In the case of the “Evidence” entity, the feature conjunction is then [2,X,O].

Second, we add the PoS tag of the antecedent candidate to the feature conjunction. The PoS tag of an entity mention reflects the givenness of an entity. The theory on antecedent accessibility by Ariel (1988) reflects this notion by relating the surface form of an entity mention to its degree of being familiar at a given point in discourse. This yields a hierarchy of surface manifestation forms, which can be coarsely adapted to PoS tags (Martschat and Strube, 2014, e.g.): named entity, common noun, and pronouns.[15] The PoS tag can be thought of as a relativization for the other features in the conjunction. For example, a transition of grammatical roles given a high sentence distance might be more likely to indicate coreference if the antecedent candidate is a named entity, rather than a pronoun.

We experimented with different representations of the aspect of familiarity (e.g. definiteness and coreferential status of the antecedent) and distance (sentences, markables) and found that the conjunction of sentence distance, PoS tag, and the transition of grammatical roles performs best. The final form of the feature conjunction for the "Evidence" entity from the above example would thus be [2,X,O,NN] when viewed from sentence three: It is two sentences away, it occurred with the grammatical role PN (X), now it is mentioned as an object (O), and its last occurrence was a nominal mention (NN).

Table 5.9 lists the 5 highest weighted features derived from the conjunctions for personal pronouns. As the feature conjunction yields many sparse features, we only list those that are seen at least 50 times during training. We see that the feature also captures parallelism of grammatical roles, a feature in the standard set, but calculates weights for parallelism of specific roles and sentence distances.

Weight	**SD**	**GF A**	**GF P**	**PoS A**
3.92	0	SUBJ	OBJD	PPER
3.91	0	SUBJ	SUBJ	PPER
3.87	0	OBJA	OBJA	PPER
3.87	0	PN	SUBJ	PPER
3.86	0	SUBJ	SUBJ	PPOSAT

Table 5.9: Top weighted instantiations of the feature conjunction of sentence distance (**SD**), grammatical role transition from antecedent (**GF A**) to pronoun (**GF P**), and PoS tag of the antecedent (**PoS A**) for personal pronouns.

For possessive pronouns, we model two specific configurations. Since the grammatical role of possessive pronouns is DET (determiner), we are not able to further diversify weight calculation on the syntactic context of possessive pronouns. Therefore, we additionally calculate a weight for the grammatical transition of the antecedent candidate to the syntactic head of the possessive pronoun (**Sequence gram. funct. PPOSAT**

[15] For example, if a discourse refers to Barack Obama, it uses a named entity mention if Obama has not been recently mentioned. If Obama has been mentioned very recently, a pronoun can be used for reference. It could use "the president" if the entity is not completely out of focus etc.

head1). Furthermore, we calculate weights for the specific cases where the head of the possessive pronoun is governed by the same verb as the antecedent candidate (**Sequence gram. funct. PPOSAT head2**). For example, consider the following segment from the test set:

(5) Landesvorsitzende Wedemeier: Ein Buchungsfehler. Im Januar hat die Arbeiterwohlfahrt$_1$ Bremen ihren$_1$ langjährigen Geschäftsführer Hans Taake fristlos entlassen [...]

Regional chairwoman Wedemeier: An accounting error. In January, the Worker Welfare Association$_1$ Bremen has laid off its$_1$ long-term CEO Hans Taake without notice [...]

For the possessive pronoun *ihren*, we create the feature for the two compatible antecedent candidates *Landesvorsitzende Wedemeier* and *Arbeiterwohlfahrt Bremen.* The feature for the former is [1,ROOT,OBJA,NE], since the candidate is in the previous sentence, its grammatical role ROOT, the possessive pronoun's head role is direct object (OBJA) and the candidate is a named entity. For the second candidate, *Arbeiterwohlfahrt Bremen* we also instantiate the feature, i.e. [0,SUBJ,OBJA,NN]. Additionally, we create this feature without the sentence distance, i.e. [SUBJ,OBJA,NN], since the *Arbeiterwohlfahrt Bremen* candidate is governed by the same verb as the possessive pronoun's head *langjährigen Geschäftsführer Hans Taake*, i.e. *entlassen.* Doing so, we intend to capture specific transitions from an antecedent to the possessive pronoun's head, given that both are governed by the same head. Table 5.10 lists weights for the most frequent instantiations of the feature.

Weight	GF A	GF P	PoS A
7.97	SUBJ	OBJA	PPER
7.86	SUBJ	PN	PPER
7.69	SUBJ	OBJA	NN
6.07	SUBJ	PN	NN
0.54	PN	PN	NN

Table 5.10: Weights for the most frequent instantiations of the feature conjunction of grammatical function transition from antecedent (**GF A**) to pronoun (**GF P**) and PoS tag of antecedent (**PoS A**) for possessive pronouns, specifically for cases where the possessive pronoun's head is governed by the same verb as the antecedent candidate.

The table shows that the transition evoked by the second candidate in our example (*Arbeiterwohlfahrt Bremen*) is very frequent and receives a high weight. The weight suggests that when possessive pronouns are determiners of direct objects, the likelihood of these pronouns referring to the subject of the verb governing the direct objects is high. By contrast, the likelihood of possessive pronouns whose heads are in prepositional noun phrases to refer to antecedents in prepositional noun phrases governed by the same verb as the pronouns' head is rather low (last row).

5.3.3.3 Approximation of syntactic embedding

Related work has often used depth of embedding of the antecedent candidate as a feature. The feature denotes numerically the depth of embedding of the clause that the candidate occurs in, i.e. the main clause has depth 0 etc. Since we work with dependency parses and do not transform them into constituents, we approximate the feature by taking the grammatical function of the verb governing the candidate as a feature. This also captures how antecedent candidates and pronouns are embedded in the sentences on the clause level. The feature value for **Clause type** is thus the grammatical function of the verb governing the antecedent candidate. The intuition behind the feature is that antecedents governed by verbs in subclauses, such as relative clauses, are less salient and therefore less likely to be pronominalized than candidates governed by root verbs. The weights seem to support this intuition. Candidates in main clauses (root: 1.07) are more likely to be pronominalized compared to candidates in relative clauses (rel: 0.69).

Additionally, we create a feature conjunction of the clause type of the antecedent candidate and the clause type of the pronouns and relate it to whether the candidate and the pronoun are in the same sentence (**Clause type sequence**). We assume that there are typical patterns of antecedents in e.g. root clauses that are pronominalized in subclauses in the same sentence. The weights indicate that there are such regularities, e.g. the weight for arguments of root verbs acting as antecedents of arguments in subclauses in the same sentence is relatively high (sentence distance=0, root→neb: weight=2.37), whereas arguments of verbs in object clauses in the previous sentence are unlikely to be antecedents of pronoun arguments of verbs in subclauses in the current sentence (sentence distance=1, objc→neb: weight=0.18) etc.

5.3.3.4 Semantic and morphosyntactic features of the antecedent

German pronouns can be used to refer to inanimate entities, as opposed to e.g. English, where pronouns with gender refer to animate entities (*he, she, his, her*). This introduces an additional layer of ambiguity, and the set of candidates that need be considered for German pronouns becomes larger.

In Tuggener and Klenner (2014), we introduced two features to account for this ambiguity. We added a feature conjunction of animacy and gender, and the feature named entity class to the set. Here, we slightly modify the animacy feature by adding number to the conjunction. As the conjunctions yields sparse instantiations, we show the weights for those instantiations for personal pronouns which occur at least 50 times in table 5.11.

Weight	**Animacy**	**Gender**	**Number**
1.95	ANIM	FEM	SG
1.50	ANIM	MASC	SG
1.19	ANIM	*	PL
0.97	INANIM	*	*
0.87	ANIM	*	SG
0.80	INANIM	*	PL
0.43	INANIM	FEM	SG
0.28	INANIM	MASC	SG

Table 5.11: Weights for the most frequent instantiations of the feature conjunction on animacy, gender, and number for antecedents of personal pronouns. '*' indicates underspecified values.

The weight distribution shows that candidates with singular number and specific gender which are animate ([ANIM, FEM, SG]: 1.95, [ANIM, MASC, SG]: 1.50) are more likely to be pronominalized than their inanimate counterparts ([INANIM, FEM, SG]: 0.43, [INANIM, MASC, SG]: 0.28). This weight distribution can be seen as a reflection of the topics of the underlying discourse in the training data. Newspaper articles often report on person entities, such as political figures, which therefore are likely to be pronominalized. We exploit this bias by featurizing it and thereby addressing the animacy ambiguity of German pronouns.

The second feature introduced in Tuggener and Klenner (2014) is the named entity class of the antecedent candidate. We add this feature without any modification, but only count and apply it when the candidate is actually a named entity. Table 5.12 shows named entity class weights for all pronouns.

	PPER	**PPOSAT**	**PRELS**	**PDS**	**PRELAT**
PERSON	1.84	2.93	0.90	1.24	1.09
ORGANIZATION	0.80	1.09	0.83	1.00	1.14
GEO-POL. ENTITY	0.45	0.43	0.58	0.30	0.69
LOCATION	0.15	0.17	0.56	0.65	0.07
OTHER	0.48	0.52	0.71	0.12	0.07

Table 5.12: Weights for named entity classes (rows) of antecedent candidates per pronoun type (columns).

The table indicates that personal and possessive pronouns tend to bind to person entities (PER), which corresponds with the weights of the animacy features. Except for organizations (ORG), the other named entity types (GPE=geopolitical entity, LOC=location, OTH=other) are weighted low, meaning their overall weight is decayed by the named entity class feature.

While working on the development set, we found that calculating prior weights for number and gender features also helps performance. Therefore, we add number and gender of the antecedent candidates as single features to the set.

Finally, we add the preposition of antecedent candidates in prepositional phrases as a feature. Candidates in PPs generally receive a low weight. However, we observed that different prepositions tend to affect salience to different amounts. Table 5.13 shows the five top and lowest weighted prepositions for personal and possessive pronouns.

Weight	**Preposition**
0.69	für
0.58	neben
0.46	gegen
0.42	von
0.33	bei
0.08	aus
0.04	nach
0.04	in
0.01	während
0.00	seit

Weight	**Preposition**
0.60	gegenüber
0.52	zwischen
0.47	neben
0.44	für
0.38	gegen
0.01	trotz
0.01	während
0.01	ohne
0.00	seit
0.00	vor

Table 5.13: Top and lowest five weights for prepositions of antecedent candidates for personal pronouns (left) and possessive pronouns (right).

The tables show that all preposition weights are below 1, which diminishes the candidates' overall weight products . However, there is a large difference between the top and lower weights. Amongst the lower weights, we see that "während" (while) and "seit" (since), two prepositions denoting PPs related to periods of time, are shared by the personal and possessive pronouns.

5.3.3.5 Features made available by the entity-mention model

The last two additions in the extended feature set are enabled by the incremental architecture of the entity-mention model. The **Discourse status** feature indicates whether a candidate is already in a coreference chain (discourse-old) or stems from the buffer list of non-anaphoric markables (discourse-new). The **Entity age** feature measures how "old" the discourse-old entities are that appear as antecedent candidates of pronouns. That is, the feature only triggers when the antecedent candidate is part of a coreference chain. The value for the **Entity age** feature is calculated by subtracting the sentence number of the first mention of the candidate's coreference chain from the sentence number of the first markable in the document. The value of the **Discourse status** feature is a binary one (discourse-old, discourse-new). The intuition behind these features is that entities introduced early in the discourse (e.g. in headlines) are likely to appear frequently throughout the discourse and are, therefore, likely to be pronominalized (Mitkov, 1998, Uryupina, 2007, inter alia). Furthermore, Strube and Hahn (1999) showed that in the Centering framework, determining salience of antecedent candidate entities based on information status (hearer new vs. hearer old) instead of grammatical functions improved pronoun resolution in their evaluation.

In Klenner and Tuggener (2010), we reported that other features derived from the entity-mention model, such as the length of the chain that a discourse-old candidate belongs to, had ambivalent impact on performance. During our experiments on the development set, we found that only the two features reported above increased performance overall.

5.4 Evaluation of pronoun resolution performance

In this section, we compare performance of the mention-pair and the entity-mention architecture and evaluate the impact of different feature sets discussed in section 5.3.3 for both models.[16] Thereafter, we validate our approach for weight calculation introduced in section 5.3.2 by comparing it to three machine learning approaches with different expressiveness. At the same time, these machine learning approaches serve as implementations of the different antecedent selection strategies discussed in chapter 2.

We first compare the models and apply several measures in order to determine the strengths of the classifiers and system performance overall. We start by comparing the classifiers in the mention-pair and the entity-mention models. Beforehand, we discuss our take on statistical significance testing of differences in evaluation scores.

5.4.1 A remark on statistical significance testing

In the coreference resolution literature, researchers often perform statistical significance tests if improvements over a baseline are small or seem marginal. A significance test then serves as an instrument to substantiate the validity of the presented work. However, there are several problems involved.

A fundamental issue is how samples are defined. The tests are usually applied to Recall and Precision figures. For example, to calculate the statistical significance of the

[16] We here evaluate full sets of features. For an evaluation of the impact of individual features, see Tuggener and Klenner (2014).

differences in Recall scores of two system responses, the document-wise Recall scores of the responses are compared in a contingency table. That is, the statistical significance of the difference regarding the overall Recall between two systems is assessed by comparing their document-wise Recall scores in a pair-wise manner. This setup is problematic, since the overall Recall is not simply the average of the document-wise Recall figures. With regard to pronouns for example, a short document might only contain one pronoun and one system resolves it correctly. This yields a 100% Recall. Another document might contain 100 pronouns and the system correctly resolves half of them, i.e. Recall is 50%. Averaging the document-wise Recall figures would yield an overall Recall of 75%, although the system has only resolved 51 of the 101 pronouns in the test set. The second system under scrutiny would incorrectly resolve the pronoun in the short document, but also 50% of the pronouns in the longer document. This would yield Recall figures of 0% for the first and 50% for the second document. In a document-wise comparison, the systems would then be quite different, i.e. 100% vs. 0% Recall for the first document and 50% vs. 50% for the second document.[17] Therefore, there is no direct correspondence between document-wise and overall Recall figures. The common way of establishing statistical significance of differences in system responses is therefore problematic and not applicable for our purposes.

An alternative is to establish the system differences on the level of the mentions instead of the documents. This is exactly what the ARCS system difference metric does. The metric takes as input a key and two system responses. It then iterates over all mentions, i.e. the gold mentions and the mentions in the system responses and compares the classifications in the system responses. The metric counts how often a mention that is classified as e.g. a wrong linkage (WL) in one response is counted as a true positive (TP) in the other response etc.[18] The percentage of mentions classified differently in one response w.r.t. the other then serves as the difference estimate.

The question now is how we should use this approach to calculate statistical significance of measured improvements, i.e. the better performance of a system compared to a baseline. When we compare system responses w.r.t. resolution performance, we neglect the fine-grained mention class distinction of the ARCS framework. That is, we are not interested in what kind of error a system makes regarding a mention (i.e. WL, FN, FP) but rather whether it makes an error or not. Thus, we map the fine-grained mention classification to a binary one, which is 1 if the mention is classified as correctly processed (TP, TN) and 0 otherwise (WL, FN, FP). We can then construct a $n \times 2$ contingency

[17] Naturally, the overall Recall is calculated on the global statistics in evaluation, i.e. in this example $\frac{51}{101} = 50.5\%$ and $\frac{50}{101} = 49.5\%$, respectively.

[18] Cf. 3.2.1

table where the first row lists the binary classification of the n mentions in the first system response and the second row their binary classification in the second response:

mention ID	**class in sys1**	**class in sys2**	**transition**
mention 1	1	1	tp→tp
mention 2	0	1	fn→tp
mention 3	1	0	tp→wl
mention 4	0	0	fp→fp
mention 5	0	1	fp→tn
...	...	...	...

Now, we can straight-forwardly apply the t-test for paired samples over the class rows. As we measure performance on the mention level, we can test significance on any measured performance change, i.e. on a specific PoS level or for a certain lemma.

However, it is noteworthy that the test cannot make statements specific to Recall and Precision changes because the error types are obscured. The test rather assesses the statistical significance of the changes in the binary classification scheme. We will indicate whether two responses vary significantly w.r.t. the test in the following section.

5.4.2 Classifier performance

In this section, we compare the performance of the classifiers in the entity-mention and the mention-pair models. The comparison is based on classifiers relying on weights calculated with our approach described in section 5.3.2. The classifiers score each candidate of a given pronoun, and we count how often the correct candidate is scored highest (i.e. is chosen as the antecedent) using the ARCS success rate[19], i.e. $\frac{|correctly\ resolved\ pronouns|}{|resolvable\ pronouns|}$. That is, we only evaluate on pronoun instances where the correct antecedent is among the candidates, since punishing the classifiers for making faulty decisions when they are not able to make the correct one introduces noise into the analysis. We saw in table 5.4 that the count of these instances where the correct antecedent is reachable is roughly the same for the mention-pair and the entity-mention models.

To put the performance of the classifiers into perspective, we establish the following four baselines commonly used in related work:

1. **Random candidate**: This is, of course, a rather crude baseline, but it transforms the analysis of the antecedent counts into performance scores. For example, given that the average number of candidates for personal pronouns is roughly 4, we can expect a resolution performance of 25% accuracy from the random baseline.

[19] Cf. section 3.4.3

2. **Most recent candidate**: This baseline selects the closest compatible candidate to the left of the pronoun as antecedent. Proximity has always been deemed an important factor in pronoun resolution which makes this a reasonable baseline often used in related work.

3. **Most recent subject candidate**: Hinrichs et al. (2005) and Wunsch (2006) found that the grammatical role of the antecedent candidates is one of the major features for identifying salient entities which are likely antecedents for pronouns, especially for German. That is, Hinrichs et al. (2005) and Wunsch (2006) reported that giving more weight to the subject emphasis in their G-RAP system compared to the English version improved their results. Furthermore, Wunsch (2010) used a postfilter which selected the most recent subject candidate as antecedent if the kNN classifier did not produce one. Doing so, Wunsch drastically improved Recall of his hybrid approach. Therefore, we consider selecting the most recent subject candidate as antecedent a strong baseline. The baseline checks whether there are subject candidates and returns the closest one to the pronoun. If no subject-bearing candidates are present, the most recent candidate is selected.

4. **Related work baseline**: For this baseline, we culminate features found in related work.[20] That is, we revert the incremental entity-mention system to a mention-pair system by i) removing the disambiguation step applied after resolution, and ii) postponing the creation of the coreference partition until the end of a document is reached and then transitively merge all found pairs. Doing so, we implement the algorithm 1 presented in section 5.2, which enables a direct comparison of both models in the same setting. We use a set of features generally found in other approaches to German coreference and pronoun resolution, as presented in table 5.5.

5. **Related work extended features baseline**: This baseline uses the extended feature set we implement in the incremental entity-mention model in a mention-pair architecture. Comparison to this baseline quantifies the effect of exchanging the mention-pair model with the incremental entity-mention model. Using approximately the same feature set restrains the feature set as a source for performance differences. Also, this comparison will show if and how much our extended feature set for German pronoun resolution improves the common mention-pair model.

The performance results for the various baselines and the mention-pair and entity-mention architecture on the development and test set are given in table 5.14. We discuss the results from top to bottom of both tables simultaneously.

[20] Cf. section 5.3.3

DEVELOPMENT SET

	PPER	**PPOSAT**	**PRELS**	**PDS**	**PRELAT**	**ALL**
random candidate baseline						
M-P	29.46	18.02	77.58	51.09	79.31	38.29
E-M	43.62	24.04	78.95	56.93	77.19	47.00
most recent candidate baseline						
M-P	68.43	62.69	92.66	83.21	91.38	72.95
E-M	69.94	62.79	92.64	**84.44**	91.38	73.72
most recent subject candidate baseline						
M-P	65.63	70.97	84.57	59.12	81.03	71.44
E-M	82.13	80.18	85.02	63.24	81.03	81.84
standard feature set						
M-P	83.40	85.40	92.73	81.02	**96.55**	86.16
E-M	87.98	85.18	92.65	83.70	**96.55**	88.28
extended feature set						
M-P	84.93	85.81	92.80	81.02	91.38	86.95
E-M	**91.59**	**90.32**	**92.94**	79.41	91.38	**91.28**

TEST SET

	PPER	**PPOSAT**	**PRELS**	**PDS**	**PRELAT**	**ALL**
random candidate baseline						
M-P	28.33	16.90	77.67	53.90	88.00	37.84
E-M	39.18	22.83	79.49	54.35	84.00	44.34
most recent candidate baseline						
M-P	67.65	62.42	92.90	74.47	92.00	72.16
E-M	68.28	62.94	93.06	**76.09**	92.00	72.74
most recent subject candidate baseline						
M-P	66.39	71.60	81.42	60.28	88.00	71.28
E-M	81.29	80.33	82.06	65.47	88.00	80.85
standard feature set						
M-P	83.69	85.33	93.30	73.05	**94.00**	86.13
E-M	88.53	85.47	**93.54**	**76.09**	**94.00**	88.52
extended feature set						
M-P	85.05	87.26	92.90	73.05	92.00	87.20
E-M	**90.42**	**89.14**	93.14	75.35	92.00	**90.31**

Table 5.14: Success rate (percentages) of different antecedent selection methods for pronouns given the mention-pair (**M-P**) and entity-mention (**E-M**) architectures.

The first important observation we make is that the performance of the random candidate baseline is higher in the entity-mention than in the mention-pair model (ALL; dev set: 47.00 vs. 38.23, test set: 44.34 vs. 37.84). The chances of picking an incorrect candidate as the antecedent are lower in the entity-mention model, since there are fewer candidates to choose from on average.[21] We also see that the random baseline performs surprisingly well for relative pronouns (PRELS and PRELAT). This is again explainable by the low candidate count for these pronoun types.

[21] Cf. section 5.3.1

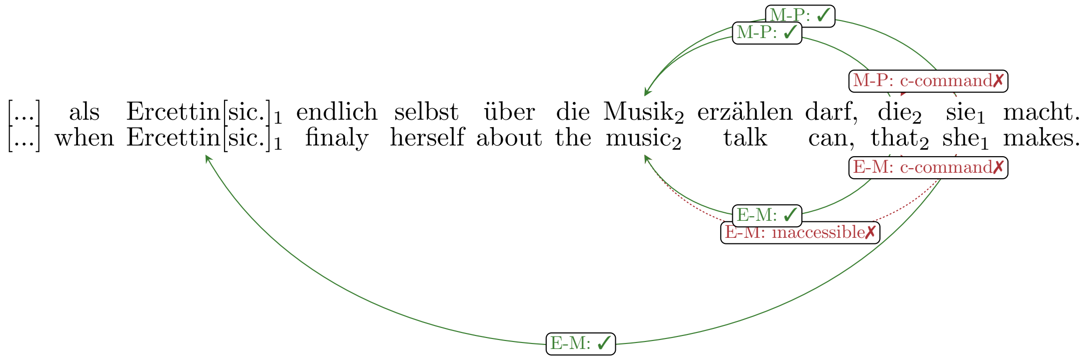
M-P: ✓
M-P: ✓
M-P: c-command✗
[...] als Ercettin[sic.]1 endlich selbst über die Musik2 erzählen darf, die2 sie1 macht.
[...] when Ercettin[sic.]1 finaly herself about the music2 talk can, that2 she1 makes.
E-M: c-command✗
E-M: ✓
E-M: inaccessible✗
E-M: ✓

Figure 5.5: Differences in pronoun resolution between the mention-pair (M-P) and entity-mention (E-M) model given the most recent candidate baseline.

Regarding the most recent candidate baseline, we note that the entity-mention model performs slightly better. Consider the segment from the test set shown in figure 5.5, where the pair-wise links established by the mention-pair model are shown above the text, and the links by the entity-mention model underneath the text. Note that neither the mention-pair nor the entity-mention model pair the personal pronoun “sie$_1$” with the relative pronoun “die$_2$” because the pronouns are exclusive due to binding constraints (i.e. c-command). However, the mention-pair model resolves both the relative pronoun “die$_2$” and “sie$_1$” to “Musik$_2$”, because “Musik$_2$” is the most recent compatible antecedent candidate for both pronouns. After the transitive closure of the pairs into coreference chains, both pronouns end up in the same coreference chain.

The entity-mention model correctly resolves “die$_2$” to “Musik$_2$” and “sie$_1$” to “Ercettin$_1$”, because it first links “die$_2$” to “Musik$_2$”. Thereafter, “Musik$_2$” is no longer accessible for the “sie$_1$” pronoun (indicated by the dashed red arc), because the exclusiveness between “die$_2$” and “sie$_1$” imposed by the c-command is transitively shared between “die$_2$” and “Musik$_2$”. Therefore, the closest compatible candidate for “sie$_1$” is the correct one, i.e. “Ercettin$_1$”. Effects like these caused by the different pair generation mechanics in the models let the entity-mention model outperform the mention-pair variant to a small degree for the most recent candidate baseline.

The most recent subject baseline extends the most recent baseline by checking whether there are candidates bearing the grammatical role subject. If so, the most recent one of them is selected as antecedent, else the most recent candidate is chosen. Table 5.14 shows that in the mention-pair model, only the performance of possessive pronouns increases (PPOSAT; dev set: 62.29 vs. 70.97, test set: 62.42 vs. 71.60). The scores of the relative (PRELS, PRELAT) and demonstrative pronouns (PDS) decrease substantially, indicating that recency is a more important factor than subjecthood for these pronouns.

