TOWARDS MULTILINGUAL COREFERENCE RESOLUTION

DESISLAVA ZHEKOVA

Computational Linguistics
Faculty 10: Languages and Literary Studies
University of Bremen

Submitted in accordance with the requirements for the degree
Doctor of Philosophy

Supervisors:
Prof. John A. Bateman, PhD and Prof. Dr. Sandra Kübler

Submitted: May 24, 2013
Defended: December 20, 2013

```
$myWorld = David;
print "To $myWorld . . . ";
```

ABSTRACT

The current work investigates the problems that occur when coreference resolution is considered as a multilingual task. We assess the issues that arise when a framework using the mention-pair coreference resolution model and memory-based learning for the resolution process are used. Along the way, we revise three essential subtasks of coreference resolution: mention detection, mention head detection and feature selection. For each of these aspects we propose various multilingual solutions including both heuristic, rule-based and machine learning methods. We carry out a detailed analysis that includes eight different languages (Arabic, Catalan, Chinese, Dutch, English, German, Italian and Spanish) for which datasets were provided by the only two multilingual shared tasks on coreference resolution held so far: SemEval-2 and CoNLL-2012.

Our investigation shows that, although complex, the coreference resolution task can be targeted in a multilingual and even language independent way. We proposed machine learning methods for each of the subtasks that are affected by the transition, evaluated and compared them to the performance of rule-based and heuristic approaches. Our results confirmed that machine learning provides the needed flexibility for the multilingual task and that the minimal requirement for a language independent system is a part-of-speech annotation layer provided for each of the approached languages. We also showed that the performance of the system can be improved by introducing other layers of linguistic annotations, such as syntactic parses (in the form of either constituency or dependency parses), named entity information, predicate argument structure, etc. Additionally, we discuss the problems occurring in the proposed approaches and suggest possibilities for their improvement.

ZUSAMMENFASSUNG

Die vorliegende Arbeit untersucht die Problematik der multilingualen Anwendung der Coreference Resolution (CR). Es wird bewertet, wie sich eine Rahmenstruktur verhält, die sowohl auf das "Mention-pair CR"-Modell als auch auf Memory-based Learning zurückgreift. Im Rahmen dessen werden drei wichtige Teilaufgaben der CR überprüft: Mention Detection, Mention Head Detection und Feature Selection. Für jede Teilaufgabe werden heuristische, regelbasierte und Machine-Learning Lösungen aufgestellt, und einer detaillierten Analyse, die acht Sprachen umfasst (Arabisch, Katalanisch, Chinesisch, Holländisch, Englisch, Deutsch, Italienisch und Spanisch) unterzogen. Die dazu benötigten Datensätze entstammen den beiden bisher einzigen Shared Tasks, die sich mit multilingualer CR auseinandersetzen: SemEval-2 und CoNLL-2012.

Die Forschungsrgebnisse lassen folgende Schlussfolgerungen zu: Erstens, dass die komplexe Aufgabe der CR auf multilingualem und sprachunabhängigen Wege durchführbar ist. Dazu wurden Machine-Learning-Methoden auf die genannten Teilaufgaben angewandt und ihre Ergebnisse mit den Resultaten der heuristischen und regelbasierten Herangehensweisen verglichen. Zweitens wird gezeigt, dass für jede Sprache ein entsprechendes Set von Part-of-Speech-Annotationen ausreicht, um sprachunabhängige Methoden aufzubauen. Die Leistungsfähigkeit diese Methoden kann durch das Heranziehen weiterer linguistischer Annotationen, z.B. durch syntaktische Analysen, Named-entity Informationen, Prädikat-Argument Strukture, etc., verbessert werden. Neben den Forschungsergebnissen werden Leistungen und Grenzen der behandelten Herangehensweisen diskutiert und Möglichkeiten für deren Verbesserung aufgezeigt.

PUBLICATIONS

Some of the ideas and figures presented in the following work have appeared previously in the following publications:

Sandra Kübler and Desislava Zhekova. Singletons and Coreference Resolution Evaluation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 261–267, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee. Available online at http://aclweb.org/anthology/R11-1036.pdf

Desislava Zhekova. Instance Sampling for Multilingual Coreference Resolution. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 150–155, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee. Available online at http://aclweb.org/anthology/R11-2024.pdf

Desislava Zhekova and Sandra Kübler. UBIU: A Language-Independent System for Coreference Resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99, Uppsala, Sweden, July 2010. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/S10-1019.pdf

Desislava Zhekova and Sandra Kübler. UBIU: A Robust System for Resolving Unrestricted Coreference. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 112–116, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Available online at http://www.aclweb.org/anthology/W11-1918.pdf

Desislava Zhekova, Sandra Kübler, Joshua Bonner, Marwa Ragheb, and Yu-Yin Hsu. UBIU for Multilingual Coreference Resolution in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 88–94, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4509.pdf

ACKNOWLEDGMENTS

The creation of this work would not have been possible without the help and support of numerous remarkable people. I would first like to thank my fiancee, David Münzing, who had to put up with me and my temper along the years of experimental work and shared task participations. David, thank you so much for your fare share of proofreading and advising as well as your never-ending support and help.