Furthermore, the performance gap between the mention-pair and entity-mention model for the most recent subject baseline is large. Performance of the entity-mention model drastically increases compared to the most recent baseline and the model surpasses the mention-pair variant by large margins for the personal (PPER) and possessive pronouns (PPOSAT). The main reason for the performance gap is that the entity-mention model less frequently selects an incorrect subject candidate and falls back more often to selecting the most recent candidate. Consider the example in figure 5.6.

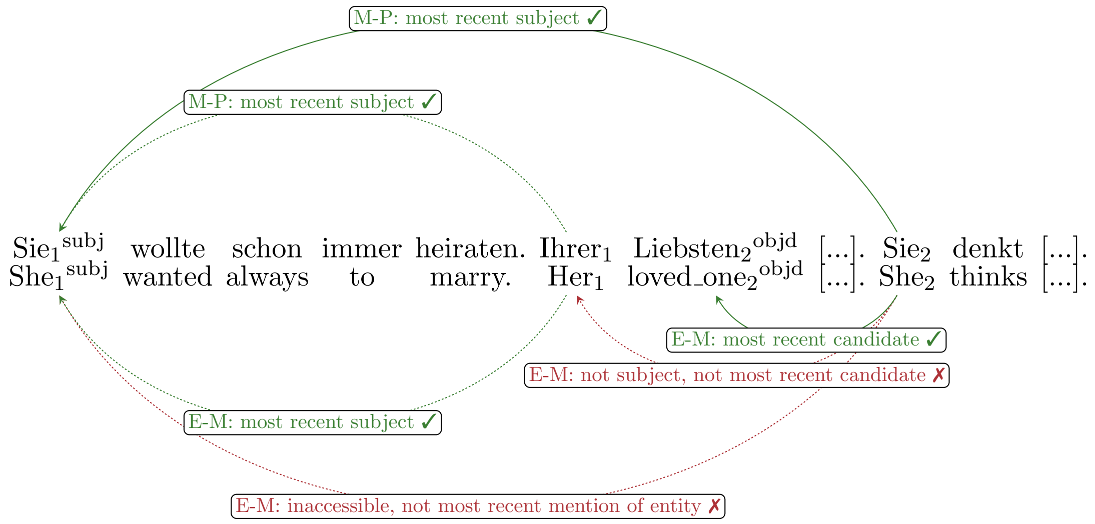
M-P: most recent subject ✓
M-P: most recent subject ✓
Sie_1^{subj} wollte schon immer heiraten. Ihrer_1 Liebsten_2^{objd} [...]. Sie_2 denkt [...].
She_1^{subj} wanted always to marry. Her_1 $\text{loved\_one}_2^{objd}$ [...]. She_2 thinks [...].
E-M: most recent candidate ✓
E-M: not subject, not most recent candidate ✗
E-M: most recent subject ✓
E-M: inaccessible, not most recent mention of entity ✗

Figure 5.6: Differences in pronoun resolution between the mention-pair and entity-mention model given the most recent subject-bearing candidate baseline.

Both models first correctly select “Sie_1” as the antecedent for “Ihrer_1” since “Sie_1” is the closest subject candidate. For “Sie_2”, the mention-pair model incorrectly selects “Sie_1” again as antecedent, because it is still the closest subject candidate. In entity-mention model, however, “Sie_1” is no longer accessible when resolving “Sie_2”, since the entity denoted by “Sie_1” is represented by its last mention, in this case “Ihrer_1”, which is not a subject-bearing candidate. Since there are no other subject-bearing candidates in the context, the entity-mention model falls back to selecting the most recent compatible candidate, “Liebsten_2”, which in this case is the correct one.

To quantify this effect, we counted how often the selected antecedent is subject-bearing in the most recent subject-bearing candidate baseline for the models. For personal pronouns, the mention-pair model selects a subject candidate in 91.00% of the cases and for possessive pronouns in 96.83% of the cases, while the entity-mention model selects a subject candidate for only 83.30% of the personal pronouns and for 93.27% of the possessive pronouns. That is, the entity-mention model less frequently has access to incorrect subject-bearing candidates because they are often hidden behind more recent mentions of the denoted entity.

Finally, the results of the feature set-based classification show that i) the entity-mention model outperforms the mention-pair variant for both the standard and the extended feature set, and ii) the extended feature set improves performance in both models.[22] Furthermore, both models strongly improve over all baselines for personal (PPER) and possessive (PPOSAT) pronouns. For relative pronouns (PRELS), performance in both models is only affected to a negligible degree by using a feature-based resolution approach. The most recent candidate baseline performs on par. For demonstratives (PDS) and attributing relative pronouns (PRELAT), the results are less conclusive. Notably, the sample sizes are considerably smaller than for the personal, possessive, and relative pronouns. Both demonstrative and attributing relative pronouns seem to perform well under the most recent candidate baseline, and we see no improvement, or even performance degradation, given the feature set-based resolution. Attributing relative pronouns (PRELAT) perform better using the standard feature set than when using the extended feature set. Performance of the extended feature set is equivalent to the most recent candidate baseline.

Overall, the performance of the classifiers in the entity-mention model seems satisfactory, with an overall ARCS success rate of 91.28 on the development set and 90.31 on the test set. Recall that we measure performance on pronoun instances that are resolvable, i.e. where the correct antecedent is among the candidate. That is, if the classifiers are able to make the correct choice, they do so in 91.28% and 90.31% of the cases in the development and test set, respectively. To assess performance of our approach as a full pronoun resolution system, and not only of its classifiers, we will next evaluate its output w.r.t to all mentions in the gold standard and the system output.

5.4.3 System response evaluation

The evaluation of classifier performance in the previous section was conducted for pronouns where the correct antecedent was among the candidates. To measure performance on all pronoun in the test set, we apply our ARCS metric (Tuggener, 2014).[23] ARCS evaluates coreference system outputs, i.e. it compares system response files to a gold key file, like the official reference scorer for coreference resolution evaluation (Pradhan et al., 2014) for the commonly used metrics.

[22] Note that we cannot model the two features discourse status and entity age in the mention-pair model, since the model does not provide access to them (cf. section 5.3.3). Besides these two features, both the entity-mention and mention-pair model share the same features in the extended features set.

[23] Cf. section 3.2

5.4.3.1 Functional evaluation

As in Tuggener and Klenner (2014), we choose the ARCS inferred antecedents metric for evaluating pronoun resolution performance. This metric is related to the functional evaluation proposed in Müller (2008) and evaluation of non-pronominal anchors in Stuckardt (2001) in requiring pronouns to (transitively) link to a correct nominal antecedent.[24] More precisely, the closest nominal antecedent to the left of the pronoun must be a member of the gold coreference chain of the pronoun. This requirement improves the estimation of performance from the perspective of downstream applications, since systems are not rewarded for linking pronouns to other pronouns, because this does not help in inferring the underlying entity.

Our analysis of classifier performance in table 5.14 showed that relative and demonstrative pronouns do not benefit from a feature based resolution strategy, i.e. their performance leveled when applying the method of selecting the most recent candidate as the antecedent. Therefore, we limit our detailed investigation to personal and possessive pronouns. Performance of all pronouns, including relative and demonstrative pronouns, is thus subsumed in the ALL scores.

To relate the weight calculation introduced in section 5.3.2 to other machine learning-based methods, we include three additional approaches. These approaches also reflect different antecedent selection strategies outlined in chapter 2.

- **Maximum Entropy (MaxEnt) / Best-first heuristic**: We apply a maximum entropy approach to learn weights and score antecedent candidates. To do so, we create a feature vector for each candidate for a pronoun and add a feature indicating whether the vector denotes a correct antecedent or not. Since we cannot directly express constraints like “only trigger the markable distance feature if antecedent candidate and pronoun are in the same sentence” in this representation, we create feature conjunctions to e.g. condition the markable distance weight to sentence distance. That is, we create a feature conjunction of markable distance and sentence distance to make markable distance dependent on sentence distance.

  Table 5.15 shows such a vector, where feature values with slashes indicate feature conjunctions. The MaxEnt classifier learns from these individual feature instances. During testing, the classifier assigns each of the candidate vectors a weight, and the highest ranked one is selected as antecedent. This corresponds to the *best-first* heuristic for selecting antecedents. Our weighting scheme also uses this heuristic. The MaxEnt approach thus serves as a direct competitor for our feature weighting approach.

[24] Cf. section 3.4.2

markable ID	**conn./sent. dist.**	**sent./mark. dist.**	**cand. index**	**GF ante**	**GF seq.**	**discourse status**	**anim./num./gen. ante**	**PoS/NE ante**	**Gender ante**	**Number ante**	**Intro. sent. ante**	**clause type ante**	**clause type seq.**	**class**
MaxEnt instance														
42	-/1	1/5	1	subj	1/SUBJ/OBJD/NE	old	ANIM/M/S	NE/PER	M	S	0	aux	1/aux/root	ante
CRF instance														
46	-/3	3/10	2	pn	3/PN/SUBJ/NN	new	*/M/S	NN/-	M	S	140	aux	1/aux/cj	-
54	-/1	1/2	1	pn	1/PN/SUBJ/NN	new	*/M/S	NN/-	M	S	142	s	1/s/cj	-
55	-/1	1/1	0	subj	1/SUBJ/SUBJ/PPER	old	ANIM/M/S	PPER/PER	M	S	0	root	1/root/cj	ante

Table 5.15: Example of feature vector instance used by the MaxEnt classifier and instance of vector sequence used for the CRF.

- **Conditional Random Fields (CRF)**: The linear-chain CRF[25] we apply uses the same vector representation as the MaxEnt classifier. However, training and testing instances are no longer isolated vectors, but sequences of vectors. An instance consists of all candidate vectors for a pronoun in the linear order of their appearance in the document (i.e. the candidate the furthest away from the pronoun denotes the first vector etc.), with only one of them signifying the correct antecedent. Table 5.15 shows such a sequence.

  CRFs are usually applied to sequence labeling problems, like PoS tagging. What kind of sequence properties can we hope to learn from our instance representation? The sequence we learn is the class labels of the candidates (i.e. the last dimension of the vectors). There is always exactly one correct antecedent among the candidates. Compared to the MaxEnt classifier, we can now learn that the 'ante' label always follows either a 'non-ante' label or the sequence start. Also, the 'ante' label can only be followed by either a 'non-ante' label or the sequence end. This can be seen as a weak approximation of enforcing a global constraint, i.e. that only one of the candidates can be the antecedent.

  Perhaps more importantly, we can access features of the current, previous, and next candidate vector while traversing the sequence. Hence, we can learn feature weights for assigning the 'ante' label to the current candidate vector given features from the current, the previous, and the next candidate vector, and combinations thereof. For example, for the grammatical role feature, we learn weights that denote:

  - What is the weight for labeling the current candidate as 'ante' given that its grammatical role is e.g. *subject*? (This is what we learn in the MaxEnt model.)

[25] For both the MaxEnt and the CRF classifier, we use *wapiti*: `https://wapiti.limsi.fr/`

  - What is the weight for labeling the current candidate as ‘ante’ given that its grammatical role is e.g. *subject* and that of the previous candidate is *direct object*?
  - What is the weight for labeling the current candidate as ‘ante’ given that the grammatical role of the previous candidate is *subject*?
  - What is the weight for labeling the current candidate as ‘ante’ given that the grammatical role of the next candidate is *direct object*?

  Since now features and vector instances directly compete in a pair-wise fashion, the CRF approach can be seen as a variant of the twin candidate model Yang et al. (2008b).[26] The twin candidate model has been shown to outperform a single candidate mention-pair model (Yang et al., 2008b). Also, it can be considered an intermediate stage when moving from a mention-pair model to a mention ranking model. That is, antecedent candidates compete in a pair-wise fashion, unlike in the mention-pair model where each candidate is considered in isolation. Still, not all candidates are considered at once, like in the mention ranking model.

- **Markov Logic Networks (MLN)**: Our approach of using MLNs for pronoun resolution is detailed in Tuggener and Klenner (2014). The main idea is to convert the features outlined in section 5.3.3 to first-order logic predicates and combine them in formulas. For example, the feature for sentence distance given the presence of a discourse connector yields the following formula:

$$\begin{aligned} &w(sd, conn, pos): \\ &\qquad in\_sent(a, s1) \wedge in\_sent(p, s2) \wedge sd = s2 - s1 \\ &\qquad \wedge\, has\_pos(p, pos) \wedge has\_connector(p, conn) \\ &\qquad\qquad \Rightarrow anaphoric(a, p) \end{aligned}$$

  where a denotes an antecedent candidate and p a pronoun. The formula is assigned weights for each instantiation combination of $sd, conn, pos$ during training.

  A training instance in the MLN approach consists of a pronoun and all its antecedent candidates, including the correct one. Weights for the formulas are inferred globally over all candidates in such an instance and over all instances. During testing, all candidates for a given pronoun are considered competitively,

[26] Cf. section 2.2.1. The twin candidate model actually pairs candidates in a round-robin tournament, and the candidate with the most wins is selected as antecedent. We do not implement such a tournament mode, but adapt from the twin candidate model the notion that weights for features should be learned in direct competition of the features. During testing, the highest scored candidate is selected as antecedent.

> and the highest ranking one is selected as antecedent. Thus, the MLN approach directly implements a mention ranking model.[27]

The aim of the following evaluation is three-fold, i.e. i) juxtaposition the mention-pair and the entity-mention model, ii) compare different weight inference schemes which reflect concepts inherent in different coreference models, iii) assess system performance w.r.t upper bounds. Also, we include results for using the standard feature set. Note that all the competing machine learning frameworks use our extended feature set. Furthermore, in each machine learning framework, we learn a separate classifier for each pronoun type.[28]

We saw in section 5.3.1, table 5.4 that the correct antecedent is not accessible for a portion of pronoun instances, which affects the upper bound of the system. To identify the upper bounds, we use the gold annotation to identify the correct antecedent among the candidates, if present. If not, the classifiers are used to determine a (necessarily incorrect) antecedent. The absence of the correct antecedent mainly stems from two issues. First, preprocessing might not have extracted the correct markable that represents the correct antecedent. Second, the pronoun at hand is not anaphoric, i.e. there is no correct antecedent in the gold standard.

Table 5.16 shows the upper bounds and actual system performance as measured by the ARCS inferred antecedent metric on the development and test set, respectively. We see that the entity-mention model outperforms the mention-pair model in all regards. That is, the better performance of the classifiers evaluated in section 5.4.2 carries over to the functional evaluation of the system responses. Over all pronouns (ALL), the entity-mention model outperforms the mention-pair model by 3.7 F1 points on the development set (73.68 vs. 69.98) and by 2.71 F1 points on the test set (75.20 vs. 72.49). The performance differences are larger when we focus on personal (PPER) and possessive pronouns (PPOSAT), respectively. On the development set, the best entity-mention response outperforms the best mention-pair response by 5.4 F1 points (E-M vs. M-P std.feat.set; 70.15 vs. 64.75) for personal pronouns and by 5.35 F1 points regarding possessive pronouns (E-M vs. M-P; 73.88 vs. 68.53). On the test set, the difference in performance are smaller, but still present. Also, the performance scores of all system responses are higher on the test set.[29]

[27] Cf. section 2.2.1

[28] For the MLN, we include the PoS tag of the pronoun in the formula to learn pronoun-specific formula weights (cf. Tuggener and Klenner (2014))

[29] In the lemma-based error analysis in section 5.5.1, we show that our resolution performance for feminine and plural pronouns is significantly lower than our performance on masculine pronouns. While our test and development sets contain roughly the same number of personal and possessive pronouns, it turns out that the test set contains fewer feminine and plural pronouns and more masculine pronouns. Thus, our systems achieve higher performance on the test set than on the development set.

DEVELOPMENT SET

	PPER			PPOSAT			ALL		
	R	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
E-M Upper bound	*86.72*	*79.82*	*83.13*	*90.63*	*82.01*	*86.11*	*88.15*	*81.23*	*84.55*
E-M MLN	72.89	66.79	69.71	**77.92**	**70.24**	**73.88**	**76.94**	**70.68**	**73.68**
E-M	**73.35**	**67.21**	**70.15**	76.86	69.35	72.91	76.66	70.44	73.42
E-M CRF	71.36	65.41	68.25	75.35	67.91	71.44	75.64	69.50	72.44
E-M MaxEnt	69.12	63.33	66.10	74.97	67.50	71.04	74.41	68.34	71.25
E-M std.feat.set	68.41	63.02	65.61	70.88	64.14	67.34	73.12	67.39	70.14
M-P Upper bound	*87.96*	*80.60*	*84.12*	*91.32*	*82.31*	*86.58*	*88.88*	*81.60*	*85.08*
M-P	66.13	60.55	63.21	72.33	65.12	68.53	72.55	66.56	69.43
M-P std.feat.set	67.70	62.04	64.75	71.70	64.63	67.98	73.11	67.11	69.98

TEST SET

	PPER			PPOSAT			ALL		
	R	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
E-M Upper bound	*89.38*	*83.59*	*86.39*	*92.63*	*82.54*	*87.30*	*89.47*	*82.26*	*85.71*
E-M MLN	**77.10**	**72.05**	**74.49**	**79.02**	**70.33**	**74.42**	**78.53**	**72.13**	**75.20**
E-M	75.93	70.97	73.39	78.69	69.99	74.09	77.89	71.53	74.57
E-M CRF	75.57	70.65	73.03	77.76	69.21	73.23	77.47	71.19	74.20
E-M MaxEnt	73.05	68.29	70.59	77.95	69.47	73.47	76.42	70.23	73.19
E-M std.feat.set	73.28	68.88	71.08	71.98	64.45	68.01	75.02	69.20	71.99
M-P Upper bound	*90.15*	*84.17*	*87.05*	*92.83*	*82.67*	*87.46*	*89.87*	*82.54*	*86.05*
M-P	71.21	66.46	68.75	77.56	69.07	73.07	75.57	69.38	72.34
M-P std.feat.set	73.41	69.04	71.16	73.64	65.82	69.51	75.57	69.65	72.49

Table 5.16: Functional evaluation of third-person pronoun resolution with the ARCS inferred antecedent metric on the development set (top) and the test set (bottom).

	E-M MLN	E-M	E-M CRF	E-M MaxEnt	E-M std.feat.set	M-P	M-P std.feat.set
E-M MLN		-	-	+	+	+	+
E-M	+		+	+	+	+	+
E-M CRF	+	-		+	+	+	+
E-M MaxEnt	+	+	-		-	+	-
E-M std.feat.set	+	+	+	+		+	+
M-P	+	+	+	+	-		+
M-P std.feat.set	+	+	+	+	-	-	

	E-M MLN	E-M	E-M CRF	E-M MaxEnt	E-M std.feat.set	M-P	M-P std.feat.set
E-M MLN		-	+	+	+	+	+
E-M	-		-	+	+	+	+
E-M CRF	-	-		+	+	+	-
E-M MaxEnt	+	-	-		-	+	-
E-M std.feat.set	+	+	+	+		+	-
M-P	+	-	-	-	+		+
M-P std.feat.set	+	+	+	+	+	+	

Table 5.17: Significance test on the performance measure differences w.r.t the functional evaluation for the development set (left) and test set (right). + indicates significant differences ($p < 0.05$), - signifies insignificant differences ($p > 0.05$)

Looking at the upper bounds, we see that both approaches have similar ones, with the mention-pair having a slightly higher upper bound. This corresponds to the analysis of the availability of the correct antecedent in section 5.3.1, where we found that the availability of the correct antecedent is slightly lower in the entity-mention model. The slightly lower availability here translates into a marginally lower upper bound. That is, for approximately 88-89% of the pronouns in the gold standards the models have the possibility to identify a correct nominal antecedent (Recall of upper bounds). And for all pronouns that the systems resolve, around 81-83% have an identifiable nominal antecedent (Precision of upper bound). Given these upper bounds, we see that the entity-mention model lacks roughly 10 F1 points behind (e.g. EM MLN 75.20 vs. E-M Upper bound 85.71 on the development set). This again corresponds to the classifier performance of around 90% ARCS success rate evaluated in section 5.4.2.

Overall, we see that Recall is higher than Precision. This indicates that our approach resolves too many pronouns, i.e. pronouns not annotated as anaphoric in the gold standard. Error analysis in section 5.5 will elaborate on this point.

Furthermore, we observe that our weight calculation approach outperforms the CRF and MaxEnt competitors and only marginally under-performs compared to the best performing classifier, the MLN. However, the significance tests suggest that the smaller differences among the top scoring responses are not significant.

Table 5.17 shows the result of the statistical significance test introduced in section 5.4.1. The tables contain both the results for personal and possessive pronouns, separated by a diagonal line of empty cells. The results for personal pronouns are listed above the diagonal line of empty cells and results for possessive pronouns are listed below the diagonal. The responses are listed along the X and Y axis according to the evaluation scores, i.e. starting with the best performing system responses in the top left corner and ending with the responses performing lowest on the bottom right corner.

To check e.g. whether improvements of our own weighting scheme are significant on the development set (left table in table 5.17), we check the according row (E-M). Our weighting scheme outperforms five other responses. Reading the E-M row from left to right, we see that all improvements are significant. To check significance of the improvements on the development set regarding possessive pronouns, we read the column of our weighting scheme (again E-M) from top to bottom. Here, we see that the improvement of our weighting scheme over the CRF approach (E-M vs. E-M CRF) is not statistically significant, but that all other improvements of E-M are.

We see that for both personal and possessive pronouns, the differences among the top two performing responses are not statistically significant (E-M MLN vs. E-M), except for possessive pronouns on the development set. This suggests that the performances of our own weighting scheme and the MLN-based approach are very similar.

Furthermore, for personal pronouns, we observe that all improvements over the mention-pair model with the extended feature set (M-P) are statistically significant (second last column in both tables). However, on the test set (right table), compared to the mention-pair model using the standard feature set (M-P std.feat.set, last column), only the top two performing responses are significantly better. This has to do with the observation that the extended feature set does not improve performance of the mention-pair model for personal pronouns. That is, on the test set, the MaxEnt variant and our weighting approach using the standard feature set perform worse than the mention-pair model employing the standard feature set. However, these lower performances are not statistically significant.

Concerning possessive pronouns, all improvements over the mention-pair model using the standard feature set are statistically significant on the test set (last row in right table). Also, all improvements over the entity-mention model with the standard feature set are significant (third last row), which is not surprising, since the response performs worse than the mention-pair responses w.r.t. possessive pronouns.

Overall we see that the magnitude of the differences in the performance evaluation correlates with the statistical significance of the results. That is, larger differences in the evaluation are significant, while smaller ones are not. While not all smaller individual differences are statistically significant, the overall improvements of the entity-mention model over the mention-pair model are (bottom left and top right of the tables).

5.4.3.2 Anchor mention evaluation

Our ARCS evaluation framework features another metric that is relevant for assessing coreference resolution performance from the perspective of downstream applications, namely the ARCS anchor mention evaluation. This metric is an extension of the ARCS inferred antecedent metric. Instead of requiring mentions to link to correct nominal antecedents, this metric requires mentions to link to a so-called anchor mention of the entity. This anchor mention is deemed to fully describe the entity that is denoted by the coreference chain. As an approximation, we select the first nominal mention in a coreference chain as the anchor mention. First nominal mentions of entities tend to introduce the entities into discourse and are likely to contain strings that higher-level applications query. An ideal anchor mention would thus be e.g. “Barack Obama, the president of the United States”.

This metric is especially relevant for downstream applications that perform queries targeted at finding contexts in which a specific entity is mentioned, i.e. Sentiment Analysis. Thus, linking a pronoun to a nominal antecedent, like “the president” would not be useful to a Sentiment Analysis query targeted at “Barack Obama”, because it is not clear which president is denoted by the mention “the president”.

The ARCS anchor mention metric first links key coreference chains to system response chains by aligning the anchor mentions. Once aligned, the metric measures how many of the key mentions are contained in the response chain to determine Recall, and how many response mentions are in the key chain to determine Precision.[30]

Since we measure performance on the mention level, we are able to measure performance specifically for different named entity types. We apply the ARCS anchor mention metric to the entity-mention and mention-pair model and assess performance for coreference chains that denote person entities, i.e. whose anchor mention is a named entity of class *PER*. The other named entity classes, like *ORG* and *GPE*, are not pronominalized often enough to provide a solid basis for evaluation.

		DEV SET			**TEST SET**		
		R	**P**	**F1**	**R**	**P**	**F1**
E-M	**ED**	77.38	86.77	81.81	76.93	88.01	82.10
	PPER	46.00	81.00	58.68	57.00	82.00	67.25
	PPOSAT	59.00	84.00	69.31	64.00	79.00	70.71
M-P	**ED**	77.50	86.33	81.67	76.83	88.32	82.17
	PPER	38.00	78.00	51.10	50.00	80.00	61.54
	PPOSAT	53.00	82.00	64.39	60.00	78.00	67.83

Table 5.18: Evaluation of third person pronoun performance for linking to anchor mentions of person entities.

Table 5.18 provides the performance details. Both models achieve almost identical performance w.r.t entity detection (ED), i.e. aligning anchor mentions in the coreference chains in the key to coreference chains in the response. We see that the entity-mention model outperforms the mention-pair competitor by a large margin regarding personal pronouns (PPER) and surpasses it w.r.t. possessive pronouns (PPOSAT) performance.

As argued before, we believe this analysis is of importance since downstream applications interested in specific target entities rely on the identification of these anchor mentions and require a coreference system to link other mentions to these anchors. We can read the results in the following way in this view. Both models identify almost 77% of the key anchor mentions for person entities (ED Recall). Of the anchor mentions that the systems produce, 88% are relevant (ED Precision). Regarding the found and aligned anchors, the entity-mention model finds 57% of the pronoun mentions (PPER Recall), the mention-pair model 50%. 82% of the identified pronoun mentions are correct in the entity-mention model, 80% in the mention-pair model (PPER Precision). The entity-mention model identifies 64% of the possessive pronoun mentions of person entities correctly, the mention-pair model 60%, and 79% and 78% of the identified possessive pronoun mentions are correct in the entity-mention and mention-pair model, respectively.

[30] For a concise and formal presentation of the respective algorithms cf. Tuggener (2014).

The overall range of the scores indicate that correctly resolving pronouns to such anchor antecedents, which we deem relevant for at least a subset of higher-level applications, is a difficult task. Performance is lower than for the functional evaluation given in table 5.16 and much lower than the pair-wise performance scores usually reported in related work.[31] That is, while we can achieve classifier accuracies of 90% and upwards concerning the identification of local antecedents of pronouns, as e.g. reported in table 5.14, this does not mean that a system will be useful to such a large extent for higher-level applications which might require global antecedents, such as anchor mentions.

5.4.3.3 Standard evaluation

Finally, we evaluate our entity-mention approach using the standard coreference evaluation scheme.[32] This evaluation also compares a key to a response file, but does not distinguish between mention types and includes nominal (common nouns and named entities)[33] and all other mentions contained in the key, i.e. pronoun types that our system does not resolve (e.g. reflexive pronouns). Table 5.19 gives the results. We show results for mention detection (**MD**) and the three metrics used to calculate the average F-score (the MELA metric, i.e. $\varnothing = \frac{MUC+B^3+CEAFE}{3}$).

	MD			**MUC**			**B^3**			**CEAFE**			$\varnothing$
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
E-M UPB	*79.07*	*77.27*	*78.16*	*70.32*	*73.45*	*71.85*	*61.91*	*69.45*	*65.46*	*71.34*	*60.31*	*65.37*	*67.56*
E-M MLN	77.95	76.26	77.10	67.26	70.26	68.73	59.28	66.42	62.65	68.64	58.21	62.99	64.79
E-M	77.83	76.20	77.01	67.11	70.09	68.57	59.08	66.26	62.46	68.42	58.15	62.87	64.63
M-P	77.98	75.90	76.92	66.64	69.68	68.13	58.30	66.06	61.94	68.73	57.30	62.50	64.19

Table 5.19: Performance in the common evaluation framework on the test set.

Since a large portion of the mentions in the gold standard are nominal mentions and all our approaches employ the same strategy to resolve nominal mentions, the differences between the different pronoun resolution approaches are marginal when compared based on the standard evaluation framework. Still, the results shows that the MLN approach performs best in all regards.

The difference between the MLN approach and the upper bound here is much smaller than in our functional evaluation of pronoun resolution. This is again due to the presence of the large number of the nominal mentions which are not affected by the pronoun resolution strategies. The upper bound (**E-M UPB**) indicates the system performance when gold standard information is used to resolve pronouns, given that the correct

[31] Cf. the next section 5.4.5

[32] Cf. Chapter 2

[33] For the details of our approach to resolve nominal mentions, cf. section A.2.

candidate is accessible. That is, it shows how well our system would fare if the classifiers used for pronoun resolution would always choose the correct antecedent if available.

While we cannot compare our results to those of English coreference resolution research, it is still noteworthy that current state-of-the-art approaches to English coreference resolution achieve scores very similar to ours. For example, Björkelund and Kuhn (2014) report a MELA F-score of 61.63, Martschat and Strube (2015) achieve 62.47 MELA F-score, and Fernandes et al. (2014) report an F-score of 63.37.

The only recent German coreference resolution paper that reports standard evaluation scores is Rösiger and Riester (2015).[34] Rösiger and Riester implemented the HOT-Coref system by Björkelund and Kuhn (2014) for German and investigated the utility of prosodic features for coreference resolution.[35] They used the same test set as we did, i.e. the first 690 documents of the TüBa-D/Z corpus to evaluate their baseline system which achieved a MELA F-score of 51.61. However, they did not apply a gold mention boundary match. If we remove the gold mention boundary alignment from our approach, we achieve a MELA F-score of 61.65.

5.4.4 Performance impact of real preprocessing components

As mentioned in section 1.5.2, preprocessing is an integral part of any coreference resolution system. It has been shown that performance of coreference resolution, including pronoun resolution, depends heavily on the quality of the preprocessing components such as PoS tagging, syntactic parsing, morphological analysis etc. (Schiehlen, 2004, Klenner et al., 2010, inter alia). So far, we have assumed perfect preprocessing information. While doing so eliminates noise when investigating coreference resolution performance, it presents an unrealistic setting for real-world applications of a coreference system. Therefore, we report the performance of the entity-mention and the mention-pair model w.r.t. pronoun resolution when real preprocessing components are used to extract the markables and their features. However, we keep our method of aligning gold mention boundaries to boundaries of the corresponding extracted markables, because we argue that the precise identification of the boundaries (i.e. including or excluding a PP or a relative clause) is irrelevant for higher-level applications.

[34] The Semeval 2010 Shared task on coreference resolution for multiple languages (Recasens et al., 2010) also featured a German data set compiled from an earlier version of the TüBa-D/Z corpus. However, the participating systems did not fare particularly well regarding German pronoun resolution, as reported in Tuggener and Klenner (2014). This is not surprising, since German pronoun resolution was not the focus of the task. Therefore, we refrain from re-running our system on the Semeval data and from comparing it to the participating systems.

[35] Cf. section 4.10

We apply the *ParZu* parser (Sennrich et al., 2013) which provides PoS tagging and morphological analysis, besides dependency parsing. The parser is an adaption of the *Pro3Gres* dependency parser for English (Schneider, 2008) to German. For named entity recognition, we use the Stanford Named Entity Recognizer[36] with the model for German provided by Faruqui and Padó (2010). We keep the tokenization given by the gold standard to avoid token alignment problems in evaluation. That is, all preprocessing is fully automated. The evaluation thus represents the performance of our system in a real-world setting. We compare the entity-mention and mention-pair models and use our antecedent selection strategy introduced in section 5.3.2. Furthermore, we found that training the weights on the training set preprocessed with the real components gives slightly better results than using the weights obtained over gold preprocessed training data.

Apart from using the TüBa-D/Z development and test set, we evaluate the systems on the Potsdam Commentary Corpus (PCC).[37] The corpus does not feature annotation of relative pronouns, but for personal and possessive pronouns.

	PPER			**PPOSAT**			**ALL**		
	R	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
TüBa-D/Z Development set									
E-M	64.80	62.87	**63.82**	68.68	62.19	**65.27**	65.52	64.00	**64.75**
M-P	59.19	57.66	58.42	64.78	58.82	61.66	62.24	60.73	61.48
TüBa-D/Z Test set									
E-M	64.37	62.87	**63.61**	68.13	60.64	**64.17**	65.15	63.45	**64.29**
M-P	59.33	58.03	58.67	67.60	60.20	63.68	63.13	61.44	62.27
Potsdam Commentary Corpus									
E-M	70.55	66.67	**68.55**	70.62	60.68	**65.27**	-	-	-
M-P	66.34	62.88	64.57	66.10	56.80	61.10	-	-	-

Table 5.20: Functional evaluation of pronoun resolution performance using real preprocessing components.

Table 5.20 shows the results of the functional evaluation that requires pronouns to (transitively) link to nominal antecedents (i.e. the ARCS inferred antecedent metric). On the TüBa-D/Z data, we see that performance is lowered by roughly 6 to 10 percentage points in F-score when using real preprocessing compared to using gold preprocessing. In table 5.16, we saw that the entity-mention model (**E-M**) achieved F-scores of 70.15 and 73.39 on the TüBa-D/Z development and test set, respectively, for personal pronouns (**PPER**). Given real preprocessing, F-scores drop to 63.82 and 63.61, respectively. The same magnitude of loss is observed w.r.t. possessive pronouns and the performance over all pronouns. For possessive pronouns, the entity-mention model reached 72.91 and 74.09

[36] `http://nlp.stanford.edu/software/CRF-NER.shtml`

[37] Cf. section 5.1.1. The corpus currently only provides coreference annotation in the CoNLL format. Thus, we were not able to evaluate our approaches based on gold preprocessing on this corpus.

F-scores on the development and test set, respectively. Here, performance is lowered to F-scores of 65.27 and 64.17.

The evaluation of our systems on the Potsdam Commentary Corpus (PCC) shows that the performance does not change significantly regarding possessive pronouns compared to the TüBa-D/Z data sets. For personal pronouns, performance is higher on the PCC. However, the PCC is significantly smaller than the TüBa-D/Z test set.[38] Still, the results suggest that our system does not overfit the TüBa-D/Z data and can be applied to other corpora.

Furthermore, the table shows that the entity-mention model outperforms the mention-pair model on all data sets. Thus, the improvements achieved on gold data carry over to an evaluation in a real-world setting.

This evaluation shows that pronoun resolution in a real-world setting, where pronouns have to meet the requirements of downstream applications (i.e. identify nominal antecedents) and where the systems have to rely on automated preprocessing, remains a challenging task. Pair-wise evaluation and the use of gold preprocessing do not adequately reflect these requirements.

5.4.5 Comparison of the evaluation results to related work

Related work on German pronoun resolution in Hinrichs et al. (2005), Wunsch (2010) performed evaluation by scoring instance vector representations of pairs of antecedent candidates and pronouns as presented to a classifier. The labels assigned by the classifier were compared to the gold labels obtained from the corpus. Each instance vector was then evaluated in the following way.

- TP: The classifier labeled the instance as positive, the gold label is positive. The antecedent and the pronoun denoted by the instance are in the same coreference chain.
- FP: The classifier labeled the instance as positive, but the gold label is negative. This subsumes both the case where the pronoun is not anaphoric, and the case where the selected antecedent and the pronoun are members of different coreference chains.
- FN: The classifier labeled the instance as negative, the gold label is positive.

[38] The TüBa-D/Z test set contains 2223 personal and 1506 possessive pronouns with an identifiable nominal antecedent. The PCC features 309 personal and 177 possessive pronouns with nominal antecedents.

Two main issues arise from this evaluation method, based on the fact that the output is not directly compared to a gold key, but to an intermediate representation of the gold standard, i.e. the pair instances. First, gold pronouns not represented in any instance vector are omitted, as vectors are only created for instances that have candidates. That is, Recall does not extend over all annotated pronouns in the gold standard, but only over the pronouns that the system extracts and creates vector instances for. Second, multiple positive instance vectors denoting the same pronoun get scored multiple times. It is possible that a pronoun is represented by multiple positive pair vectors of different correct antecedents. If all these instances are labeled correctly, the pronoun is scored multiple times as TP. Likewise, the classifier might label one of the instances that denote the same pronoun as positive and others as negative. This yields problems when merging the pairs to obtain the coreference chains, since the transitivity property of coreference is violated. That is, a classifier might label mention A and B as coreferent, and B and C as coreferent, but negatively classify the pair A and C. It is left unclear how such contradictions affect the formation of coreference chains in a pairwise evaluation, which is undesirable from the perspective of higher-level applications. Therefore, it is not guaranteed that the improvements made on pairwise decisions carry over to the goodness of the coreference chains formed in the pair merging step (Ng and Cardie, 2002a, Ng, 2010).

Given the differences in the evaluation methods, it is difficult to compare the results of previous work to our work. However, we have tried to approximate related work by reverting the incremental entity-mention model to a pair-wise mention-pair model and compiled a set of commonly used features. Furthermore, we can loosen the requirements for the correctness of a pronoun antecedent in the ARCS evaluation. We approximate the evaluation of pair-wise instances in related work by simply requiring that a pronoun has an antecedent in its coreference chain that is also in the key chain which the pronoun is part of, i.e. any correct antecedent can be correct. We deem this the *ARCS any antecedent metric* and use it to score our system responses. Table 5.21 shows the results.

We see that the ranking of the responses does not change compared to the evaluation using the *ARCS inferred antecedent metric*, which requires pronouns to (transitively) link to nominal antecedents. However, as expected, the scores are higher in general and the differences between the responses become slightly smaller.

Using this analysis, we can cautiously compare our results to related work in a more direct manner. Wunsch (2010), p. 201, reported an F-score of 61.7% for personal pronouns and an F-score of 52.2% for possessive pronouns for the TiMBL classifiers trained with random sampling (best results) in a ten-fold cross-validation experiment on an earlier version of the TüBa-D/Z corpus. Wunsch, p. 185 further reports an average F-score of

DEV SET

	PPER			**PPOSAT**			**ALL**		
	R	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
E-M Upper bound	*92.08*	*86.21*	*89.05*	*95.79*	*88.36*	*91.93*	*91.50*	*85.57*	*88.44*
E-M	**83.35**	**78.04**	**80.61**	**86.34**	**79.64**	**82.86**	**83.12**	**77.74**	**80.34**
E-M MLN	82.78	77.51	80.06	86.22	79.53	82.74	82.99	77.61	80.21
E-M CRF	81.90	76.71	79.22	85.18	78.57	81.74	82.40	77.07	79.65
E-M MaxEnt	80.69	75.58	78.05	84.94	78.35	81.51	81.64	76.37	78.92
E-M std.feat.set	80.45	75.33	77.81	81.71	75.37	78.41	80.77	75.55	78.07
M-P Upper bound	92.92	86.94	89.83	95.85	88.41	91.98	91.90	85.85	88.77
M-P	79.97	74.82	77.31	83.84	77.33	80.46	81.10	75.76	78.34
M-P std.feat.set	81.13	75.91	78.44	82.87	76.43	79.52	81.41	76.05	78.64

TEST SET

	PPER			**PPOSAT**			**ALL**		
	R	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
E-M Upper bound	*92.81*	*88.10*	*90.40*	*95.29*	*86.74*	*90.81*	*92.16*	*85.47*	*88.69*
E-M MLN	**84.40**	**80.15**	**82.22**	**85.24**	**77.58**	**81.23**	**84.23**	**78.12**	**81.06**
E-M	83.65	79.41	81.48	**85.24**	**77.58**	**81.23**	84.08	77.97	80.91
E-M CRF	83.52	79.28	81.35	84.66	77.05	80.68	83.84	77.75	80.68
E-M MaxEnt	82.47	78.32	80.34	**85.24**	**77.58**	**81.23**	83.49	77.44	80.35
E-M std.feat.set	82.38	78.20	80.24	81.95	74.59	78.10	82.75	76.74	79.63
M-P Upper bound	*93.78*	*88.94*	*91.30*	*95.74*	*87.15*	*91.24*	*92.67*	*85.91*	*89.17*
M-P	83.13	77.07	79.98	84.91	77.29	80.92	83.13	77.07	79.98
M-P std.feat.set	82.65	78.39	80.46	82.53	75.12	78.65	83.00	76.95	79.86

Table 5.21: Approximation of pair-wise evaluation of third-person pronoun resolution with the ARCS any antecedent metric.

82.5% for all pronoun types after applying a postfilter which links pronouns that the TiMBL classifier did not resolve to the most recent compatible subject markable. The postfilter resolved over 30% of all pronouns processed by the system and doubled the Recall score. However, the 82.5% figure includes reflexive pronouns which are comparatively trivial to resolve and which we do not process. Unfortunately, Wunsch did not give separate evaluation scores for personal and possessive pronouns of applying the postfilter. The adaption of the rule-based RAP system (Lappin and Leass, 1994) to German yielded an overall F-score of 76.6% for personal, possessive, and reflexive pronouns in the experiments of Wunsch.

Schiehlen (2004) performed experiments on the Negra corpus (Skut et al., 1997) which is compiled from articles from the German newspaper *Frankfurter Rundschau.* Schiehlen reported F-scores of 78.2% and 79.0% for personal and possessive pronouns as his best results. However, he did not disclose how he calculated Recall and Precision. The usage of a different corpus makes it difficult to compare our results. Still, the Negra corpus stems from the same domain, i.e. newspaper text.

Strube et al. (2002), Kouchnir (2004) reported results for their experiments on the Heidelberg Text Corpus (HTC), a collection of short texts containing information about sights, historic events and persons in Heidelberg. Strube et al. presented a full coreference resolution system, but also provided performance results for pronouns, evaluated in the same way as (Hinrichs et al., 2005, Wunsch, 2010). Strube et al. reported F-scores of 82.79% and 84.94% for personal and possessive pronouns. Kouchnir (2004) improved over these results on the same data by applying boosting. She reported F-scores of 87.4% and 86.9% for personal and possessive pronouns. Without the semantic features and using real preprocessing, performance dropped to an average of 67.2% accuracy[39] on a held-out HTC test set. Furthermore, Kouchnir evaluated her system on a 40-article sample form the *Spiegel* magazine using real preprocessing and found that performance dropped to 34.4% accuracy (best results), which only marginally improved over a most recent candidate baseline (31.1%).

Finally, Hartrumpf (2001) presented the only entity-mention or clustering-based system for coreference resolution for German. He evaluated his approach on roughly 500 anaphoric mentions from the German newspaper *Süddeutsche Zeitung* and reported a MUC F-score of 66.00%. However, Hartrumpf did not give separate evaluation results for pronouns.

In summary, related work on German pronouns resolution investigated pronoun resolution in a pair-wise manner and evaluated resolution performance in that perspective. We have argued that results obtained in this manner are hard to interpret regarding the potential benefit for downstream applications and have provided an alternative evaluation framework that addresses this issue. Due to different test sets and different evaluation methods, it is difficult to directly compare our results to related work. However, we have implemented an approximation of related work and showed that our incremental entity-mention model outperforms this baseline w.r.t. different evaluation settings.

5.5 Error analysis

In this section, we identify and classify errors regarding pronoun resolution in our approaches. As discussed earlier[40], error analyses in coreference resolution has recently gained attention, because it is a useful complement to performance assessment based on the common evaluation framework.

Kummerfeld and Klein (2013) and Martschat and Strube (2015) presented approaches that quantify errors on the coreference chain level. These approach are not applicable to our scenario, since we are interested in analyzing pronoun resolution specifically. The

[39] Kouchnir (2004) did not elaborate on the measure. We assume she used the terms F-score and accuracy interchangeably.

[40] Cf. section 3.3

tool presented by Martschat and Strube (2015) also allows mention-type specific error analysis, i.e. pronoun resolution errors. However, there is no fine-grained analysis beyond Precision and Recall errors and no other error categorization.

We have so far assessed performance of our approaches in different regards, i.e. classifier performance when selecting an antecedent (table 5.14) and coreference model and feature set comparisons using different criteria for the correctness of an antecedent (tables 5.16, 5.18, 5.21). Here, we try to identify common error sources of our systems.

5.5.1 Lemma-based performance comparison

In Tuggener and Klenner (2014), we performed lemma-based evaluation of pronoun resolution and found that there is a large performance gap between the masculine and feminine/plural pronouns. We here repeat this analysis.

PoS	**lem.**	**Rec.**	**Prec.**	**F1**	**Av.**	**TP**	**WL**	**FN**	**FP**	**GM**
					Entity-Mention					
PPER	er	82.59	80.10	81.33	91.95	930	189	7	42	1126
	sie	69.19	62.26	65.54	85.00	759	316	22	144	1097
PPOSAT	sein	84.18	75.83	79.78	87.53	665	119	6	93	790
	ihr	72.63	63.73	67.89	84.98	520	194	2	102	716
					Mention-Pair					
PPER	er	76.02	73.79	74.89	93.73	856	262	8	42	1126
	sie	66.27	59.79	62.86	86.63	727	345	25	144	1097
PPOSAT	sein	82.15	74.17	77.96	88.43	649	133	8	93	790
	ihr	72.63	63.88	67.97	85.98	520	192	4	102	716

Table 5.22: Lemma-based performance analysis of third person pronouns based on the ARCS inferred antecedents metric.

Table 5.22 provides a detailed lemma-based performance analysis on the test set, as given by the ARCS inferred metric. This metric requires pronouns to (transitively) link to nominal antecedents. The table gives the usual measures to the left for the entity-mention (**E-M**) and the mention-pair (**M-P**) model, but also indicates the percentage of resolved pronouns for which the correct antecedent is available (**Av.**), the ARCS mention class distribution (**TP, ..., FP**), and the gold mention count (**GM**).

Looking at the F-score column, we make two interesting and related observations. First, like in Tuggener and Klenner (2014), we see a large performance gap between the masculine pronouns and their feminine/plural counterpart. For example, the entity-mention model achieves an F-score of 81.33% for *er*, but only 65.54% F-score for *sie*. A similar difference can be observed for the possessive pronouns. Also, these differences are reflected in the F-scores for the mention-pair model. The right side of the table gives insight into the reason for these performance gaps. We see that there is a large difference in the percentages of the availability of the correct antecedent (**Av.**). For *er*, the entity-mention model has access to the correct antecedent in 91.95% of the resolved instances. For *sie*, it can only resolve to the correct antecedent in 85% of the cases. Since all pronoun instances are always resolved if there are compatible candidates in our approach, we more frequently select an incorrect candidate for *sie* than for *er*. This is also reflected in the differences in the wrong linkage counts (**WL**), i.e. the entity-mention model selects an incorrect antecedent for *sie* in 316 cases, while it only chooses the wrong candidate in 189 cases for *er*.

The unavailability of the antecedent is mainly caused by issues in the markable extraction step. Since *sie* can refer to plural entities, it can also refer to coordinated NPs, like "Paul and Mary". Such coordinated markables are much harder to extract than singular entities, since retrieving them from parse trees can be cumbersome and error-prone. We manually inspected several cases where *sie* lacked access to the correct antecedent and found that incorrectly extracted or non-extracted coordinated NPs where often the source of the problem. Other problems are non-matching markable boundaries and antecedents that have a PoS tag that we do not consider during markable extraction, e.g. substituting indefinite pronouns (PIS).

In table 5.22 we see furthermore that the false positive counts (**FP**) are much higher for *sie* than for *er*. That is, the system produces more errors due to unannotated instances for *sie*. A manual inspection of some of these cases revealed that often the pronoun simply lacked gold annotation.

Secondly, the entity-mention model mainly outperforms the mention-pair model w.r.t. the masculine pronouns. For personal pronouns, the entity-mention model supersedes the mention-pair model by 6.44 F-score points (81.33% vs. 74.89%). For possessive pronouns, the difference is present, but smaller. Also, we see from the false negative and false positive counts that the entity-mention model does not perform better due to anaphoricity detection. That is, both models have identical or very similar counts for these error classes. The performance difference between the entity-mention and the mention-pair model thus stems from the different counts for true positives and wrong linkages, which substantiates that the entity-mention model is the better performing resolution strategy.

5.5.2 Error examples

In this section, we investigate a set of hand-selected examples of pronoun resolution errors that the models produce. Although we do not provide an extensive error analysis and we do not present a systematic error categorization, we pick examples that subsume several similar errors we encountered during manual error tracking. We limit our analysis to the wrong linkage class, since false positive and false negative errors are due to preprocessing errors, which we do not cover here.

The first example shows an instance of a possessive pronoun that the entity-mention model correctly resolves which the mention-pair model incorrectly handles.

(6) Sie$_1$ denkt an den Tag danach, wenn die KollegInnen womöglich ihr$_1$ Bild in der Zeitung sehen.

She thinks about the day after, when the colleagues perhaps see her/their$_1$ picture in the newspaper.

In this example, the instance of the possessive pronoun [*ihr*]$_1$ is genuinely ambiguous. That is, without discourse context, the pronoun can be linked to either [*Sie*]$_1$ or [*KollegInnen*], and both resolutions yield semantically valid utterances. However, the previous sentences in this context have only mentioned the entity denoted by [*Sie*]$_1$, but not [*KollegInnen*], i.e. [*KollegInnen*] is a discourse-new entity, while [*Sie*]$_1$ denotes a discourse-old entity this context. Due to its local confinement, the mention-pair model has no notion of information status of entities and NPs denoting them. Therefore, the mention-pair model incorrectly attaches [*ihr*]$_1$ to [*KollegInnen*] based on proximity preferences. The entity-mention model, which incorporates the discourse-new and discourse-old distinction, correctly links it to [*Sie*]$_1$ based on its bias towards favoring discourse-old entities as antecedents.

The next example shows a relative pronoun which both models incorrectly resolve.

(7) Hier werden Beiträge$_1$ kleiner Leute veraast, die$_1$ von ehrenamtlichen Kassierern fünf Mark weise gesammelt werden.

Donations$_1$ of ordinary people are being wasted, that$_1$ are collected by volunteers, 5 Mark per donation.

Here, the strategy to select the most recent compatible antecedent candidate for relative pronouns fails. Both models resolve [*die*$_1$] to [*Leute*]. To correct this error, the models would need a notion of semantic compatibility to infer that volunteers generally collect donations rather than ordinary people. However, a simple model of verb selectional preferences might not suffice, since people might be a relevant direct object of "sammeln" in such a model. The discriminatory factor here is the adverbial phrase "5 Mark a piece" which modifies the verb, since it is unlikely that one collects or gathers people at a 5 Mark rate.

In the next example, the models fail because the pronoun occurs in direct speech and refers to an antecedent in a previous direct speech segment.

(8) Endlich kommt einer und fragt, warum der Mann$_1$ nicht sein zweites Bein benutzt. “Warum wohl?”, fragt Kilian das Publikum und gibt sich selbst die Antwort. “Er$_1$ hatte es einfach nicht bemerkt.”

Finally, someone approaches and asks why the man$_1$ does not use his second leg. “Why?”, Kilian asks the audience and gives the answer himself. “He$_1$ simply had not noticed it.”

Both models here resolve [Er_1] to [$Killian$] because given our feature set, it is a more probable antecedent when direct speech is disregarded. Clearly, our approach needs an elaborate strategy to disentangle direct speech segments from other parts in the discourse to resolve such pronoun instances.

Finally, the following example in figure 5.7 illustrates the major differences of the models by looking at a longer discourse segment, i.e. beyond pair-wise decisions.[41] We use the subscript to indicate entity IDs and superscript to enumerate mentions, i.e. lexeme$^{mentionID}_{entityID}$. Thus, lexemes with identical entity IDs are coreferent. ’*’ in the responses denotes that mentions or entities are invented or overlooked. Table 5.23 lists the coreference chains of the key and the two responses.

Key:	$[Jusef^{1}, Er^{2}, er^{3}, seine^{4}, seinen^{5}, Jusef^{9}, seine^{10}, ich^{11}, meiner^{12}, meiner^{13}, Jusef^{14}]_{1}$ $[Vater^{6}, ihn^{7}, er^{8}]_{2}$
E-M:	$[Jusef^{1}, Er^{2}, er^{3}, seine^{4}, seinen^{5}, ihn^{7}, er^{8}, Jusef^{9}, seine^{10}, Jusef^{14}]_{1}$ $[ich^{11}, meiner^{12}, meiner^{13}]_{2}$
M-P:	$[Jusef^{1}, Er^{2}, er^{3}, seine^{4}, Jusef^{9}, seine^{10}, Jusef^{14}]_{1}$ $[Regime^{*}, seinen^{5}, ihn^{7}, er^{8}]_{2}$ $[ich^{11}, meiner^{12}, meiner^{13}]_{3}$

Table 5.23: Gold and predicted coreference chains for the segment.

We see that the entity-mention model makes two errors. First, it links ihn^7 and er^8 to the $Jusef_1$ entity instead of to $Vater_2$. Here, the weight for favoring discourse old entities misguides resolution. Secondly, the model is unable to attach the first person pronouns in the direct speech segment to the $Jusef_1$ entity, since we only allow matching first person pronouns to antecedents governed by a communication verb at most one sentence away. The right mention here would be $Jusef_1^{14}$. However, it occurs after the first person pronouns and we do not allow pronouns to link to postcedents.

[41] English translation: Jusef gets up. He moves his eyes away from the wall and starts to talk. Of Afghanistan, where he and his family led a happy life. Until the new regime came. One day, the Taliban stood at the door, took his father and shot him, because he was a communist, allegedly. The mother thereafter had Jusef and his sister taken out of the country by their uncle. “No I am here and I have no contact with my mother and my family”, Jusef says.

(9) **Key:**

Da steht Jusef$_1^1$ auf. Er$_1^2$ wendet den Blick von der Wand und fängt an zu erzählen. Von Afghanistan, wo er$_1^3$ und seine$_1^4$ Familie ein glückliches Leben führten. Bis das neue Regime kam. Eines Tages standen die Taliban vor der Tür, nahmen seinen$_1^5$ Vater$_2^6$ mit und erschossen ihn$_2^7$, weil er$_2^8$ angeblich Kommunist war. Die Mutter ließ daraufhin Jusef$_1^9$ und seine$_1^{10}$ Schwester Abeda vom Onkel außer Landes schaffen. “Jetzt bin ich$_1^{11}$ hier und habe keinen Kontakt zu meiner$_1^{12}$ Mutter, auch nicht zu meiner$_1^{13}$ Familie”, sagt Jusef$_1^{14}$.

Entity-mention response:

Da steht Jusef$_1^1$ auf. Er$_1^2$ wendet den Blick von der Wand und fängt an zu erzählen. Von Afghanistan, wo er$_1^3$ und seine$_1^4$ Familie ein glückliches Leben führten. Bis das neue Regime kam. Eines Tages standen die Taliban vor der Tür, nahmen seinen$_1^5$ Vater$_*^6$ mit und erschossen ihn$_1^7$ *, weil er$_1^8$ angeblich Kommunist war. Die Mutter ließ daraufhin Jusef$_1^9$ und seine$_1^{10}$ Schwester Abeda vom Onkel außer Landes schaffen. “Jetzt bin ich$_2^{11}$ * hier und habe keinen Kontakt zu meiner$_2^{12}$ Mutter, auch nicht zu meiner$_2^{13}$ Familie”, sagt Jusef$_1^{14}$.

Mention-pair response:

Da steht Jusef$_1^1$ auf. Er$_1^2$ wendet den Blick von der Wand und fängt an zu erzählen. Von Afghanistan, wo er$_1^3$ und seine$_1^4$ Familie ein glückliches Leben führten. Bis das neue Regime$_2^*$ kam. Eines Tages standen die Taliban vor der Tür, nahmen seinen$_2^5$ * Vater$_*^6$ mit und erschossen ihn$_2^7$, weil er$_2^8$ angeblich Kommunist war. Die Mutter ließ daraufhin Jusef$_1^9$ und seine$_1^{10}$ Schwester Abeda vom Onkel außer Landes schaffen. “Jetzt bin ich$_3^{11}$ *hier und habe keinen Kontakt zu meiner$_3^{12}$ Mutter, auch nicht zu meiner$_3^{13}$ Familie”, sagt Jusef$_1^{14}$.

FIGURE 5.7: Example responses of the models. Incorrectly resolved pronouns are marked by a star (*).

The mention-pair model, on the other hand, makes three resolution errors. It also incorrectly resolves the two personal pronouns (ihn^7, er^8) that denote the $Vater_2$ entity. However, it identifies $Regime^*$ as the antecedent entity. This happens because the model forms the pairs [$Regime^* - seinen^5$], [$seinen^5 - ihn^7$], and [$ihn^7 - er^8$]. The transitive merge then yields the coreference chain which features inconsistent morphology (i.e. $Regime^*$ and ihn^7, and $Regime^*$ and er^8 are exclusive). Finally, the mention-pair model also fails to link the first person pronouns to the $Jusef_1$ entity, since we apply the same heuristics as in the entity-mention model.

5.5.3 Candidate ranking performance

Finally, we analyze the quality of the ranking of the antecedent candidates in those cases where the entity-mention model selects an incorrect one. This analysis provides an additional insight into the ranking strength of our approach.

We analyse the rank index of the correct antecedent in cases where our approach selects an incorrect antecedent candidate by tracking the frequency of the different rank indices of the correct antecedent for incorrectly resolved pronouns.

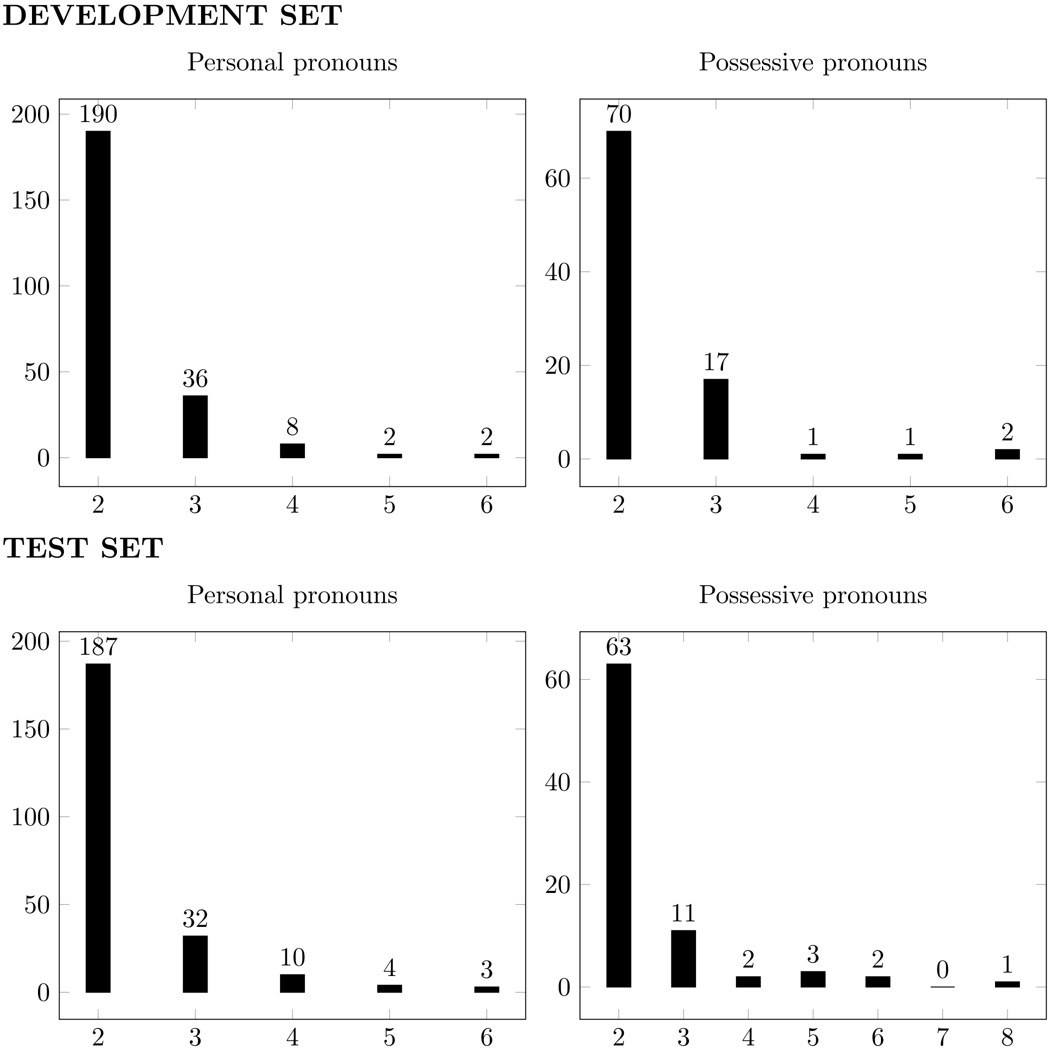

Figure 5.8: Rank index frequency of the correct antecedent for third person pronouns where the entity-mention model ranks an incorrect candidate highest. A lower rank means a higher score for a candidate w.r.t. denoting the antecedent.

Figure 5.8 shows the rank frequency of the correct antecedent for such incorrectly resolved personal and possessive pronouns. Note that the ranking is inversed, i.e. the lower the rank, the better the candidate, i.e. rank 1 denotes the selected antecedent. We see that in the cases where the entity-mention model selects an incorrect antecedent, the correct one is ranked 2nd in 79.83% (development set; i.e. 190 of 238 total cases) and 79.24% (test set) of the personal pronoun instances and in 76.92% (development set) and 76.82% (test set) of the possessive pronoun cases. That is, selecting the second best candidate as antecedent in the cases where the first one is incorrect would reduce almost 80% of the errors made by the classifiers in the entity-mention model. Thus, re-ranking the top two candidates seems to be a fruitful notion. We explore such a re-ranking attempt based on distributional semantics in the next chapter.

5.6 Chapter summary

In this chapter, we empirically validated our incremental entity-mention approach for coreference resolution for German. We evaluated pronoun resolution in detail and applied different evaluation measures and strategies that assess performance on different levels.

Throughout our evaluation, we substantiated the proclaimed theoretical advantages of our incremental entity-mention model over the commonly used mention-pair model with empirical evidence. In all our experiments, the entity-mention model outperformed the mention-pair competitor.

We introduced an extended feature set for German pronoun resolution and compared it to a standard feature set typically encountered in related work. Our extended set improved performance of the classifiers in both the mention-pair and the entity-mention model. In subsequent evaluation, the extended set did improve performance of the entity-mention model. However, for the mention-pair model, the extended features only improved resolution performance for possessive pronouns, while lowering the performance on personal pronouns.

Furthermore, we investigated and compared different machine learning frameworks that correspond to different antecedent selection methods. We introduced our own simple feature weighting scheme which performed on par with the top ranking machine learning approaches. Overall, we found that the different machine learning approaches and the respective antecedent selection strategies did not show substantial performance differences in the top ranks. We found larger performance differences between the mention-pair and the entity-mention model.

In error analysis, we demonstrated that the mention-pair model does indeed produce coreference chains with inconsistent morphological properties due to underspecification of certain German pronouns. In our entity-mention approach, such errors are avoided.

Chapter 6

Semantics for pronoun resolution

In this chapter, we explore distributional semantics as a device to determine the degree of compatibility between an antecedent candidate and a pronoun's context. As we saw in the analysis of error examples in section 5.5.2, our approach to ranking candidates, which is primarily based on the discourse salience of the candidates, sometimes selects an antecedent that is either incompatible or less compatible with the pronoun's context than the correct antecedent.

In one of our previous examples, the salience-based approach selected the antecedent "people" for the pronoun "them" in the verb-argument tuple "collect them", although the correct antecedent "donations" was accessible. Obviously, "donations" is a more likely candidate for the direct object slot of the verb "to collect" than "people". Thus, selectional preferences of verbs are potentially beneficial to resolving pronouns. However, successfully incorporating these preferences into a real-world pronoun resolution system has proven to be notoriously difficult (Kehler et al., 2004, Wunsch, 2010, inter alia).

We explore two frameworks to model compatibility of antecedent candidates and a pronoun's context, a co-occurrence graph and word embeddings. The co-occurrence graph estimates compatibility by traversing weighted co-occurrence paths between nodes that denote nouns and verbs. The word embedding model represents words as vectors and estimates compatibility of words based on the cosine of their vectors.

We expand on previous work by estimating not only the compatibility between the verb argument slot of a pronoun and the antecedent candidate, but by also taking into account the additional verb argument of the verb governing the pronoun, i.e. the syntactic co-argument of the pronoun in cases of (di-)transitive verbs. We first outline the rationale for exploring verb semantics for pronoun resolution and then present our models for doing so.

6.1 Pronominalization as a discourse phenomenon

From a discourse perspective, pronoun resolution can be viewed as the task of determining which entity is most likely to be mentioned at a given point in discourse, i.e. when a pronoun is encountered. That is, the task of pronoun resolution is to assess which of the previously mentioned entities in the discourse is likely to be discussed at the very point of encountering a pronoun.

In general, pronoun resolution approaches, including our own, work by modeling the salience of discourse entities based on a set of features that captures relevant aspects of their occurrences, such as grammatical functions etc. If a pronoun is encountered, the most salient entity is chosen as antecedent.

However, these models do not exclusively model pronominalization, but provide a general model of entity salience in discourse.[1] For example, given we have established the salience record of entities in a particular discourse, we could point to any position within the discourse and blank out the subsequent sentence. Querying the salience model, we could then determine which entities are likely to be mentioned in the subsequent sentence, based on the salience configuration at the point we have chosen, without any hint regarding the specifics of the subsequent sentence. If we knew that there is a pronoun in the next sentence, we could determine the entity in the current sentence which is most likely to be pronominalized in the next sentence. Neglecting many important details, this can be argued to be the general model for pronoun resolution in the majority of approaches.

That is, from the perspective of pronoun resolution, there is an overlooked aspect in this model, i.e. the local context in which a pronoun occurs. The majority of our features are concerned with bookkeeping the salience of the discourse entities, regardless of particular pronouns and their respective, specific contexts. If a pronoun is encountered at a certain point in discourse, the most salient entity at that point is looked up in the salience record and chosen as antecedent, ignoring the specific context that surrounds the pronoun.

This is somewhat surprising, given that the antecedent of a pronoun has to be compatible with the context surrounding the pronoun, because, compared to its nominal antecedent, the pronoun is simply an altered linguistic manifestation of the same underlying entity. Therefore, it can be argued that the pronoun's context itself emits certain expectations regarding the antecedent. Consider the following example:

(10) Er bellt.
He barks.

[1] Cf. Centering Theory (Grosz et al., 1995), for example.

Although we are not given any discourse history of entities and their salience, we automatically infer, based on the selectional preference of the verb *bellen*, that *Er* is likely to refer to something canine-like. This likelihood is ignored in the purely salience-based model.

Thus, there exists a line of research dating back to the pioneering era of automated pronoun resolution that has attempted to incorporate the context surrounding the pronoun in the antecedent selection process.[2] The main road that researchers have taken in this direction is to represent the context of a pronoun by the verb that governs the pronoun. The selectional preferences of that verb are then taken to rank the antecedent candidates.

6.1.1 Selectional preferences of verbs for pronoun resolution

The underlying idea of approaches that incorporate selectional preferences into pronoun resolution is straight-forward. As the pronoun is simply an altered surface representation of the underlying entity, its nominal antecedent has to be compatible with the selectional preference of the verb governing the pronoun. A typical approach would rank a set of antecedent candidates according to their degree of preference for the verb argument slot of the pronoun and suggest the most preferred one as the antecedent.

A famous example where such a model has clear benefits over a salience-based approach is the following:

(11) If the baby doesn't thrive on raw milk, boil it.

Although *baby* is in the subject position, which increases its salience, we know that it cannot be the antecedent of the pronoun *it*, because it violates the selectional preference of *to boil*.

Formally, for applying verb selectional preference to the example above, we formulate that the probability of choosing *baby* as the antecedent is lower than choosing *milk*, based on the verb selectional preference. That is, we want to express that $p(baby|boil, obja) < p(milk|boil, obja)$. The conditional probabilities can be harvested from large corpora in a distributional approach.

[2] We discuss these in more detail in section 6.6.

6.1.2 Issues of selectional preferences and potential solutions

Deriving an adequate model of selectional preferences of verbs is a challenging task in itself (Erk et al., 2010, Schulte im Walde, 2010, inter alia). The crucial issue w.r.t. pronoun resolution is whether the preference ranking of antecedent candidates given the verb governing a pronoun is always the right choice. In other words, if we assume a perfect model of selectional preferences of verbs, is the most preferred candidate always the best one, or are there contexts in which other factors override the preferences and favor a less preferable candidate? The tentatively negative results in related work that incorporates verb selectional preferences into pronoun resolution suggest the latter.

However, we argue that there are certain aspects that related work has so far not taken into account and that the (non-)impact of modelling selectional preferences for pronoun resolution is closely related to the three following issues.

1. **Selectional preferences alone are not (always) a sufficient representation of a pronoun's context**: While there are verbs that have an intuitively "narrow" selection regarding its arguments, like *bellen (to bark)* towards its subjects, there are verbs that do not feature a selection narrow enough to give a clear preference ranking to a set of antecedent candidates. For example, we cannot expect to derive distinctive preferences for *machen (to make)* and its subjects that would allow us to prefer a specific candidate. To address this issue, we incorporate the additional verb arguments of the verb governing the pronoun into the compatibility model of candidates and pronoun contexts. By doing so, we aim to narrow down the selectional preferences of verbs that have an otherwise wide array of permissible arguments in the pronoun's argument slot.

2. **Sparseness of co-occurrence counts:** Although large corpora are available, we cannot expect to see every permissible argument for a given grammatical slot (i.e. we might never see the tuple $(boil, milk, obja)$ even in large corpora). Related approaches have tried to alleviate this problem by either clustering verbs based on their arguments (Bergsma et al., 2008a) or by performing class abstraction on the arguments using a word net (Wunsch, 2010).[3]

   An uncertainty in clustering the verbs is the number of clusters to choose. Also, infrequent verbs are likely to be merged into clusters that feature low similarity between the verbs.

   When performing class abstraction of arguments, the arguments of a verb that are found in a corpus are mapped to their hypernyms in a word net. Selectional

[3] Cf. section 6.6.

preferences are then learned for verbs and word net hypernyms. During pronoun resolution, antecedent candidates are also mapped to their hypernyms and compatibility of the candidates is approximated by the compatibility of their hypernyms with those selected by the verbs governing the pronouns. This approach has the common downside of manually crafted resources that not all candidates and arguments can be mapped to word net senses. Perhaps more importantly, the level of abstraction has to be determined, i.e. how far up the hypernym hierarchy arguments should be projected. If a low level is selected, sparsity might not be reduced sufficiently. If the level of abstraction is too high, the selectional preferences might become very coarse-grained and cannot distinguish antecedent candidates anymore.

Thus, we present methods to alleviate the sparseness problem that do not rely on class abstraction and clustering.

3. **Pronominal antecedents:** A major issue in related work is how pronominal antecedents are processed. Since the selectional preferences of a verb governing a pronoun can only be applied to nominal antecedents, it is unclear how pronominal antecedents (antecedents that are pronouns themselves) are handled in related work. Here, our incremental entity-mention model has a clear advantage, since pronominal antecedents are likely to have been resolved before becoming antecedents. Thus, we can query the incrementally established coreference chain of a pronominal antecedent to retrieve a nominal mention and use this nominal mention to assess compatibility of the underlying entity and the pronoun's context.

With these issues in mind, we next present our two models that estimate compatibility of antecedent candidates and a pronoun's context.

6.2 A graph-based representation of verb-argument tuples

The first distributional compatibility model we investigate is a graph-based representation of verb-argument tuples. This model will enable us to estimate the selectional strength of a verb and an argument based on their first-order co-occurrence[4] which we use to estimate the compatibility of an antecedent candidate and the verb governing a pronoun. Furthermore, the graph representation allows us to model compatibility between arguments, i.e. an antecedent candidate and an additional argument of verb

[4] First-order co-occurrence denotes that e.g. two words occur together in a specified context. This context can be defined in different ways, e.g. documents, sentences, or word windows. In our case, we focus on syntactic structures and relations. A verb and an argument show a first-order co-occurrence if they occur together given a specified grammatical relation, e.g. a verb and its subject.

governing the pronoun, based on their second-order co-occurrence. To determine the degree of compatibility between an antecedent candidate and a pronoun's context, we combine the selectional preference of the verb governing the pronoun w.r.t. the candidate, and the candidate's compatibility with the additional arguments of the pronoun's verb.

A straight-forward way of measuring compatibility of e.g. a noun w.r.t. the subject position of a transitive verb that governs a specific direct object would be to extend a common association measure, such as Pointwise Mutual Information (PMI), to three variables, as in Van De Cruys (2011):

$$PMI(subj, verb, object) = \log \frac{P(subj, verb, object)}{P(subj)P(verb)P(object)} \tag{6.1}$$

While such an approach is suited for identifying association strengths of seen triples, it is bound to be sparse for many possible but unseen combinations. That is, we cannot expect to find all triple combinations *(subj,verb,object)* for cases where e.g. a pronoun is in subject position and where we want to calculate the PMI for each antecedent candidate in the pronoun's slot, i.e. $PMI(ante\_noun_{subj}, verb, object)$. Therefore, we aim for a compositional approach that combines the pair-wise compatibilities of the verb and the candidate, $comp(ante\_noun_{subj}, verb)$, and the candidate and the additional argument of the verb, $comp(ante\_noun_{subj}, add\_arg_{object})$, into an overall compatibility score.

We next outline how we calculate these compatibility scores.

6.2.1 Verb selectional preferences as first order co-occurrences

In its simplest form, a compatibility function $comp(\cdot,\cdot,\cdot)$ of an antecedent candidate noun a_i and the verb v_j governing a pronoun in the grammatical slot gf_k is their first-order co-occurrence count:

$$comp(v_j, a_i, gf_k) = |v_j, a_i, gf_k| \tag{6.2}$$

Selecting the most suited antecedent from a set of candidates simply involves determining the antecedent candidate a_i with the highest first-order co-occurrence count with the verb governing the pronoun among all candidates a:

$$ante = \arg\max_{a} comp(v_j, a_i, gf_k) \tag{6.3}$$

However, the count does not measure associativity of the verb and its argument w.r.t the overall occurrence distribution of each. That is, we might see a verb-argument tuple with a high count and conclude that the two are strongly associated. However, the high count can stem from the overall high occurrence of both argument and verb. Thus, the raw count of a verb-argument tuple needs to be normalized in relation to the individual, overall occurrence counts of verb and argument.

There are association measures that perform such (point-wise) normalizations, e.g. DICE coefficient, TF IDF, Jaccard coefficient, and Pointwise Mutual Information (PMI), among others.[5] In the realm of the vector space models, the Positive Pointwise Mutual Information measure (PPMI) is a popular choice (Turney et al., 2010). PMI of two words x and y is defined as the logarithm of the joint probability divided by the product of the marginal probabilities. It thus takes into account the co-occurrences of two words and the individual occurrences of the words.

$$PMI(x;y) = \log \frac{p(x,y)}{p(x)p(y)} \tag{6.4}$$

PPMI replaces all negative values in a word co-occurrence matrix with zeros and has been shown to outperform PMI and other association measures on several tasks (Bullinaria and Levy, 2007).

The main issue for our purposes is that PMI can have negative values for seen, but loosely associated word pairs. One option would be to use PPMI and replace all negative values with zero. However, we want to reward seen verb-argument pairs that have a negative PMI value over non-seen pairs. Non-seen pairs are not represented by PMI, and we would have to assign a dummy value for these verb-argument combinations, i.e. zero. However, zeroing out negative values and assigning zero to unseen pairs would obscure whether a zero-valued pair had a negative PMI or was not seen at all. Thus, we would not be able to favor an antecedent candidate with a negative PMI over a candidate that has never been seen as an argument of the verb governing the pronoun. Therefore, we aim for a compatibility measure with a range of $0 < comp(\cdot,\cdot,\cdot) \leq 1$ for all seen pairs and zero for unseen pairs.

All the association measures mentioned above share the characteristic that they normalize the first-order co-occurrence count of the word pair by the individual occurrence counts of the words. We aim for a simple association measure and a simple scaling procedure along this line. To do so, we model selectional preferences in a bidirectional manner. We model how strong the verb's selection of the argument is and how strong

[5] Cf. e.g. Evert (2004)

the argument's selection of the verb is given a specific grammatical relation. Note that these two associations are asymmetrical, since a verb and an argument typically have a different overall count of first-order co-occurrences.

We formulate this bidirectional association in relation to conditional probabilities, i.e. the likelihood of seeing the argument a_i given the verb v_j and the grammatical slot gf_k, and the likelihood of seeing the verb given the argument and the grammatical slot, but do not actually model a probability distribution. To derive a compatibility score, we simply take the arithmetic mean of the bidirectional scores:

$$comp(a_i, v_j, gf_k) = \frac{1}{2} * \big(score(a_i|v_j, gf_k) + score(v_j|a_i, gf_k)\big) \tag{6.5}$$

where we calculate the scores by:

$$\begin{aligned} score(a_i|v_j, gf_k) &= \frac{|(a_i, v_j, gf_k)|}{|\max(a, v_j, gf_k)|} \\ score(v_j|a_i, gf_k) &= \frac{|(a_i, v_j, gf_k)|}{|\max(a_i, v, gf_k)|} \end{aligned} \tag{6.6}$$

That is, we normalize the first-order co-occurrence counts, i.e. $|(a_i, v_j, gf_k)|$, by division by the highest count, i.e. $|\max(a, v_j, gf_k)|$, instead of taking the sum in the denominator (which would yield a probability distribution). This is a common replacement in e.g. TF IDF calculation, where in-document term frequency of each word is divided by the count of the most frequent word instead of the overall word count. Doing so decreases the denominator and thus yields higher scores overall.

Consider the following example. We want to score the triple $(Hund, bellen, subj)$, i.e. the compatibility between *Hund* (*dog*) and *bellen* (*bark*) given the grammatical relation subject. Figure 6.1 shows an excerpt of the co-occurrence graph that depicts first-order co-occurrence for *Hund* and *bellen* given the grammatical relation subject. The nodes in the graph denote words, i.e. nouns and verbs, and edges signify grammatical relations of seen co-occurrences and their counts.

The first-order co-occurrence count is $|(Hund, bellen, subj)| = 261$. The maximal count of *Hund* as a subject is $|\max(Hund, v, subj)| = 440$ (for v = *kommen* (*to come*))

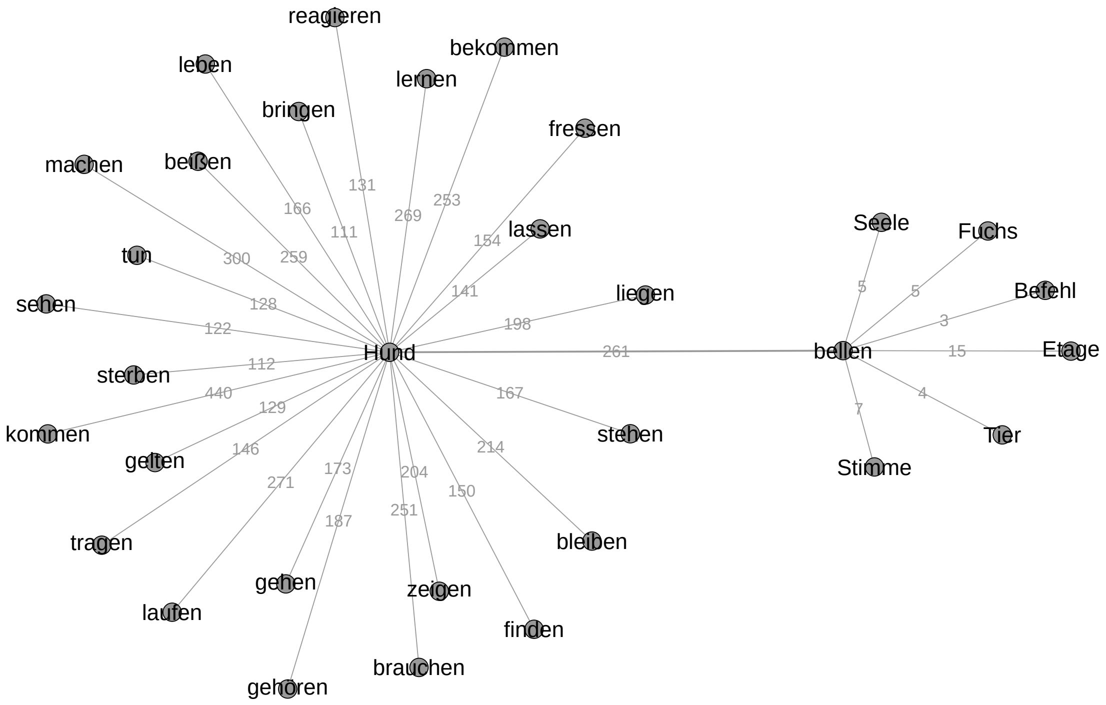

FIGURE 6.1: Excerpt of the co-occurrence graph, showing first-order co-occurrence subject relations for the noun *Hund* and the verb *bellen*. Numbers on edges denote absolute counts.

and the count of the most frequent subject of *bellen* is $|\max(a, bellen, subj) = 261|$ (i.e. $a = Hund$). We then get a compibility score of $\frac{1}{2} * (\frac{261}{261} + \frac{261}{440}) = 0.8$.[6]

This compatibility score has the advantage over (P)PMI that it ranges from 0 to 1, i.e. low counts for seen combinations are still above 0 and the maximum score is 1 (for a hypothetical combination of a noun and a verb that exclusively occur together), and we can assign zero to unseen verb-argument combinations.

6.2.2 Addressing sparsity

We still face the sparsity problem, since we cannot expect to see all permissible verb-argument combinations even in a large corpus. In Tuggener and Klenner (2012), we proposed an approach that uses non-negative matrix factorization to estimate counts of unseen verb-argument combinations. Here, we explore three simpler approaches that are expressed naturally in the graph representation of word co-occurrences.

[6] Taking the sum of all counts in the denominators would yield $\frac{1}{2} * (\frac{261}{300} + \frac{261}{14738}) = 0.4$, which is rather low. We want combinations that are intuitively strong, like *dog* being the subject of *bark* to have a high compatibility, i.e. close to 1. Taking the max count in the denominator gets us closer than taking the sum. Also, the Jaccard and DICE coefficients yield rather low scores for the $dog, bark, subj$ combination, i.e. $Jacc = \frac{A \cap B}{A \cup B} = \frac{261}{300+14738} = 0.0174$ and $DICE = \frac{2*|A+B|}{|A|+|B|} = \frac{2*261}{300+14738} = 0.0347$.

6.2.2.1 Estimating compatibility through distributional siblings

First, we compare the seen argument fillers (i.e. the nouns $args$ in a grammatical slot gf_k for a verb v_j) with the antecedent candidate noun under scrutiny, a_i, and select the compatibility of the argument $arg \in args$ that is most similar to a_i as the compatibility rating of a_i and the pronoun verb v_j in slot gf_k.

$$\begin{gathered} comp(v_j, a_i, gf_k) \approx comp(v_j, arg, gf_k) \\ \text{where } arg = \underset{args}{\arg\max}\; sim(a_i, arg, gf_k) \end{gathered} \tag{6.7}$$

To determine the verb argument most similar to the antecedent candidate noun at hand, we need a similarity measure of nouns w.r.t. a specific grammatical relation, i.e. $sim(\cdot,\cdot,\cdot)$. For our purposes, we want to define noun similarity in relation to the verbs they occur with, since we want nouns to be of high similarity if they frequently occur with the same verbs w.r.t. the grammatical function of interest.

For example, suppose we have not seen “banana” as the direct object of “to eat”. We then want to find the direct object of “to eat” which is most similar to “banana”, given the grammatical relation direct object, e.g. “apple”. To achieve this, we model noun similarity based on second-order co-occurrence with verbs. As mentioned, we want nouns to be of high similarity if they frequently occur in the same grammatical argument slots of the same verbs.

A straight-forward approach for this purpose is to consider the number of verbs that the two nouns co-occur with in the given grammatical relation. This second-order co-occurrence then serves as the basis for calculating the distributional similarity of the nouns.

We calculate the similarity of two nouns n_i, n_j given a set of verbs v and an argument slot gf_k, as the ratio of verbs that they share as first-order co-occurrences divided by all their individual first-order co-occurrences w.r.t. gf_k, i.e.:

$$sim(n_i, n_j, gf_k) = \frac{|\forall v \in V| : |(v, n_i, gf_k)| > 0 \wedge |(v, n_j, gf_k)| > 0}{|\forall v \in V| : |(v, n_i, gf_k)| > 0 + \big|\forall v \in V| : |(v, n_j, gf_k)| > 0} \tag{6.8}$$

where the numerator counts the verbs that the two nouns share as first-order co-occurrences and the denominator counts all first-order co-occurrences of the two nouns.

In the graph representation, this corresponds to the count of verb nodes that the nouns share as neighbors, divided by the total number of neighboring nodes the two nouns connect to, given a specific grammatical relation. Figure 6.2 shows this overlap of first-order co-occurrences in the graph. Here, we determine similarity of $Hund$ (dog) and $Fuchs$ (fox) given the grammatical relation subject. The similarity of the two nouns is depicted by the ratio of nodes that both nouns connect to (in the center of the figure) divided by the total number of verb nodes they connect, given the subject relation. That is, the more verb nodes they share as neighbors, the more similar the two nouns are.

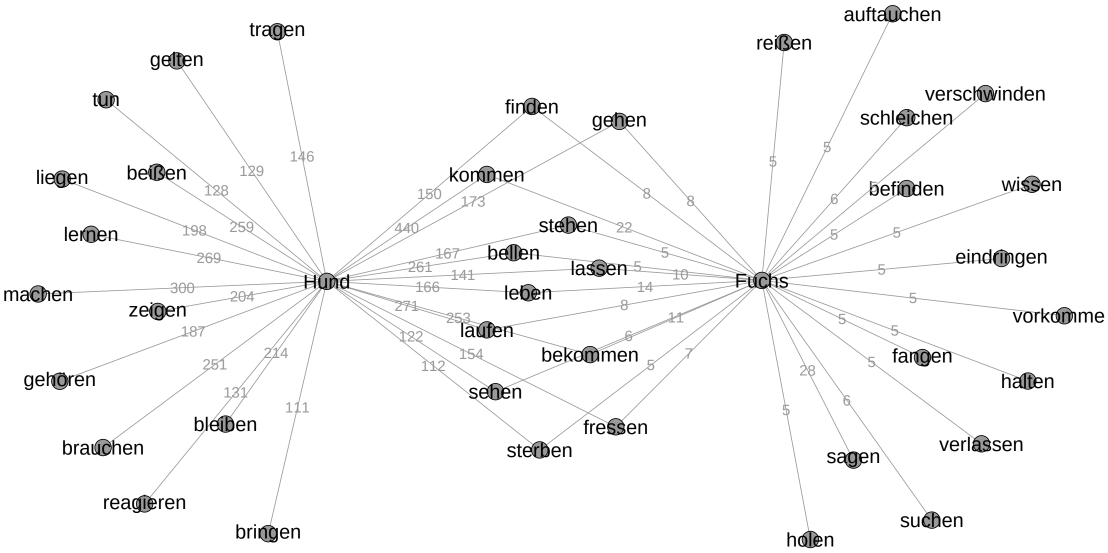

FIGURE 6.2: Excerpt of the co-occurrence graph, showing second-order co-occurrence of $Hund$ and $Fuchs$ in subject position. Numbers on edges denote absolute counts.

However, there are nodes (verbs) that the nouns more strongly associate with than others. In the previous section, we have defined an association measure for nouns and verbs, i.e. $comp(noun, verb, gram.\ funct.)$. We can use this measure in our similarity measure to weight the importance of the nodes that two nouns share. That is, a shared node that both nodes strongly associate with should have more impact on the similarity measure than a node with lesser association strengths w.r.t. the nouns. Thus, we replace the counts in equation 6.8 by the sum of the compatibility scores:

$$sim(n_i, n_j, gf_k) = \frac{\sum_{v \in V:|(v,n_i,gf_k)|>0 \wedge |(v,n_j,gf_k)|>0} comp(v, n_i, gf_k) + comp(v, n_j, gf_k)}{\sum_{v \in V:|(v,n_i,gf_k)|>0} comp(v, n_i, gf_k) + \sum_{v \in V:|(v,n_j,gf_k)|>0} comp(v, n_j, gf_k)} \tag{6.9}$$

where the numerator simply sums all edge weights (where the weights are calculated by $comp(\cdot, \cdot, \cdot)$) from the nouns to the shared verbs, and the denominator sums all edge weights that the two nouns are connected to.

For our sparsity problem, where we have not seen a specific noun-verb combination, we can now identify the seen noun argument of the verb that is most similar to our target noun, based on equation 6.7.

Returning to our previous example, where we would like to estimate the compatibility of "banana" as the direct object (*obja*) of the verb "to eat" and where we have not seen the combination, we first identify all first-order co-occurrences of the verb "to eat" with the grammatical relation direct object, i.e. $args$. Then, we calculate $sim(banana, arg, obja)$ for all these first-order co-occurrences $arg \in args$. The arg most similar to "banana", say "apple" with a similarity of 0.45, then serves as the distributional sibling of "banana", and we can take compatibility score $comp(apple, eat, obja)$, which is 0.55, as the score for "banana". However, since we have not seen "banana" as the direct object of "to eat" we want to exercise caution in taking over the score of "apple". While in our example the similarity between the target noun and its distributional sibling is obvious, we might identify siblings that are less similar to the target noun. Therefore, we multiply the compatibility score of "apple" as the direct object of "to eat" with the similarity between "apple" and "banana". This product then serves as the final compatibility score of the unseen pair:

$$\begin{gathered} comp(v_j, a_i, gf_k) \approx comp(v_j, arg, gf_k) * sim(a_i, arg, gf_k) \\ \text{where } arg = \underset{args}{\text{arg max}}\ sim(a_i, arg, gf_k) \end{gathered} \tag{6.10}$$

That is, if the target noun and its sibling are very similar, the compatibility score of the sibling will not be lowered significantly, but it will decrease with increasing dissimilarity between the target noun and its sibling.

6.2.2.2 Similarity to $n$best arguments

As a second measure for compatibility between an antecedent candidate noun and a verb governing a pronoun at hand, we measure the similarity of the candidate noun to the n most strongly associated arguments of the verb in the grammatical function slot of the pronoun. We determine n to be the 10 most strongly associated arguments if there are more than 100 seen arguments in the specific grammatical slot of the verb. If there are less than 100, we take the top 10% of the arguments (three at least).

In our example, we would measure similarity of "banana" to the n most strongly associated direct objects of "to eat", i.e. *Fleisch (meat), Brot (bread), Obst (fruit), Gemüse (vegetables), Eis (ice), Kleinigkeit (snack), Pizza, Schokolade (chocolate), Mittag (lunch), Salat (salad).* The average similarity then serves as the compatibility score, in this case 0.41.

6.2.2.3 Compatibility of verbs

Thirdly, we measure the similarity of the verb governing the antecedent candidate and the verb governing the pronoun w.r.t. their grammatical functions, i.e. $sim(v_i, gf_k, v_j, gf_l)$. Since our similarity score is based on shared arguments, it can be interpreted as a measure of how likely it is to see an argument of the antecedent verb as an argument of the pronoun verb. Assume for example the following sentences, where we want to resolve the last pronoun, i.e. sie_1^4:

(12) Sie^1 schenkt ihm eine Banane^2. Er schält die Banane$_1^3$ und isst sie$_1^4$.

She presents him with a banana. He peels the banana$_1$ and eats it$_1$.

Our similarity measure between verbs is geared at estimating how likely it is for a noun that is peeled ($Banane_1^3$ is the direct object of "to peel") to be eaten (sie_1^4 is the direct object of "to eat") versus how likely it is for someone that presents someone with something (Sie^1 is the subject of "to present") to be eaten etc. In this case $sim(peel, obja, eat, obja) = 0.14$. By contrast, subjects of "to present" are less likely to be eaten, i.e. $sim(present, subj, eat, obja) = 0.04$. Also, nouns that are presented to someone are less likely to be eaten than nouns that are peeled, i.e. $sim(present, obja, eat, obja) = 0.10$. Thus, the compatibility of the verb governing the antecedent candidate and the verb governing the pronoun can be an additional cue to identify the correct antecedent.

6.2.3 Compatibility with additional verb arguments

Finally, we estimate the compatibility of an antecedent candidate noun and the pronoun's context by taking into account additional arguments of the verb governing the pronoun. We do so not only to address sparsity, but to render the pronoun's context more specific to address the problem that certain verbs do not feature a narrow selection w.r.t. the grammatical slot of the pronoun. Given intransitive verbs, there are no other content words (i.e. non-stop words) in the pronoun's context that can be accessed in order to measure an antecedent candidate's compatibility with the pronoun's context.[7] However, in the case of (di-)transitive verbs, there is always at least one additional verb argument in the context of the pronoun, namely the other complement or adjunct of the pronoun verb.

Here, we again access our similarity metric to determine the compatibility of the antecedent n_i in the pronoun's grammatical function gf_k and the additional argument n_j of the verb governing the pronoun and its grammatical role gf_l, i.e. $sim(n_i, gf_k, n_j, gf_l)$. That is, we take into account the amount of verb nodes in the graph that the antecedent noun n_i connects to with grammatical relation gf_k and to which the additional verb argument n_j connects to with the grammatical role gf_l.

As said, we deem this approach useful in those cases where the selection of the pronoun's verb is broad, i.e. where it allows a diverse set of nouns as argument. For example, consider the verb *machen (to make)*. Almost any noun can be the subject of *machen.* However, if there are direct objects involved, e.g. *Kuchen (cake)* and *Lärm (noise)*, the selection of subjects is narrowed down. For example, *Bäcker (baker)* is an obvious subject for *Kuchen machen (make cake)*, and *Motor* is a likely subject of *Lärm machen (make noise).* However, *Motor* is a very unlikely subject given *Kuchen machen.* In such cases, we would obtain low selectional preference scores regarding the verb, but high and distinctive compatibility scores between the subjects and the direct objects. In our example, we calculate the following compatibilities:

- Verb selectional preferences $comp(noun, machen, subject)$:
    - *comp*(Motor,machen,subject) = 0.1
    - *comp*(Bäcker,machen,subject) = 0, since we have not seen the combination. We identify the seen subject argument of *machen* most similar to *Bäcker*, which is *Hausfrau* with similarity 0.58.

[7] Another category of content words would be adjectives, but adjectives are never used to modify anaphoric pronouns.

We take $comp(\text{Hausfrau,machen,subject}) = 0.48$ and multiply it with the similarity between the two nouns to get the approximated compatibility score:

$$\begin{aligned} &comp(\text{Bäcker,machen,subject}) \approx \\ &\quad comp(\text{Hausfrau,machen,subject}) * sim(\text{Bäcker,Hausfrau,subject}) \\ &\qquad = 0.58 * 0.48 = 0.29 \end{aligned}$$

- Compatibility of the nouns with the direct object $sim(noun, gf_k, arg, gf_l)$:
  - $sim(\text{Bäcker,subject,Kuchen,object}) = 0.36$
  - $sim(\text{Motor,subject,Kuchen,object}) = 0.06$
  - $sim(\text{Bäcker,subject,Lärm,object}) = 0.21$
  - $sim(\text{Motor,subject,Lärm,object}) = 0.11$

To calculate the score of *Bäcker* as the subject of *Kuchen machen*, we thus would take the average ($\varnothing$) of the verb selectional preference and the similarity to the additional argument, which is $\varnothing(0.29, 0.36) = 0.32$. To calculate the score of *Motor* as subject of *Kuchen machen*, we obtain $\varnothing(0.1, 0.06) = 0.08$. Thus, *Bäcker* is a much more likely subject in this context. For the score of *Bäcker* and *Motor* as subjects of *Lärm machen*, we get $\varnothing(0.29, 0.21) = 0.25$ and $\varnothing(0.1, 0.11) = 0.105$. That is, *Bäcker* is deemed the more likely subject in this case, as well. This does not correspond to our initial intuition, but, on the other hand, it is not an implausible outcome. However, *Motor* as the subject of *Kuchen machen* is highly counter-intuitive, and our compatibility scores reflect this.

6.2.4 Ranking antecedent candidates based on the compatibility

In summary, when we compare an antecedent candidate noun to a pronoun's context, we construct the following features based on the graph representation:

- Compatibility of the verb governing the pronoun and the antecedent candidate noun w.r.t. the grammatical function of the pronoun.
- Three features, including the similarity of the verb governing the antecedent candidate and the pronoun verb, that address the sparsity problem of the above feature.
- Similarity between the antecedent candidate noun and the additional verb arguments in cases where the pronoun is governed by a (di-)transitive verb.

To score the antecedent candidates, we simply take the arithmetic mean of the features and select the candidate with the highest mean as antecedent.

Before we evaluate this approach empirically, we introduce an alternative way of calculating the compatibility of an antecedent noun and a pronoun's context in the realm of vector space models.

6.3 Word embeddings as a compatibility framework

In the previous sections, we have outlined an approach to determine an antecedent's compatibility with a pronoun's context that relies directly on the co-occurrence counts of the respective nouns and verbs. Here, we will explore an approach that transforms word co-occurrences into a vector representation in the domain of vector space models. Cosine similarity between the word vectors that represent the antecedent candidate noun and the words in the pronoun's context then serves as a means to determine the compatibility between the antecedent and the pronoun context.

6.3.1 *word2vec* as a framework to derive word embeddings

While there are numerous approaches to construct vector representations of words based on first-order co-occurrences, we make use of *word2vec*[8] (Mikolov et al., 2013c,b,a), a state-of-the-art tool for this task. *word2vec* learns vector representations of words in a vector space model, much like in latent semantic analysis (LSA) (Landauer et al., 1998) and other related approaches. In related vector space models, word vector representations are obtained by constructing vectors with n dimensions, where n denotes the n most frequent nouns in a corpus, for example. The co-occurrence count of each word with these n most frequent words then comprises the vector representation of any given word. Cosine distance between these vector representation can then be used to estimate similarity of words. In LSA, singular value decomposition factorizes the co-occurrence matrix (constructed using the words in the vocabulary as rows and the n most frequent words as columns) to retrieve compressed representations with fewer dimension. Three smaller matrices are constructed, where one represents a clustering of the rows, and one denotes a clustering of the columns. The third matrix describes how these compressed matrices can be combined to retrieve an approximation of the original matrix.

[8] To learn the vectors, we use *word2vec* itself (`https://code.google.com/p/word2vec/`). To integrate the vectors into our Python code, we use *gensim* (`https://radimrehurek.com/gensim/models/word2vec.html`).

The main difference of *word2vec* (and related neural word embeddings) to traditional models, such as LSA, is that the meaning of the vector dimensions are latent to begin with. Unlike LSA, the dimensions cannot be thought of as clusterings of rows and columns of a co-occurrence matrix. The dimensions and their values are mathematical artifacts of the optimization process during training. Instead of taking the n most frequent words, *word2vec* takes n latent dimensions and learns values for them in a deep learning/neural network-inspired fashion. However, compared to related work, the approach removes the need for hidden layers, which makes it fast to train. The basic idea is to learn vector representations of words based on positive and negative examples of context words. Co-occurrence contexts of words are mined from large corpora. Positive examples of context words for a given word w are drawn from a window whose size is defined as one of the model parameters. Negative examples are drawn from outside the co-occurrence window of word w. Given the positive and negative evidence, the objective is then to learn a set of parameters so that the vector denoting the word w, i.e. $\vec{w}$, has a high similarity with the vectors of the context words $\vec{c}$ within the window, and a low similarity with context words outside the window.[9] The word embeddings produced by *word2vec* out of the box have been shown to outperform several traditional word vector models on a variety of tasks (Baroni et al., 2014, inter alia).

6.3.2 Application of word embeddings to pronoun resolution

For our purpose of assessing the compatibility of a given antecedent candidate a_i and a pronoun's context, we use the vector representations of the antecedent head word, i.e. $\vec{a_i}$ and the (relevant) words in the pronoun's context and calculate the average cosine-based similarity $cos\_sim(\cdot,\cdot)$ of the antecedent word vector and the word vectors in the pronoun's context, i.e. the vector of the verb governing the pronoun $\vec{v_j}$ and the vector of its additional argument, $\vec{arg_k}$:

$$comp(a_i, v_j, arg_k) = \varnothing\big(cos\_sim(\vec{a_i}, \vec{v_j}), cos\_sim(\vec{a_i}, \vec{arg_k})\big)$$

However, as stated above, our interest is not to model general word similarity, the typical test task for word embeddings. Our aim is to determine the compatibility of a given

[9] This describes the workings of the skipgram approach using negative sampling on a very basic conceptual level. Discussing the learning algorithm in detail is beyond the scope here. We refer to the original papers cited above and recommend Goldberg and Levy (2014) for an approachable mathematical exposition. Furthermore, Levy and Goldberg (2014b) showed that *word2vec* seems to factorize a PMI matrix internally. Also, Levy et al. (2015) have shown that given the right parameter settings, traditional approaches to word vector estimation, such as PMI and singular value decomposition of the PMI matrices perform on par with *word2vec*. However, *word2vec* achieves state-of-the-art results right out of the box and does not need extensive parameter tuning.

antecedent candidate with a given pronoun context. Therefore, we aim to specifically model the compatibility of an antecedent candidate's head noun and the position of the pronoun in its context. Therefore, we must take into account the grammatical role of the pronoun in its context.

For example, consider a compatibility model that simply takes into account the selectional preferences of the verb governing a pronoun to determine compatibility with the given antecedent candidates. Let us assume two test instances:

(13) Er bereitet den Braten zu.

He prepares the roast.

(14) Der Koch bereitet ihn zu.

The cook prepares him* (it).

We want to resolve each pronoun in turn and we have the antecedent candidate *Koch* for both examples. In the first example (13), *Koch* is a very likely antecedent for *Er* in the subject argument slot of the verb *zubereiten*, because cooks normally prepare food. Thus, we can assume that the vector based similarity of *Koch* and *zubereiten* would be high and we would select *Koch* as antecedent for *Er*.

For the second example (14), however, where we want to resolve *ihn*, which is the direct object of the verb *zubereiten*, we would perform exactly the same query as in the first example given our word vector representation. That is, we would look up the similarity of the *Koch* vector and the *zubereiten* vector, which would be the same as in the first example. We would thus assume that we can select *Koch* as the antecedent of *ihn*. However, *Koch* is a very unlikely direct object of *zubereiten*.

That is, we cannot straight-forwardly apply the *word2vec* approach to assess compatibility of antecedent candidates with a pronoun's context, since the model lacks any notion of grammatical functions.[10] To alleviate this, we perform a simple transformation of the input that is fed into *word2vec*. We concatenate the words (i.e. their lemmas) in a sentence with their grammatical functions, like in the following example:

(15) Original sentence: Der Koch bereitet den Braten zu.

Input sentence: Der$_{det}$ Koch$_{subj}$ zubereiten$_{root}$ Braten$_{obja}$

[10] Note that there are approaches that incorporate syntax into *word2vec*, notably Levy and Goldberg (2014a), or vector representations in general, e.g. Rothenhäusler and Schütze (2009). However, these approaches make use of syntax to identify relevant context words, i.e. syntactic co-arguments for target words, but do not derive vectors for individual grammatical functions that a word occurs with.

Now, we will learn separate vector representations for the subject (e.g. "Koch$_{subj}$") and direct object ("Koch$_{obja}$") instantiations of words, which will enable us to more specifically determine compatibility of antecedent candidates and pronoun contexts.

In our examples above (13,14), a vector representation that is ignorant of grammatical functions would yield maximal similarity for the antecedent candidate "Koch", since the antecedent word itself occurs in the pronoun context as an additional verb argument. Clearly, this high similarity would boost the "Koch" candidate as antecedent. Selecting it would, however, produce a non-sense sentence, i.e. "Der Koch bereitet den Koch zu." Including the grammatical roles in the vector representations lowers the similarity of "Koch" and "Koch" from 1 to "Koch$_{obja}$" and "Koch$_{subj}$", i.e. 0.55.

In comparison to the graph-based approach, the *word2vec* model has the advantage that it is able to calculate similarity between any two words in its vocabulary, since it does not rely on first and second-order co-occurrences directly. Thus, we expect the *word2vec* model to have a broader coverage and applicability. On the other hand, the graph-based model more directly implements the notion of compatibility of verbs and their arguments based on co-occurrence, because it explicitly models the co-occurrences in designated syntactic contexts. Therefore, we expect the graph-based approach to provide high precision.

6.4 Data and preprocessing

To gather verb-argument tuples, we make use of the *SdeWaC* corpus.[11] The corpus is a cleaned version of the *dewac* corpus, which is part of the web-as-corpus initiative[12] (Baroni et al., 2009). The *(S)deWaC* corpora contain sentences crawled from various German websites. In the *SdeWaC* corpus, duplicate sentences from the same domain (URL) have been removed, and sentences were selected that can be parsed with an automatic parser. The *SdeWaC* contains roughly 45 million sentences. We apply the *ParZu* parser[13] (Sennrich et al., 2013) to obtain dependency parses of the sentences.

6.4.1 Construction of the co-occurrence graph

Similar to Scheible et al. (2013), we apply a set of heuristics to extract verb-argument tuples from the parses. For example, we use a rule that recognizes passive voice, and heuristics that identify the main verb in auxiliary constructions etc.

[11] `http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/sdewac.en.html`

[12] `http://wacky.sslmit.unibo.it/`

[13] `https://github.com/rsennrich/parzu`

While the *sdewac* has undergone cleaning, it still contains noise. Thus, we only extract tuples from verbs that occur at least ten times with an identifiable subject. Also, like Wunsch (2010), we exclude the highly frequent verbs *sein* and *haben*. Since their selection of arguments is very broad, we would require a more sophisticated model to represent them adequately. Furthermore, we only extract arguments that are nouns, i.e. that have the PoS tag *NN*, because named entities are often sparse. Additionally, we discard any arguments that we have seen less than three times.

Given these tuples, we construct the graph.[14] We iterate over all verb lemmas and their arguments and add nodes and edges for the grammatical relations subject, direct object, and indirect object. Nodes signify words, i.e. nouns and verbs, and edges between them denote grammatical relations. In addition to the grammatical relation, we add the first-order co-occurrence counts to the edges, which are used to calculate our compatibility and similarity scores. Since there can be multiple edges between a noun and a verb node, denoting different grammatical relations, the graph constitutes an undirected multigraph (without loops, as no node is connected to itself).

After its instantiation, the graph contains a total of 71'460 nodes, of which 8299 are verbs and 63'161 are nouns. It features 2'222'500 edges, of which 1'068'795 denote subject relations, 1'022'623 direct object relations, and 131'082 indirect object relations. Given the initial 45 million sentences, this seems rather small, but recall that we use frequency thresholds for verbs and arguments to be integrated into the graph.

Since we cannot expect to map all encountered nouns in the test set into our resources, we apply compound splitting in order to check whether their heads are represented in the resources.[15] Given an unseen noun, we use this head for the compatibility estimation if it is represented in our resources.

6.4.2 *word2vec* model

For the *word2vec* model, we use the same data basis, i.e. the 45 million sentences from the *sdewac* parsed with *ParZu*. As shown in example 15, we transform the initial raw texts by lemmatizing the words and concatenating them with their grammatical function. We run *word2vec* with vanilla settings and train a skipgram model using negative sampling over three iterations. We set the minimal word count to 50, i.e. words encountered less than 50 times are omitted. This yield a total of 378'511 vectors. 10'784 of these denote verbs, 48'548 signify subjects, 28'608 direct objects, and 6744 indirect objects.

[14] We use *networkx* (`https://networkx.github.io/`) for implementing the graph in Python.

[15] Our procedure for compound splitting and its evaluation is outlined in appendix A.3.

6.5 Evaluation of the models on a word similarity task

In this section, we evaluate our graph-based model of syntactic compatibility and our *word2vec* model of similarity on a word association test set in order to get an estimate of the quality of our models.

For this experiment, we use the German Relatedness Dataset[16] (Gurevych, 2005, Zesch and Gurevych, 2006). The data set provides three different test sets. The first one (Gur65) consists of 65 word pairs. Each pair was assigned a similarity rank between 0 and 4 by human subjects, where 0 denotes fully dissimilar and 4 fully similar. The second (Gur350) and third set (ZG222) contain 350 and 222 such pairs, respectively. The latter two sets are aimed at semantic relatedness, rather than (direct) similarity. The pairs do not only consist of nouns, but words of other part of speeches. We here limit our investigation to pairs of nouns.[17] To assess the performance of a system, Pearson correlation of the system scores and the human judgements is measured.

Our graph-based model of compatibility is aimed at rewarding pairs of nouns with a high score if they display high second-order co-occurrence with verbs given a specific argument slot, i.e. a specific grammatical relation. Test sets of noun associativity and similarity, by contrast, provide gold standards of noun similarity where the semantics of the similarity is generally underspecified, i.e. they are not geared towards a specific semantic relation like hyponymy, meronymy, or synonymy, but subsume different such relations. Our graph model is therefore not necessarily suited to capture these relations and is not aimed at outperforming the state-of-the-art for this task. Still, it is interesting to see how well the model fares, especially compared to the *word2vec*.

Since our models are specific to grammatical relations, we test them using the subject and direct object relation. That is, we calculate how similar a pair of nouns is regarding the subject and the direct object role in our models, respectively.

Table 6.1 gives the results. The top table shows the Pearson correlations of the similarity estimations of our models with human judgements given the grammatical role subject, the lower table indicates correlations given the grammatical relation direct objects. The first column indicates the data set, along with the inter-annotator correlation.[18] We also indicate the correlation of our models (**Graph-W2V**) to see how similarly they judge the pairs. The right part of the tables gives the counts of the pairs where both

[16] `https://www.ukp.tu-darmstadt.de/data/semantic-relatedness/german-relatedness-datasets/`

[17] We identify noun pairs by checking whether both words in a test instance start with an uppercase character.

[18] This can be read as the average pair-wise correlation between the human judgements, although the actual calculation is more complicated, cf. Gurevych (2005).

Data set	**Graph**	**W2V**	**Graph-W2V**	**Appl.**	**NN**	**Total**
Subject relation						
Gur65 (0.81)	0.38	0.76	0.64	53	63	65
Gur350 (0.69)	0.31	0.75	0.48	108	168	350
ZG222 (0.49)	0.23	0.54	0.22	68	118	222
Direct object relation						
Gur65 (0.81)	0.56	0.77	0.69	49	63	65
Gur350 (0.69)	0.46	0.74	0.63	100	168	350
ZG222 (0.49)	0.24	0.53	0.29	65	118	222

Table 6.1: Pearson correlation of similarity/relatedness estimations by our models and human judgements.

our models are applicable (**Appl.**), i.e. for which both models have representations for the nouns, the count of noun pairs (**NN**), and the total count of pairs (**Total**).

We see that for both grammatical relations, the *word2vec* model achieves a higher correlation with the human judgements than our graph model. Still, the graph model yields a positive, moderate correlation with the human judgments. Also, we have to take into account that the inter-annotator correlation for two of the three data sets is rather low, especially for ZG222.

Interestingly, the correlation of the graph-based model increases drastically when the direct object relation is used to determine noun similarity, i.e. from 0.38 to 0.56 for the Gur350 test set. This relates to e.g. Wunsch (2010), who excluded subject-verb relations from his selectional preference model, since he found that verbs hardly feature designated preferences towards their subjects.[19] That is, the results suggest that the direct object relation is a more precise relation than subject when it comes to determining semantic similarity of nouns based on their distributions as arguments of verbs.

The *word2vec* model is not affected by the choice of the grammatical relation, however. This is not too surprising, since the model learns the vector representations based on all context words within a given window, while the graph model only relies on the specific co-occurrences of nouns and verbs given a specific grammatical relation. Since the grammatical relations yield different verb-argument pairs, the similarity judgements differ as well.

Concerning applicability (**Appl.**), we see that similarity judgements relying on the subject relation is slightly higher than for the direct object relation. Note that the applicable count only counts pairs where both models have representations of both nouns in the pair. Thus, the applicability of the individual models is potentially higher.[20] Given the

[19] Cf. section 6.6.

[20] Note further that we have not applied compound splitting in this experiment.

subject relation, our models cover 84% of the pairs in the Gur65 set, 64% of the pairs in the Gur350 set, and 58% of the pairs in the ZG222 set.

Overall, we conclude from this evaluation that both our models are capable of producing similarity judgements that correlate with human judgements. We also see that the models' judgements correlate with each other (**Graph-W2V**), but not perfectly so. Thus, there is potential for complementary use of the models. We next investigate how well the model fare w.r.t. the task they were designed for, i.e. identifying antecedents of pronouns. Before doing so, we discuss related work on selectional preferences of verbs w.r.t. pronoun resolution.

6.6 Related work

In this section, we give an overview of related work on pronoun resolution that incorporated verb selectional preferences and distributional semantics. Also, we briefly discuss related work that employs graph representations for word co-occurrence modeling.

6.6.1 Selectional preferences for pronouns resolution

Using selectional preference of verbs in pronoun resolution dates back to the pioneering era of automated pronoun resolution, and there is a long-standing debate among researchers whether incorporating them is beneficial.

Dagan and Itai (1990) showed that simple co-occurrence frequency can potentially aid pronoun resolution systems. Based on the work of Dagan and Itai, Lappin and Leass (1994) and Dagan et al. (1995) reported an experiment that incorporate a statistical model of selectional preferences of verbs. The selectional preference of a verb v towards and argument noun n was modelled as a conditional probability of seeing the verb given the noun, i.e. $P(v|n) = \frac{|n,v|}{|n|}$. The experiment used the salience-based ranking of the antecedent candidates as calculated by the RAP algorithm and re-ranked the top two candidates. If the selectional preference of the second best candidate exceeded a threshold, it was promoted to first rank and selected as antecedent. This improved the accuracy of the salience-based approach from 86% to 89% accuracy in evaluation. In an evaluation setting where a pronoun was deemed resolved correctly if either the salience-based or the verb selection-based resolution was correct, accuracy improved to 93%. Thus, the approach showed empirically the benefits from incorporating selectional preferences of verbs. However, the test set used in the experiments stemmed from a restricted domain, i.e. computer manuals. In such a domain, there is a specific nomenclature and verb selectional preferences are arguably more easily captured and applied to pronoun resolution than in more open domains, such as newspaper texts. An example given by Lappin and Leass concerned selecting either *message* or *display* as the direct object of the verb *to send*. It is not clear if this approach had benefited the salience-based approach in a more open domain, like newspaper texts. However, Lappin and Leass introduced an experimental protocol which we follow in our own experiments.

Kehler et al. (2004) implemented the approach presented in Dagan et al. (1995) in a pronoun resolution system for newspaper texts. Their findings were less optimistic regarding the utility of incorporating selectional preferences into a state-of-the-art pronoun resolution system. They used a MaxEnt classifier trained with a standard feature set and applied verb selectional preferences either in a postfilter setting, where the top two ranking candidates were re-evaluated, like Dagan et al. (1995), or incorporated the selectional preferences as features for the MaxEnt classifier. In evaluation, Kehler et al. found that neither the postfilter nor the additional features had a substantial impact on performance. In the best configuration, verb selectional preferences improved performance from 76.16% to 76.63% accuracy. On a second test set, the selectional preferences improved performance from 75.72% to 76.77% accuracy when used as a postfilter. Kehler et al. thus concluded that selectional preferences have little to offer given a state-of-the-art pronoun resolution system. They argued that while selectional preference obviously seem applicable to crisp textbook examples, such cases are rarely found in real-world texts. Also, they acknowledged the difficulty of determining when to "trust" the selectional preferences if used as a postfilter. In summary, Kehler et al. raised doubts about the utility of incorporating selectional preferences into a real pronoun resolution systems.

A more optimistic view was presented in Bergsma et al. (2008a). They trained a binary support vector machine classifier using seen verb-argument pairs as positive examples and unseen combinations as negative examples. A separate classifier was trained for each verb that decided whether a given noun was a permissible argument for the verb a hand. Verbs with sparse occurrence were clustered into smaller groups which then received a cluster-level classifier. For each verb classifier, Bergsma et al. constructed a large feature vector that described properties of their arguments, i.e. upper- and lowercasing, digits, person names, semantic class based on a preprocessed clustering etc. Using this large feature set and the verb specific binary classifiers, Bergsma et al. outperformed previous work in a pseudo-disambiguation task, where a system has to choose one of two possible nouns as the correct argument of a given verb, i.e. scoring *article* and *fashion* as direct objects of *to read*. They implemented a baseline using only Mutual Information (MI) which also outperformed related work and which was not outperformed by the SVM-based approach by a large margin. They also performed an evaluation w.r.t. pronoun resolution. Using the MUC corpus, they identified 39 pronouns in direct object position that had an antecedent in the preceding or same sentence as the pronoun. Using their classifiers to identify the correct antecedent, they achieved an accuracy of 38.5%, which was higher than the baseline that selected the most recent noun as antecedent (17.9%). However, this baseline often failed because the faulty antecedent was simply the subject of the verb governing the pronoun in direct object position. Usually, such candidates are not considered as antecedents because they violate binding constraints. Bergsma et al. left it to future work to determine a means to incorporate their approach into a pronoun resolution system. Also, they limited their investigation to pronouns in direct object position.

Wunsch (2010) explored the integration of selectional preferences of verbs into his hybrid pronoun resolution system for German using the kNN classifier TiMBL.[21] Verb selectional preferences were gathered from the TüPP-D/Z corpus, a larger variant of the TüBa-D/Z corpus that contains roughly 12 million sentences from newspaper texts. The TüPP-D/Z was parsed with a dependency parser and verb-argument tuples were extracted using a set of heuristics that identify passive voice etc. Wunsch argued that the extracted subject-verb tuples did not seem to provide strong selectional preferences and thus focused on verb-object tuples. Wunsch performed class abstraction of the extracted direct objects by mapping them to the 22 unique beginners (i.e. the top synsets) of GermaNet, a German word net. For example, from the sentence "Ich fragte die Stewardess (I asked the stewardess)", the tuple *fragen,Mensch* was extracted, since "Stewardess" is a hyponym of the unique beginner *Mensch*. Wunsch then added binary features capturing the compatibility of an antecedent candidate and the verb governing the pronoun to the feature vector used by TiMBL. For example, if the pronoun at hand was governed by *fragen*, as in "Ich fragte sie (I asked her)", and the antecedent candidate was *Stewardess*, the binary feature would be 1, since *fragen* selects nouns that are hyponyms of the unique beginner *Mensch* and *Stewardess* is such a noun. In evaluation, Wunsch found that the inclusion of verb selectional preferences had no impact on resolution performance. However, he did not evaluate performance of the personal pronouns in direct object position separately, at which the preferences were aimed, but measured the overall effect on pronoun resolution.

In our view, the class abstraction of the seen direct objects to the unique beginners in GermaNet is problematic. Class abstraction obscures fine-grained distinctions of preferences for several verbs. Thus, the approach loses its discriminatory power if two antecedent candidates belong to the same GermaNet unique beginner. Consider the *Stewardess* example. For a verb like *fragen*, the class abstraction seems unproblematic. Let us add another noun, e.g. *Räuber (robber)* which is also part of the *Mensch* synset. It is obvious that the two will have very dissimilar distributions w.r.t. to the verbs that

[21] Cf. section 4.4.

select them as direct objects. Given the verb *verhaften (arrest)*, we would like to model that robbers are more likely to be arrested than stewardesses etc. However, mapping arguments to broad semantic classes like word net unique beginners will obfuscate such differences.

Versley (2010) presented an approach to noun coreference resolution for German that made use of selectional preferences of verbs. He calculated PMI statistics of subject-verb and verb-direct object co-occurrences and used them as a feature to determine coreferent bridging between non-string matching nouns (e.g. *the enterprise - the company*).[22] His basic idea was that swapping the verbs that govern the antecedent and the anaphor should not yield large differences in the PMI values. That is, the PMI value for the antecedent and the verb governing the anaphor should be similar to the PMI value of the anaphor and the verb governing it. In turn, the PMI value for the anaphor in the antecedent's verb slot should be roughly the same as that of the antecedent in that slot. This provided an intuitive model of context compatibility and Versley found that it slightly improves Recall from 69.7% to 70.00% in noun coreference resolution. Unfortunately, this approach is not applicable to pronoun resolution, since we cannot meaningfully calculate association strengths of pronouns, which are semantically empty, and verbs. However, we have introduced in the previous sections a compatibility measure that scores pairs of verbs w.r.t. the arguments they share, which is closely related to the approach of Versley.

Our approach using *word2vec* in pronoun resolution is most similar to Klebanov and Peter (2002) who used Latent Semantic Analysis (LSA) to derive word vectors. Klebanov and Peter also compared an antecedent candidate noun to a pronoun's context, i.e. the verb that governs it, and the additional arguments of that verb. In contrast to Klebanov and Peter, we learn vector representations for nouns given grammatical roles, which allows us to distinguish the use of a noun w.r.t. a specific grammatical role, which we argue is important for the task of measuring compatibility of the antecedent and the pronoun's context, as outlined in section 6.3.

6.6.2 Graphs in Computational Linguistics

Graph representations have been used in a variety of tasks in Computational Linguistics, ranging from sentence compression for summarization (Filippova, 2010, Olariu, 2014) to bilingual lexicon extraction Dorow et al. (2009), Laws et al. (2010) (cf. Lahiri (2014) for an overview), because they provide an intuitive formalism to represent co-occurrences of

[22] Cf. section 4.8.

words. To our knowledge, we are the first to explore graphs w.r.t. selectional preferences of verbs and the distributional compatibility of their arguments.

Our approach is most closely related to Dorow (2006) who explored semantic relatedness of nouns in a graph. Dorow used Hearst-style patterns (Hearst, 1992), i.e. conjunctions and listings etc., to derive first-order co-occurrence statistics of nouns. By contrast, we use second-order syntactic co-occurrence of nouns to determine their distributional similarity as verb arguments.

6.7 The distributional compatibility models as postfilters for $n$best candidate re-ranking

In this section, we evaluate the utility of our two models of compatibility of verbs and arguments for the task of pronouns resolution. We follow the protocol introduced in Lappin and Leass (1994) and apply our models as postfilters that re-rank the top two antecedent candidates identified by the salience-based entity-mention model. We saw earlier[23] that in the cases where the entity-mention model fails to rank the correct candidate highest, it ranks the correct candidate as the second best in roughly 80% of the cases for personal pronouns. Thus, focusing the re-ranking on the top two candidates has the potential of substantially reducing errors while limiting the error margin as compared to considering e.g. three or all candidates.

As in our evaluation of classifier performance[24], we first limit the analysis to pronoun instances where the correct antecedent is among the candidates. Furthermore, we restrict evaluation to those cases where the postfilters produce compatibility scores for both candidates and where these scores differ. Otherwise, the postfilters could select the correct candidate simply because the other candidate was not mapped to our resources. Also, it forces the models to assign different scores, i.e. cases were the top two candidates are scored equally are not rewarded.

Since, as we will see, the incremental entity-mention model achieves a higher accuracy than the postfilters, we always chose the candidate it ranks best as antecedent and do not resolve the pronouns to the antecedents proposed by the postfilters. This is relevant for those cases where a resolved pronoun serves as an antecedent candidate for a subsequent pronoun. In such a case, we try to retrieve a nominal mention of the entity that the pronominal antecedent denotes. That is, if the pronominal antecedent is incorrectly resolved, we cannot expect to retrieve an adequate nominal mention.

[23] Cf. section 5.5.3.

[24] Cf. section 5.4.2

An interesting question is whether the two models correct the same set of mistakes made by the salience-based resolution or if their corrections are complementary. To estimate their overlap, we measure the potential improvement by deeming pronouns correctly resolved if either one of the three resolution approaches is correct.

Model	**Postfilter**	**Salience**	**Either**	**Coverage**
		Development set		
Graph	67.61	88.66	93.78	42.96
W2V	60.52	88.89	93.65	59.43
Any	**74.32**	88.71	**94.80**	**61.99**
		Test set		
Graph	63.58	88.42	93.33	45.97
W2V	62.05	88.37	93.37	59.74
Any	**75.41**	88.18	**94.33**	**62.38**

Table 6.2: Evaluation of the graph and *word2vec* models as postfilters.

Table 6.2 shows the results of the re-ranking. Selecting an antecedent randomly achieves a 50% accuracy in this experiment, since we only re-rank the top two candidates. We see that the postfilter based on our models both clearly outperform the 50% margin (**Postfilter** column). Also, postfilter accuracy increases substantially when both models are considered for selecting the correct antecedent (**Any** row, **Postfilter** column). When we combine the salience-based classification and the postfilters and count the cases where either the postfilter or the salience-based antecedent choice is correct, we achieve an accuracy of almost 94% (**Either** column). Furthermore, when we consider all proposed antecedents by the resolution strategies and always select the correct one (if among them), we gain an additional percentage point of performance (**Any** row, **Either** column). These results suggest that our models have the potential of increasing performance of the salience-based antecedent selection by roughly 5 percentage points (**Any** row, moving from **Salience** to **Either** column), which corresponds to an error reduction of 55%.

Obviously, this only applies to the cases where our models are actually applicable. The coverage column shows that the *word2vec* model has better coverage than the graph-based approach. It applies to almost 60% of the pronouns in both data sets, while the graph model covers roughly 45% of the cases. However, the graph model achieves a better accuracy, as indicated in the postfilter column.

The evaluation above only measures the potential of the models w.r.t. the cases where they are applicable. To quantify their potential impact when used in a more realistic setting, we measure the upper bounds in performance given the functional (using the ARCS inferred metric)[25] and pair-wise evaluation of all personal pronoun instances in

[25] Cf. section 3.2.

the respective data sets. Table 6.3 gives the results. **E-M** denotes the entity-mention model and the salience-based antecedent selection, **+Graph** and **+W2V** the respective postfilters, and **+Any** either of them. **Up. Bd.** indicates the upper bounds, i.e. performance when the gold antecedent is selected whenever present.

Functional evaluation

Model	**Rec.**	**Prec.**	**F1**
Development set			
E-M	73.35	67.21	70.15
+Graph	75.84	69.49	72.53
+W2V	76.92	70.54	73.59
+Any	**77.79**	**71.34**	**74.42**
Up. Bd.	86.72	79.82	83.13
Test set			
E-M	75.93	70.97	73.39
+Graph	78.50	73.32	75.82
+W2V	79.22	74.05	76.55
+Any	**80.48**	**75.23**	**77.77**
Up. Bd.	89.38	83.59	86.39

Pair-wise evaluation

Model	**Rec.**	**Prec.**	**F1**
Development set			
E-M	83.35	78.04	80.61
+Graph	85.44	80.00	82.63
+W2V	86.04	80.56	83.21
+Any	**86.93**	**81.39**	**84.07**
Up. Bd.	92.08	86.21	89.05
Test set			
E-M	83.13	77.07	79.98
+Graph	85.67	81.32	83.44
+W2V	86.59	82.20	84.34
+Any	**87.38**	**82.95**	**85.10**
Up. Bd.	92.81	88.10	90.40

Table 6.3: Upper bounds in functional and pair-wise evaluation for incorporating the compatibility models into personal pronoun resolution.

The table shows that the models, in their combination, have the potential to improve performance by 3 to 4 percentage points in all settings. This improvement reduces the distance to the upper bounds (**Up. Bd.**) significantly. In the pair-wise evaluation (right table), this distance is roughly halved w.r.t. F-score. On the test set, the entity-mention model with the salience-based antecedent selection achieves an F-score of 79.98. The verb models show the potential to raise performance to 85.10 F-score, while the upper bound is at 90.40 F-score. In the functional evaluation (left table), the performance difference of the salience-based resolution (**E-M**) to the upper bound is larger than in the pair-wise evaluation. Therefore, the potential improvements by the verb-based models cover less distance w.r.t. the upper bounds. Still, the potential improvements are a substantial step towards reaching the upper bounds.

6.7.1 Learning when to apply the postfilter

The evaluation in the previous section shows that our distributional models of compatibility have the potential to substantially improve the salience-based resolution approach. However, the evaluation setting is unrealistic, since in a real-world application, a strategy is needed to decide which of the approaches to apply for a given pronoun instance and the constellation of its antecedent candidates. We saw that the resolution performance of our compatibility models is below that of the salience-based approach. That is, always selecting the antecedents that these models identify during the re-ranking will harm system performance overall.

A strategy is needed to decide which of the antecedents that are suggested by the models should be picked in the cases where they indicate different ones. In other words, we need a formal criterion to select the appropriate method for each pronoun. Our initial efforts that derived features from the compatibility models and incorporated them into the salience model to rank all candidates did not affect performance significantly. In this regard, our findings align with Kehler et al. (2004) and Wunsch (2010). Therefore, similar to Lappin and Leass (1994), we have opted for the strategy of $n$best re-ranking, i.e. re-ranking the top two antecedent candidates as identified by the salience-based approach.

We have conducted initial experiments with a classifier that is aimed at identifying which model to choose given the verb governing the pronoun and its two top-ranked antecedents. Our main idea is to combine different features that indicate the applicability of the distributional models in the cases that they disagree with the salience-based antecedent selection. Obvious features are the confidence of the models regarding their decisions. Our entity-mention model calculates scores for each candidate and we can access these scores and their differences to assert the classifier's confidence by comparing them. The compatibility models also produce scores which we can access and compare. The task is then to learn thresholds for the confidence measures and their differences in order to decide whether the salience-based antecedent selection should be revised in cases where the verb models disagree with it. This is the approach that Lappin and Leass (1994) and Dagan et al. (1995) applied, although they manually set the confidence thresholds. One problem with this approach is that it assumes that small differences in the salience-based antecedent scoring (indicating weak confidence in the antecedent selection) coincide with large differences in the scores assigned by the verb models (indicating high confidence). Since there is no clear motivation for this assumption, we argue that it is more beneficial to focus on the verb-based models and neglect the scores assigned by the salience model.

One of the features we envision in this direction is aimed at capturing the selectional narrowness of the verb governing the pronoun w.r.t. the grammatical function of the pronoun. For example, we expect the verb *bellen (to bark)* to have a more strict selection regarding its subjects than e.g. *machen (to make)*. The main motivation for investigating this feature is that a narrow selection should correlate with the trustworthiness of the antecedent selection of the compatibility models.

Another feature that we deem helpful in deciding whether to trust the distributional models' antecedents is to determine the similarity of the two antecedent candidates.

We assume that the more dissimilar the two candidates are, the more trustworthy the decision are that the models make.

Since these intuitions require further investigation, we leave it to future work to explore and parametrize them empirically. Given the long-standing debate among researchers about whether incorporating verb semantics into pronoun resolution is a fruitful endeavor, we subscribe to the camp cheering in favor of doing so.

6.8 Chapter summary

This chapter explored the use of the distributional hypothesis to model compatibility of antecedent candidates and a pronoun's context.

We have presented a graph representation of first-order co-occurrence of verbs and arguments, and second-order co-occurrence among arguments. Within this representation, we have defined compatibility metrics and similarity scores that enabled us to address the sparsity problem. We have contrasted the graph model with a state-of-the-art approach to word similarity modeling within distributional semantics, i.e. *word2vec*. We found that the *word2vec* model provides better coverage, i.e. it applies to more pronoun instances, while the graph-based model achieves slightly higher Precision. A combination of both models in an oracle setting further increased performance.

In contrast to related work that used selectional preferences of verbs as a means to semantically represent a pronoun's context, we have included additional verb arguments of the verb governing the pronoun to determine compatibility with an antecedent candidate. We have argued that verb selectional preferences are not always narrow enough to favor one candidate over the other. The inclusion of the additional verb arguments helps to narrow down the selection in cases of (di-)transitive verbs.

A clear benefit of our framework over related work is that the entity-mention model can provide nominal antecedents for pronouns that are themselves antecedents for subsequent pronouns, since the antecedents of resolved pronouns are accessible during the traversal of a document. Related work that processes markables (including pronouns) in a pairwise fashion does not have access to these antecedents. Thus, selectional preferences can only be applied to pronoun instances where the relevant antecedent candidates are all nouns.

Apart from Lappin and Leass (1994), who reported a small accuracy improvement, related work has so far reported mixed or negative results on incorporating verb semantics into pronoun resolution. Kehler et al. (2004) and Wunsch (2010) reported no performance impact when incorporating features denoting selectional preferences into their classifier. Klebanov and Peter (2002) and Bergsma et al. (2008a) showed that their models of verb semantics were able to outperform simple baselines, but did not incorporate their models into real-world pronoun resolution systems.

Although we have not yet found a way to decide when to apply our models, we have shown that they have a large potential to improve performance of a real-world pronoun resolution system which, by itself, reaches state-of-the-art performance. How much of this potential can be harvested in a fully automated setting will have to be determined by future work.

Chapter 7

Conclusions and future work

Underspecification of German pronouns. The main interest of this thesis was to develop a procedure for coreference resolution that addresses the problem of local underspecification of mentions. While underspecification poses a problem in coreference resolution in general, we argued that it is particularly problematic regarding certain German pronouns that feature underspecified morphological properties.

We presented an entity-mention model which efficiently remedies the problem of inconsistent coreference decisions by incrementally disambiguating properties of mentions. Our main hypothesis stated that performance of pronoun resolution for German improves when a consistent solution for the problem of underspecification is devised. We found empirically that the entity-mention model improves performance of pronoun resolution compared to related work which does not address this issue.

Coupled with heuristics to resolve nominal mentions, the incremental entity-mention model achieved new state-of-the-art performance in German coreference and pronoun resolution. Whether our approach of incrementally disambiguating properties of mentions is beneficial for coreference resolution in other languages has to be determined by future work.

Evaluation of coreference and pronoun resolution. We argued that the common evaluation framework for coreference and pronoun resolution is not tailored to the specific requirements of downstream applications. By devising the ARCS metrics, we aimed at developing an evaluation framework that supports the view of prospective downstream applications.

We showed that evaluation of our approach to pronoun resolution yields varying performance levels when different requirements regarding the antecedents are applied, ranging from 90% accuracy of classifiers under idealized settings to 65% F-score when pronouns are required to link to the first mentions of the entities they denote. Such a requirement is not unusual for downstream applications that seek mentions of specific target entities. Thus, pronoun resolution remains a challenging task.

In this light, we encourage future work to investigate the crucial link of pronouns to nominal antecedents, since pronouns are often followed by subsequent pronouns. If the first pronominal mention of an entity is resolved incorrectly, all pronouns linked subsequently to that first pronominal mention denote an incorrect underlying entity and are thus irrelevant from the perspective of downstream applications. We believe that paying attention to this problem will significantly improve the benefit that coreference and pronoun resolution systems provide for downstream applications.

The state-of-the-art in coreference resolution changes rapidly, and progress is often made in small steps. We outlined that evaluation of coreference is affected by a variety of factors. Therefore, it is often not clear why a particular system achieves better performance than another. In an effort to shed light on these differences, we have extended the ARCS framework to accommodate an in-depth comparison of system outputs. This comparison enables an arguably more informative view on the performance differences between systems than the comparison of small changes in averaged F-score. Thus, we encourage researchers to demonstrate in what regard their approach works better compared to related work. Together with the recent approaches on systematic and automated error analysis for coreference, we hope to have provided a tool for this purpose.

Semantics for pronoun resolution. We investigated distributional models that capture the semantic compatibility of antecedent candidates and contexts of pronouns. As an extension to related work, we proposed to take into account the additional arguments of a verb that governs a pronoun to determine compatibility with the antecedent candidates. We showed that the models have the potential of correcting a large amount of erroneous pronoun resolutions of the salience-based antecedent selection. However, we found that devising strategies to successfully integrate the models into the salience-based resolution approach in a real-world setting is difficult. Given the potential of error reduction and the leveling performance of salience-based approaches, we encourage future work to further pursue this direction.

An interesting approach would be to narrow down the set of verbs whose selectional preferences are applicable to pronoun resolution. We argued that not all verbs have a selection preference which is narrow enough to be useful for pronoun resolution. We proposed to address this issue indirectly by requiring dissimilarity between the antecedent candidates in order for the verb selectional preferences to be taken into account. A different approach would be to narrow down the set of verbs that have specific selectional preferences. Furthermore, psycholinguistic research has investigated verbs that promote either their subjects or their direct objects for subsequent mentions by pronouns. For example, constructions like “Peter accused Paul that he [...]” clearly mark the object of the matrix verb as the antecedent for the pronoun in the subordinated clause. Traditional approaches to pronouns resolution would resolve this pronoun incorrectly because salience dictates subject preference and favors parallelism of grammatical roles. Thus, we believe that identifying verbs and constructions that violate these general salience patterns is a fruitful direction.

Appendix A

Implementation details

A.1 Resolution of first person pronouns to third person antecedents

While our approach mainly focuses on third person pronoun resolution, we also accommodate for first person pronouns in the following fashion. During our traversal of the markables, we require pronouns and their antecedents to match regarding their person feature. We do so in order to reduce the number of potential antecedent candidates for a pronoun and ensures coherence of the morphological properties within a coreference chain. That is, we would never resolve a first person pronoun to its third person antecedent, as in e.g. "Peter$_1$ said: 'I$_1$ [...]' ". Therefore, we perform first person pronoun resolution to their third person antecedents as a separate step, after all markables have been processed. While there exist more complex approaches (Almeida et al., 2014, e.g.), we employ the following two heuristics.

We first try to attach first person pronouns on the coreference partition level. That is, we check if there are coreference chains consisting of first person pronouns only. For such a chain, we iterate all markables to find one that occurs at most one sentence before the first mention in the first person pronoun chain, is singular in number, and not neuter in gender (i.e. preferably a person). Also, this markable has to be the subject of a communication verb, like *to say*, which is asserted by a look-up in a list of such verbs. If such a markable is found, we check if it is a member of a coreference chain. If so, this coreference chain is merged with the first person pronoun chain at hand. Otherwise, the markable is prepended to the first person pronoun chain.

The second heuristic tries to find third person antecedents for first person pronouns in the buffer list. Recall that the buffer list contains markables that have not been resolved to an antecedent. If a first person pronoun is found on the buffer list, we again look for an antecedent markable which adheres to the same constraint as in the first heuristic.

For the mention-pair model, we only apply the second heuristic, since the coreference partition is not available after the markable traversal, i.e. before the transitive merge of the found antecedent-anaphor pairs.

A.2 Resolution of nominal markables

In this section, we describe how our system processes noun markables, i.e. markables that feature either a common noun (NN) or named entity (NE) as their syntactic head. While the focus of this thesis is on pronoun resolution, we also resolve noun markables because i) if a non-pronominal antecedent candidate denotes an entity that has already occurred multiple times in a document, it is generally likely to be pronominalized, and ii) we aim to provide a full-fledged coreference resolution system for German.

Resolution of nominal markables can be coarsely divided into two categories. The first category subsumes cases where a nominal mention can be linked to an antecedent based on string matching of the syntactic heads of the markables, as in e.g. [*A company – the company*]. The second category denotes cases where the heads of the coreferring mentions do not match, as in e.g. [*Monsanto – the company*] or [*the book – the novel*]. In machine learning-based approaches to coreference resolution, features targeted at resolving pairs of nominal markables usually encode whether the syntactic heads of two markables match to capture the first category. Soon et al. (2001) showed that the feature that was placed highest in their decision tree was indeed the head match feature. Strube et al. (2002) reported significant performance gains for German coreference resolution when encoding minimum edit distance-based features which measure string similarity for lexeme-based head matching.[1] A set of features captures semantic class compatibility based on e.g. WordNet classes to capture the second category of coreferring nominal mentions.

In our approach, we only consider pairs of nominal markables where the head lemmas of the two markables match, e.g. the first of the aforementioned categories. Resolving nominal pairs with non-matching heads requires complex models of semantic compatibility and relatedness, which is beyond the scope of this thesis. Also, Versley (2010) showed that applying such models in a real end-to-end coreference resolution system for German only marginally affects performance.

[1] Cf. section 4.2

Our approach to noun markable resolution based on head matching follows Versley (2010), who employed hard constraints for the task and does not involve machine learning. We argue that head matching-based resolution of nouns does not involve any discriminatory features for which reasonable weights can be learnt. Consider for example the two markables "a red book" and "the book". Based on the matching heads "book" we might assume that the markables are coreferent. We can extend the second markable by adding an adjective, like "the blue book". Now, clearly the two markables cannot corefer, because their adjectives express exclusiveness. However, we might exchange the adjective with another, like "the nice book". Now, the markable could be coreferent with "a red book" again. That is, the difficulty involved in deciding whether two markables with matching heads are coreferent lies in determining whether their additional arguments are compatible (including adjectives, prepositional phrases, genitive modifiers, relative clauses etc.). However, none of the features used in machine learning approaches in related work addresses this issue. The features generally capture whether the heads of two markables match, which is a binary feature. But the crucial variable is the compatibility of the additional arguments, and applying machine learning to the problem would require a model of this compatibility w.r.t. coreference relations. This model would have to capture e.g. that different color adjectives prevent coreference, but that color adjectives and qualitative adjectives, such as "nice", can license coreference. However, such a model is to date out of reach. Still, there are certain linguistic criteria which can be harvested for the decision on whether two head-matching noun markables should be considered coreferent. We will discuss these criteria and show how they can be turned into constraints.

A.2.1 Constraint-based resolution of nominal markables

We deploy two separate approaches for resolving name markables (i.e. denoting named entities) and noun markables (NPs with a common noun as syntactic head). Potentially coreferring markables are identified based on matching head lemmas. Therefore, it is important to identify relevant heads of multi-word terms. For example, person entities are often introduced in newspaper texts by their full name and are subsequently mentioned only by their last name, i.e. [$Angela\ Merkel - Merkel$]. Therefore, we mark the last name of name markables as the head.

A.2.2 Name markables

For name markables, i.e. NPs with a named entity as their head, we query all previous markables to find one with a string matching head. If the potentially anaphoric markable at hand is a single-word term, e.g. *Berlin*, and we find a matching antecedent, we link the two markables. We also do so if the markable at hand is a multi-word term but does not denote a person. If the markable is a multi-word term and denotes a person, we check if we find a first name in the potential anaphor and its potential antecedent using a list of first names. If we find first names in both markables, they have to be identical. This prevents us from linking e.g. [*Patrick Wagner – Barbara Wagner*]. In newspaper texts, persons are often mentioned by their last name. In other domains, however, persons also frequently occur with their first name. This is the case for fictional texts, as well as e.g. mountaineering reports, as in the corpus *text+berg* (Volk et al., 2010), where we often encounter pairs like *[Bergführer Peter Taugwalder - Peter] ([mountain guide Peter Taugwalder - Peter]).* To accommodate for such pairs, additional heuristics can be added to our approach.

Additionally, we accumulate and match nominal descriptors of name markables in the following fashion. We store noun predications in copulas where a name markable is the subject. For example, in the copula construction "Vita B. ist die Siegerin [...] (Vita B. is the winner [...])", we extract "Siegerin" as a nominal description of the [*Vita B.*] markable. The entity is later mentioned as "die ehemalige Gewinnerin (the former winner)". Using the extracted predication from the copula, we are able to detect the common noun mention of the named entity. We also capture appositions in similar fashion. In a markable like [*Kanzlerin Angela Merkel*], we identify *Merkel* as the head for string matching. However, we would miss subsequent noun mentions of the entity, i.e. [*die Kanzlerin*]. Therefore, we store nouns found in name markables as nominal descriptors and allow subsequent noun markables to link to them. This way, we are able to identify coreference between [*Kanzlerin Angela Merkel – die Kanzlerin*], although the markable heads do not match. Analogously, we identify nominal descriptors in the reversed construction, i.e. *Angela Merkel*, *die Kanzlerin* and make them available for string matching. Furthermore, we allow for partial (substring) matches if the string of the anaphoric markables is at the end of the antecedent string. Doing so, we are able to capture coreference between e.g. [*EU – Umweltkommissarin Ritt Bjerregaard*] ([*EU environment commissioner Ritt Bjerregaard*]) and [*die Umweltkommissarin*] ([*the environment commissioner*).

A.2.3 Noun markables

A major problem regarding the resolution of noun markables (i.e. markables whose syntactic head consists of a common noun, like [*the company*]) lies in deciding which noun markables should be considered to be anaphoric. This is known as the anaphoricity detection problem. The majority of noun phrases in newspaper articles are not anaphoric

(Recasens et al., 2013, inter alia) and thus we need strategies to determine the anaphoricity of markables before attempting to resolve them.

There are two main branches in the research that addresses this problem explicitly.[2] One branch applies machine learning to the problem. A classifier is learnt that decides for every noun markable whether an antecedent should be sought (Recasens et al., 2013, inter alia). Coreference resolution is then only applied to those markables that the classifier has deemed to be anaphoric. The other branch develops heuristics based on linguistic constraints. For example, indefinite NPs (e.g. [*a company*]) are very unlikely to be anaphoric. Therefore, most of these approaches evolve around definite NPs only. Because the literature on this topic is vast and our focus lies on pronoun resolution, exploring the problem in detail is beyond the scope of this thesis. We point to the dissertation of Versley (2010) for a thorough discussion.

Our approach to resolving noun markables relies on said linguistic heuristics to determine pairs of head-matching, coreferring markables. We basically employ the same strategy as for name markables, but apply more constraints for matching noun markables to each other.

We loop through the markables and check for each markable headed by a common noun whether we find a head-matching noun markable. If so, we apply the following filters that have to be passed in order for the pair to be processed further:

- The potential anaphoric noun markable has to be at least two tokens long. This removes e.g. bare plural markables, such as [*people*], which are rarely anaphoric.
- Antecedent and anaphor have to match in their morphological properties, i.e. gender and number.
- The potential anaphoric noun markable cannot be indefinite ([*a company*]), all-quantified ([*all companies*]), or negated ([*no company*]).

Then, we impose the following heuristics for establishing coreference:

- The antecedent is indefinite, the anaphor definite. This captures typical patterns of entities being introduced and subsequently mentioned in discourse, i.e. [*a company*–*the company*]. Note that we here do not apply constraints on potential modifiers in both antecedent and anaphor.

[2] Note that most coreference resolution system perform anaphoricity detection implicitly. That is, an antecedent is sought for every (definite) nominal markable. If one is found, the markable becomes anaphoric, else it is deemed non-anaphoric.

- If there are modifiers in the antecedent, i.e. [*the successful and growing company*], all modifiers in the anaphor have to be contained in the antecedent. This enables us to match e.g. [*the company*] or [*the successful company*], but prevents us from matching [*the troubled company*] etc. Note that we here do not apply constraints on the determiners. However, indefinite anaphors are filtered beforehand (see above).

- If the head token in the antecedent markables is not the last token in the respective NP (i.e. the head is followed by a PP), we require at least 60% of the tokens in the anaphor to match. This allows us to match [*the king of France*] to [*the king*], but prevents the match to [*the king of England*].

- Finally, if there are modifiers in the antecedent, but none in the anaphor, we allow the match, i.e. [*the elderly man with the funny hat*] – [*the man*].

Additionally, we add two heuristics, one for processing hyphenated words and one for resolving remaining noun markables featuring a demonstrative determiner. For noun markables with hyphenated head words, we remove the left-hand side and check if we find a markable whose head string matches the right-hand side. These two markables are then processed by the filter batch outlined above, like all other markables. This allows us to match e.g. [*die debis – Mitarbeiter*] ([*the debis – employees*]) to [*die Mitarbeiter*] ([*the employees*]).

For unresolved noun markables with a demonstrative determiner, we also allow the resolution to an antecedent with substring matching at string end. Note that we do not generally allow substring matching, since it yields many false positives. In this restricted setting, where anaphoricity is indicated by the demonstrative determiner, we are more safe in allowing for it. Doing so, we are able to match e.g. [*diese Reise*] ([*this trip*]) to [*Bildungsreise*] ([*educational trip*]). That is, we are able to identify heads of compounds without the need for linguistically guided, proper compound splitting as in Versley (2010).

A.2.4 Evaluation

We apply the ARCS inferred antecedents metric[3] to the nominal markables, which requires mentions to link to correct nominal antecedents and thus directly captures the performance of our approach in a pair-wise fashion. Table A.1 shows the results on our test set. Note that the results subsume both noun (i.e. common nouns) and name (named entities) markables, since ARCS does not distinguish between the two.

[3] Cf. section 3.2

Rec	**Prec**	**F1**	**Acc.**	**True Pos.**	**Wrong Link.**	**False Neg.**	**False Pos.**	**Gold mentions**
58.28	63.09	60.59	90.02	4809	533	2909	2280	8251

Table A.1: Evaluation of nominal markable resolution (nouns and named entities).

We see that we achieve higher Precision than Recall, which is represented in the higher false negative (FN) count compared to the false positive (FP) count. Accuracy (Acc.), which only evaluates gold mentions (i.e. $\frac{TP}{TP+WL}$), indicates that our matching strategy is solid, i.e. when we resolve gold mentions, we find the correct antecedent in 90.02% of the cases. The comparison of accuracy to the F1-score shows that the problem in resolving definite NPs mainly lies in determining which NPs to resolve. Once this problem is solved, our matching strategy achieves high resolution accuracy. However, in a real-world setting, the referring mentions are not known, thus the measure is unrealistic in that regard.

Our F-score lies within the range of results for same-head resolution heuristics reported by (Versley, 2010, 56.6%-66.2%), who used the first 125 articles of an earlier version of the TüBa-D/Z as a test set. We thereby conclude that our approach achieves reasonable results.

A.3 Character ngram-based splitting of sparse compound nouns

German features compound nouns which can be constructed rather freely. Therefore, we are likely to encounter compounds that are not in our resources. In these cases, we want to test whether the head noun of the compound is represented in our resources. While there exist several compound splitters for German, we briefly introduce our own character ngram-based splitter which has several advantages over related work, at least for our purposes.

The main idea behind our approach is that different character ngrams in (German) words have a likelihood of indicating a word start, a word middle, and a word end. Loosely adapting nomenclature from morphology, we call these prefix, infix, and suffix ngrams, respectively. A good position for a split of a compound is located where we encounter a low likelihood of a word middle, preceded by a character sequence with a high likelihood of indicating a word ending, followed by a sequence that has a high likelihood of indicating a word start.

We devise simple probabilities to calculate the likelihood of character ngrams to indicate each of the ngram types (prefix, infix, suffix), which can be derived from a corpus. Given a noun, we start at the first character and extract all ngrams of size $3 < size < n$ that start at the word beginning as prefix ngrams. Analogously, we extract all ngrams of size $3 < size < n$ that end at the word end as suffix ngrams. For the infix ngrams, we cut the first and last character of the word and shift all ngrams of size $3 < size < n$ over the character sequence, re-starting at each character in the sequence. For example, we extract the following ngrams for the word "Dosenöffner":

prefix	Dos, Dose, Dosen, Dosen, Dosenö, Dosenöf, ...
infix	ose, osen, osenö, ..., sen, senö, senöf, ..., öff, öffn, öffne, ..., ffn, ffne
suffix	ner, fner, ffner, öffner, ...

Table A.2: Examples of character ngrams extracted from the word "Dosenöffner" using our approach.

Given the extracted ngrams and their counts as prefix, infix, and suffix occurrences, we calculate for each ngram the conditional probability of seeing each ngram type. That is, the conditional probability of seeing a word beginning given a specific ngram n is given by $P(prefix|n) = \frac{|prefix,n|}{|n|}$ etc. Table A.3 shows the 10 most likely and the 10 most unlikely character sequences per ngram type obtained over 20 million noun instances in the *sdewac*.

Type	**likely**	**unlikely**
prefix	bundesr, bundesregier, bundesregie, bundesregi, bundesregierun, menschenrech, menschenrec, vorj, vorjah, vorja	chul, chw, chs, nsc, lle, nne, ozen, mme, roze, nst
infix	ieru, ierun, itä, roze, gieru, gierun, egieru, egierun, nsc, hrun	diskussion, dollar, niederlage, beispiel, vorjahr, situation, mio., oktober, januar, samstag
suffix	itpunkt, eitpunkt, itraum, eitraum, skanzlerin, undeskanzlerin, ndeskanzlerin, eskanzlerin, deskanzlerin, enagentur	gierun, gieru, oze, kre, itä, gru, ierun, ieru, roze, nsc

Table A.3: Likely and unlikely character ngrams at word beginnings (prefix), within words (infix), and word endings (suffix).

When splitting a word, we again shift the ngram window of size $3 < size < n$ over the word, where n is the length of the word minus 3 characters. We choose 3 characters, since that is the minimal size of a word that we consider as the head or tail of a compound. Our aim is to identify the character position within the compound where

1. the character sequence leading up to the position has a high likelihood of denoting a word ending,
2. the character sequence following the position has a high likelihood of a word beginning,
3. the infix probability is low.

Starting at the forth character of the compound under scrutiny, we check how likely it is to see the character sequence before the forth character as a word ending (1). To do so, we collect the suffix probability of all ngrams that end at the current position and take the probability of the most likely ngram.[4] Next, we collect the probability of all prefix ngrams that occur after the current position (2). Again, we take the probability of the most likely ngram. Finally, we gather all infix ngrams that start at the current position (3). Here, we take the least likely infix ngram, because a low infix ngram probability indicates a good position for a split. To calculate a score for splitting the compound at the given position n, we use the following formula which implements the idea outlined in the enumeration above, i.e.:

$$score(n) = \max p(prefix) + \max p(suffix) - \min p(infix) \tag{A.1}$$

All position within the compound are scored, and the highest scored position is taken as the split location.

To evaluate our approach, we use the set of compounds extracted from GermaNet that have manual annotation of splits (Henrich and Hinrichs, 2011).[5] The set is comprised of 54’572 compounds. Note that the purpose of our splitter is to determine the head of a compound in order to map it into our resources, i.e. the graph and the *word2vec* model, in cases where the compound is not represented in them. Therefore, we are only interested in whether our splitter correctly identifies the head of the compound and are not concerned with the lemmatization of the left part of the compound. Thus, we count a split as correct is the head matches the head in the test set. We compare our splitter

[4] We check heuristically if there is a *Fugen-S* with a regular expression. If we find one, we remove it, because *Fugen-S* do not occur at word endings, but only in compounds.

[5] `http://www.sfs.uni-tuebingen.de/lsd/compounds.shtml`

to three available tools that feature compound splitting. GERTWOL (Haapalainen and Majorin, 1995), a rule-driven morphological analyzer for German, produces word boundaries in its output and is a widely used tool within the German Computational Linguistics and NLP community. The statistical machine translation pipeline *Moses* (Koehn et al., 2007) also features a compound splitting mechanism. It relies on counts of how often the potential parts of a compounds are seen individually within a large corpus. The IMS splitter[6] expands on the splitter integrated in *Moses* by incorporating lemmas and PoS tags in the analysis. This tool can also be forced to split compounds in the cases where it would normally not do so. We also include this version (IMS F.), since all nouns in our test set are compounds. We chose these splitters for our evaluation, since they are available.[7] Table A.4 shows the results.

System	**Accuracy all**	**Accuracy selected**
MOSES	15.83	91.01 (17.39)
IMS	48.47	94.99 (50.80)
IMS F.	84.86	90.54 (92.49)
GERTWOL	88.80	**99.99** (88.80)
OUR APPR.	**95.22**	95.22 (**100.00**)

Table A.4: Compound splitting accuracy of different systems.

We list the accuracy (correctly identified head of the compound) given all compounds (**Accuracy all**) and the accuracy given only those compounds that are actually split by the systems (**Accuracy selected**). The table shows that our approach achieves the highest accuracy given all compounds. Performance in this category is especially low for IMS and MOSES, since these systems heavily rely on the corpus that is used during training. For training, we use the same data for all systems, i.e. the nouns extracted from *sdewac*. Thus, the IMS and MOSES system will have seen many of the compounds in the training data and will thus not split them in the test set. We see this behaviour given the percentages of split compounds in parentheses in the **Accuracy selected** column. The MOSES splitter only splits 17.39% of the compounds in the test set and achieves and accuracy of 91.01% for these. The IMS splitter splits more of the compounds (50.08%) and achieves a higher accuracy than MOSES. Interestingly, the GERTWOL system splits with almost perfect accuracy given the cases that it splits. However, for roughly 10% of the compounds in the test set it does not identify a word boundary.

The evaluation shows that our simple character ngram approach without linguistic knowledge performs surprisingly well for compound splitting. It outperforms the other statistical splitters and has a better coverage than GERTWOL. Henrich and Hinrichs

[6] `http://www.ims.uni-stuttgart.de/institut/mitarbeiter/wellermn/tools.html`

[7] Note that GERTWOL has a commercial licence.

(2011) also compared various splitters on this data set and measured accuracy for identifying correct split positions. The best performing approach (the combined hybrid compound splitter, CH-CS) achieved an accuracy of 94.83%. It combined the output of the other splitters in a voting scheme and incorporated knowledge on derivation morphology. It thus featured a high degree of complexity compared to our approach.

A drawback of this evaluation is that we cannot measure the identification of compounds, since all words in the test set are actually compounds.[8] We thus cannot assess how well our approach identifies compounds. The splitter outputs the scores (as given by equation A.1) for its splits. A score above 0 indicates a likely split, a score below 0 an unlikely split. Thus, the score can be used as a confidence measure. For example, for the word "Dosenöffner", the splitter outputs the following analysis:

0.31	Dosen	Öffner
-1.41	Dosenöff	Ner
-1.48	Dose	Nöffner
...	...	...

The splitter identifies the most probable position for a split. However, the score for making the split is not very high (0.31), compared to e.g. the analysis of "Autobahnraststätte":

0.79	Autobahn	Raststätte
-0.55	Auto	Bahnraststätte
-0.72	Autobahnrast	Stätte
...	...	...

For non-compounds, the scores for splitting are low, e.g. for "Beamter", we get:

-0.43	Beam	Ter
-1.01	Bea	Mter

Given the scores, we could evaluate the splitter's performance on compound identification by only considering splits above a certain score, e.g. 0. However, the intended use for our splitter is to attempt to map unseen compounds into our resources. We only apply the splitter in these cases and are thus not too worried about splitting non-compounds.

[8] The data set used in Escartín (2014) is not available, unfortunately.

Bibliography

Almeida, M. S. C., Almeida, M. B., and Martins, A. F. T. (2014). A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48, Gothenburg.

Aone, C. and Bennett, S. W. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 122–129, Massachusetts.

Ariel, M. (1988). Referring and accessibility. *Journal of linguistics*, 24(01):65–87.

Bagga, A. and Baldwin, B. (1998). Algorithms for Scoring Coreference Chains. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 563–566, Granada.

Baldwin, B. (1997). CogNIAC: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45, Madrid.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247, Baltimore.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Bergsma, S., Lin, D., and Goebel, R. (2008a). Discriminative Learning of Selectional Preference from Unlabeled Text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 59–68, Honolulu.

Bergsma, S., Lin, D., and Goebel, R. (2008b). Distributional Identification of Non-Referential Pronouns. In *ACL-08: HLT*, pages 10–18, Columbus.

Björkelund, A., Eckart, K., Riester, A., Schauffler, N., and Schweitzer, K. (2014). The Extended DIRNDL Corpus as a Resource for Coreference and Bridging Resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik.

Björkelund, A. and Farkas, R. (2012). Data-driven Multilingual Coreference Resolution Using Resolver Stacking. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 49–55, Jeju.

Björkelund, A. and Kuhn, J. (2014). Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47–57, Baltimore.

Broscheit, S., Ponzetto, S. P., Versley, Y., and Poesio, M. (2010). Extending BART to Provide a Coreference Resolution System for German. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 164–167, Valletta.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

Cai, J. and Strube, M. (2010a). End-to-End Coreference Resolution via Hypergraph Partitioning. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 143–151, Beijing.

Cai, J. and Strube, M. (2010b). Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 28–36, Tokyo.

Cardie, C., Wagstaff, K., and Others (1999). Noun phrase coreference as clustering. In *Proceedings of the Joint Sigdat Conference on empirical methods in natural language processing and very large corpora*, pages 82–89, Maryland.

Chen, C. and Ng, V. (2013). Linguistically Aware Coreference Evaluation Metrics. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1366–1374, Nagoya.

Culotta, A., Wick, M., Hall, R., and McCallum, A. (2007). First-order probabilistic models for coreference resolution. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 81–88, Rochester.

Dagan, I. and Itai, A. (1990). Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th conference on Computational linguistics*, pages 330–332, Helsinki.

Dagan, I., Justeson, J., Lappin, S., Leass, H., and Ribak, A. (1995). Syntax and Lexical Statistics in Anaphora Resolution. *Applied Artificial Intelligence*, 9:633–644.

Dalton, J., Blanco, R., and Mika, P. (2011). Coreference aware web object retrieval. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 211–220, Glasgow.

Daumé, H., Marcu, D., Daumé III, H., and Marcu, D. (2005). A Large-scale Exploration of Effective Global Features for a Joint Entity Detection and Tracking Model. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 97–104, Vancouver.

De Beaugrande, R. and Dressler, W. U. (1981). *Einführung in die Textlinguistik*, volume 28. Niemeyer Tübingen.

Denis, P. and Baldridge, J. (2007). A Ranking Approach to Pronoun Resolution. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1588–1593, San Francisco.

Denis, P. and Baldridge, J. (2008). Specialized Models and Ranking for Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 660–669, Honolulu.

Denis, P. and Baldridge, J. (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42(1):87–96.

Dorow, B. (2006). *A graph model for words and their meanings*. PhD thesis, University of Stuttgart.

Dorow, B., Laws, F., and Michelbacher, L. (2009). A graph-theoretic algorithm for automatic extension of translation lexicons. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 91–95, Athens.

Durrett, G. and Klein, D. (2013). Easy Victories and Uphill Battles in Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1971–1981, Seattle.

Erk, K., Padó, S., and Padó, U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4).

Escartín, C. P. (2014). Chasing the Perfect Splitter: A Comparison of Different Compound Splitting Tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3340–3347, Reykjavik.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, University of Stuttgart.

Faruqui, M. and Padó, S. (2010). Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of KONVENS 2010*, Saarbrücken.

Fernandes, E. R., dos Santos, C. N., and Milidiú, R. L. (2012). Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48, Jeju.

Fernandes, E. R., dos Santos, C. N., and Milidiú, R. L. (2014). Latent Trees for Coreference Resolution. *Computational Linguistics*, 40(4):801–835.

Filippova, K. (2010). Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 322–330, Beijing.

Finkel, J. R. and Manning, C. D. (2008). Enforcing transitivity in coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pages 45–48, Columbus.

Ge, N., Hale, J., and Charniak, E. (1998). A Statistical Approach to Anaphora Resolution. In *In Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170, Montreal.

Goldberg, Y. and Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

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

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pages 767–778, Jeju Island.

Haapalainen, M. and Majorin, A. (1995). GERTWOL und morphologische disambiguierung für das Deutsche. In *Proceedings of the 10th Nordic Conference on Computational Linguistics*, Helsinki.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English.* Longman.

Hamp, B. and Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid.

Hartrumpf, S. (2001). Coreference Resolution with Syntactico-Semantic Rules and Corpus Statistics. In *Proceedings of the Fifth Computational Natural Language Learning Workshop*, pages 137–144, Toulouse.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th Conference on Computational Linguistics*, pages 539–545, Nantes.

Henrich, V. and Hinrichs, E. W. (2011). Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of Recent Advances in Natural Language Processing*, pages 420–426, Hissar.

Hinrichs, E., Filippova, K., and Wunsch, H. (2005). What Treebanks Can Do For You: Rule-based and Machine-learning Approaches to Anaphora Resolution in German. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*, pages 77–88, Barcelona.

Hinrichs, E. W., Filippova, K., and Wunsch, H. (2007). A Data-driven Approach to Pronominal Anaphora Resolution in German. In *Proceedings of the 5th International Conference on Recent Advances in Natural Language Processing*, Borovets.

Hirshman, L. and Chinchor, N. (1998). MUC-7 coreference task definition. In *Proceedings of the Seventh Message Understanding Conference (MUC-7).*

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, 44(4):311–338.

Holen, G. I. G. (2013). Critical Reflections on Evaluation Practices in Coreference Resolution. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 1–7, Atlanta.

Kehler, A., Appelt, D., Taylor, L., and Simma, A. (2004). The (Non) Utility of Predicate-Argument Frequencies for Pronoun Interpretation. In *Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 289–296, Boston.

Kim, J.-D., Nguyen, N., Wang, Y., Tsujii, J., Takagi, T., and Yonezawa, A. (2012). The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. *BMC bioinformatics*, 13(11):1–12.

Klebanov, B. and Peter, W.-H. (2002). The Role of Wor(l)d Knowledge In Pronominal Anaphora Resolution. In *Proceedings of the International Symposium on Reference Resolution for NLP*, Alicante.

Klenner, M. and Ailloud, É. (2008). Enhancing coreference clustering. In *Proceedings of the Second Bergen Workshop on Anaphora Resolution (WAR II)*, pages 31–40, Bergen.

Klenner, M. and Ailloud, É. (2009). Optimization in Coreference Resolution is Not Needed: A Nearly-optimal Algorithm with Intensional Constraints. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 442–450, Athens.

Klenner, M. and Tuggener, D. (2010). Inkrementelle salienzbasierte Koreferenzanalyse für das Deutsche. In *Proceedings of the 10th Edition of the KONVENS Conference*, pages 37–46, Saarbrücken.

Klenner, M. and Tuggener, D. (2011a). An Incremental Entity-Mention Model for Coreference Resolution with Restrictive Antecedent Accessibility. In *Recent Advances in Natural Language Processing*, pages 178–185, Hissar.

Klenner, M. and Tuggener, D. (2011b). An Incremental Model for Coreference Resolution with Restrictive Antecedent Accessibility. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 81–85, Portland.

Klenner, M., Tuggener, D., Fahrni, A., and Sennrich, R. (2010). Anaphora Resolution with Real Preprocessing. *Lecture Notes in Computer Science*, 6233:215–225.

Kobdani, H. and Schütze, H. (2010). SUCRE: A Modular System for Coreference Resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 92–95, Los Angeles.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., and Others (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180, Prague.

Kouchnir, B. (2004). A machine learning approach to German pronoun resolution. In *Proceedings of the ACL 2004 Workshop on Student Research*, Barcelona.

Kübler, S. and Zhekova, D. (2011). Singletons and Coreference Resolution Evaluation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 261–267, Hissar.

Kummerfeld, J. J. K. and Klein, D. (2013). Error-Driven Analysis of Challenges in Coreference Resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277, Seattle.

Lahiri, S. (2014). Complexity of Word Collocation Networks: A Preliminary Structural Analysis. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–105, Gothenburg.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Lappin, S. and Leass, H. J. (1994). An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20:535–561.

Laws, F., Michelbacher, L., and Dorow, B. (2010). A linguistically grounded graph model for bilingual lexicon extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 614–622, Beijing.

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).

Levy, O. and Goldberg, Y. (2014a). Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 302–308, Baltimore.

Levy, O. and Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185, Montreal.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Linke, A., Nussbaumer, M., Portmann, P. R., Willi, U., and Berchtold, S. (2004). *Studienbuch Linguistik*. Niemeyer Tübingen.

Luo, X. (2005). On Coreference Resolution Performance Metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., and Roukos, S. (2004). A Mention-synchronous Coreference Resolution Algorithm Based on the Bell Tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona.

Màrquez, L., Recasens, M., and Sapena, E. (2012). Coreference resolution: an empirical study based on SemEval-2010 shared Task 1. *Language Resources and Evaluation*, 47(3):661–694.

Martschat, S. and Strube, M. (2014). Recall error analysis for coreference resolution. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2070–2081, Doha.

Martschat, S. and Strube, M. (2015). Latent Structures for Coreference Resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418.

McCarthy, J. and Lehnert, W. (1995). Using decision trees for coreference resolution. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligenc*, pages 1050–1055, Montreal.

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., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, Lake Tahoe.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). 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, Atlanta.

Mitkov, R. (1998). Robust Pronoun Resolution with Limited Knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 869–875, Morristown.

Mitkov, R. (2001). Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. *Applied Artificial Intelligence*, 15(3):253–276.

Müller, M.-C. (2008). *Fully Automatic Resolution of 'it', 'this', and 'that' in Unrestricted Multi-Party Dialog*. PhD thesis, University of Tübingen.

Müller, M.-C. and Strube, M. (2001). Annotating Anaphoric and Bridging Relations with MMAX. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–6, Aalborg.

Naumann, K. (2007). Manual for the annotation of in-document referential relations. Technical report, University of Tübingen.

Ng, V. (2010). Supervised Noun Phrase Coreference Research: The First Fifteen Years. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1396–1411, Uppsala.

Ng, V. and Cardie, C. (2002a). Combining Sample Selection and Error-driven Pruning for Machine Learning of Coreference Rules. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, pages 55–62, Pennsylvania.

Ng, V. and Cardie, C. (2002b). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111, Pennsylvania.

Nicolae, C. and Nicolae, G. (2006). BestCut: A Graph Algorithm for Coreference Resolution. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 275–283, Sydney.

Nicolov, N., Salvetti, F., and Ivanova, S. (2008). Sentiment analysis: Does coreference matter. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, pages 37–41, Aberdeen.

Olariu, A. (2014). Efficient Online Summarization of Microblogging Streams. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 236–240, Gothenburg.

Poesio, M. (2004). The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162, Cambridge.

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., and Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 30–35, Baltimore.

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., Björkelund, A., Uryupina, O., Zhang, Y., and Zhong, Z. (2013). Towards Robust Linguistic Analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning*, Jeju.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 1–27, Portland.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. D. (2010). A Multi-Pass Sieve for Coreference Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Massachusetts.

Rahman, A. and Ng, V. (2009). Supervised Models for Coreference Resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 968–977, Edinburgh.

Recasens, M. and Hovy, E. (2010). BLANC : Implementing the Rand index for coreference evaluation. *Journal Natural Language Engineering*, 17(4):485–510.

Recasens, M., Marneffe, M. D., 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, Atlanta.

Recasens, M., Martí, T., Taulé, M., Llu, M. R., Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Los Angeles.

Rösiger, I. and Riester, A. (2015). Using prosodic annotations to improve coreference resolution of spoken text. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 83–88, Beijing.

Rothenhäusler, K. and Schütze, H. (2009). Unsupervised Classification with Dependency Based Word Spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 17–24, Athens.

Scheible, S., Im Walde, S. S., Weller, M., and Kisselew, M. (2013). A compact but linguistically detailed database for German verb subcategorisation relying on dependency parses from Web corpora: Tool, guidelines and resource. In *Proceedings of the 8th Web as Corpus Workshop*, pages 63–72, Lancaster.

Schiehlen, M. (2004). Optimizing Algorithms for Pronoun Resolution. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 390–396, Geneva.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, IMS Stuttgart.

Schneider, G. (2008). *Hybrid long-distance functional dependency parsing.* PhD thesis, University of Zurich.

Schulte im Walde, S. (2010). Comparing Computational Models of Selectional Preferences - Second-order Co-Occurrence vs. Latent Semantic Clusters. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 1381–1388, Valletta, Malta.

Sennrich, R., Volk, M., and Schneider, G. (2013). Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 601–609, Hissar.

Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An Annotation Scheme for Free Word Order Languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 88–95, Washington.

Soon, W. M., Ng, H. T., and Daniel (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, pages 96–102, Barcelona.

Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. (2009). Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664, Singapore.

Strube, M. and Hahn, U. (1999). Functional Centering: Grounding Referential Coherence in Information Structure. *Computional Linguistics*, 25(3):309–344.

Strube, M., Rapp, S., and Müller, C. (2002). The Influence of Minimum Edit Distance on Reference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 312–319, Philadelphia.

Stuckardt, R. (2001). Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm. *Computational Linguistics*, 27(4):479–506.

Telljohann, H., Hinrichs, E., and Kübler, S. (2004). The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 2229–2232, Lisbon.

Tuggener, D. (2014). Coreference Resolution Evaluation for Higher Level Applications. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 231–235, Gothenburg.

Tuggener, D. and Klenner, M. (2012). Non-negative Matrix Factorisation-based Verb Semantics for 3rd Person Pronoun Resolution. In *Proceedings of the 2012 IEEE Sixth International Conference on Semantic Computing*, pages 250–254, Palermo.

Tuggener, D. and Klenner, M. (2014). A Hybrid Entity-Mention Pronoun Resolution Model for German Using Markov Logic Networks. In *Proceedings of the 12th Edition of the KONVENS Conference*, pages 21–29, Hildesheim.

Turney, P. D., Pantel, P., and Others (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Uryupina, O. (2007). *Knowledge acquisition for coreference resolution*. PhD thesis, University of Saarland.

Uryupina, O. (2008). Error Analysis for Learning-based Coreference Resolution. In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference*, pages 1914–1919, Marrakech.

Van De Cruys, T. (2011). Two Multivariate Generalizations of Pointwise Mutual Information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 16–20, Portland.

Versley, Y. (2006). Disagreement dissected: Vagueness as a source of ambiguity in nominal (co-) reference. In *Ambiguity in Anaphora Workshop Proceedings*, pages 83–89, Malaga.

Versley, Y. (2010). *Resolving Coreferent Bridging in German Newspaper Text*. PhD thesis, University of Tübingen.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Conference on Message Understanding*, pages 45–52, Columbia.

Volk, M., Bubenhofer, N., Althaus, A., Bangerter, M., Furrer, L., and Ruef, B. (2010). Challenges in building a multilingual alpine heritage corpus. In *Seventh international conference on Language Resources and Evaluation*, pages 1653–1659, Malta.

Webster, K. and Curran, J. R. (2014). Limited memory incremental coreference resolution. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 2129–2139, Dublin.

Wunsch, H. (2006). Anaphora resolution - What helps in German. In *Pre-Proceedings of the International Conference on Linguistic Evidence 2006*, pages 1–5, Tübingen.

Wunsch, H. (2010). *Rule-based and Memory-based Pronoun Resolution for German: A Comparison and Assessment of Data Sources*. PhD thesis, University of Tübingen.

Wunsch, H., Kübler, S., and Cantrell, R. (2009). Instance Sampling Methods for Pronoun Resolution. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 478–483, Borovets.

Yang, X., Su, J., Lang, J., Tan, C. L., Liu, T., and Li, S. (2008a). An Entity-Mention Model for Coreference Resolution with Inductive Logic Programming. In *Proceedings of ACL-08: HLT*, pages 843–851, Columbus.

Yang, X., Su, J., and Tan, C. L. (2008b). A Twin-Candidate Model for Learning-Based Anaphora Resolution. *Computational Linguistics*, 34(3):327–356.

Yang, X., Su, J., Zhou, G., and Tan, C. L. (2004a). An NP-cluster Based Approach to Coreference Resolution. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva.

Yang, X., Su, J., Zhou, G., and Tan, C. L. (2004b). Improving Pronoun Resolution by Incorporating Coreferential Information of Candidates. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, pages 127–134, Barcelona.

Yang, X., Zhou, G., Su, J., and Tan, C. L. (2003). Coreference Resolution Using Competition Learning Approach. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 176–183, Sapporo.

Zesch, T. and Gurevych, I. (2006). Automatically Creating Datasets for Measures of Semantic Relatedness. In *Proceedings of the Workshop on Linguistic Distances*, pages 16–24, Sydney.