I would also like to express my deepest gratitude to both my supervisors: Prof. Dr. Sandra Kübler, who has always promptly and wisely responded to my endless questions with respect to the content of my work and Prof. John A. Bateman, PhD, who has given me this possibility and has encouraged my work and research over the years.

Needless to say, my loving parents Darina Doycheva and Doycho Doychev, my great family and all my fantastic friends have been a tremendous support to me through these hard times and have always managed to motivate and encourage me to continue with my work and research.

Thank you ALL, for the help!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI Artificial Intelligence

AR Anaphora Resolution

BLANC BiLateral Assessment of Noun-Phrase Coreference

CEAF Constrained Entity Alignment F-Measure

CL Computational Linguistics

mdCP Mention Detection based on Constituent Parse

CS Construct State

CR Coreference Resolution

mdDS Mention Detection based on Dependency Structure

FS Free State

FSA Finite State Automata

GNU GPL GNU General Public License

HLT Human Language Technology

mdIOBA Mention Detection based on IOB Annotation

k-NN *k*-nearest neighbor

MBL Memory-Based Learning

MCR Multilingual Coreference Resolution

MD Mention Detection

MHD Mention Head Detection

mhdH Mention Head Detection Heuristic

mhdR Mention Head Detection Rule-Based

mhdML Mention Head Detection Machine Learning Based

ML Machine Learning

MUC Message Understanding Competition

NE Named Entity

mdNES Mention Detection based on Named Entity Structure

NL Natural Language

NLP Natural Language Processing

NP Noun Phrase

POS Part of Speech

mdPOSP Mention Detection based on Part of Speech Patterns

TiMBL Tilburg Memory-Based Learner

mdVOTE Mention Detection via a Voting Technique

WWW World Wide Web

Part I

PREFACE

CHAPTER

1

INTRODUCTION AND MOTIVATION

1.1 INTRODUCTION

Computational Linguistics ***(CL)***, also called ***Natural Language Processing*** ***(NLP)*** and ***Human Language Technology*** ***(HLT)*** is an interdisciplinary field of study that aims to develop and improve the automatic processing of ***Natural Language*** ***(NL)*** in both spoken as well as written form.

computational linguistics
natural language processing
human language technology
natural language

NLP is an immensely challenging task and thus computational linguistics has rapidly been divided into a large number of major areas or subfields. The latter differ mainly on the medium of language that they target (e.g. written, spoken, etc.) or the process that they perform on this medium (e.g. analysis, recognition, generation, etc.). Some of the most prominent subfields of computational linguistics are: Machine Translation (aims at the automatic translation of text from one NL to another), Information Extraction (the automatic extraction of structured information from unstructured or semi-structured text/speech), Coreference Resolution (the automatic identification of the real-world entities to which various discourse mentions/phrases refer to and their clustering into equivalence classes according to the referents), Word Sense Disambiguation (automatic identification of the particular sense (of polysemous words) with which a word is used in a context), Automatic Speech Recognition (the automatic transformation of speech into a written text), Natural Language

Figure 1.1: A word cloud for *computational linguistics* generated via TagCrowd[1].

Generation (the automatic generation of NL from a machine readable structured form, such as knowledge base or a logical form), Question Answering (the automatic extraction of information and formulation of an answer posed in natural language), Text Summarization (the automatic transformation of a given text by extracting and reformulating the most important information from it), Text Mining (the automatic derivation of specific information from large text collections), etc.

Even though all these areas sound like relatively distant fields of exploration and research, they are deeply connected and dependent on each other. For example, larger subfields, such as Question Answering are dependent on other areas of CL, such as Text Mining, Coreference Resolution, Word Sense Disambiguation, Information Extraction, etc. This is well demonstrated by the word cloud generated for *computational linguistics* (given in figure 1.1) that combines various subareas of the field that are otherwise not directly related (e.g. logic and statistical approaches).

rule-based approaches

machine learning approaches

In general, independent of the approach (e.g. ***rule-based approaches***, which rely mostly on manually defined rules modelling natural language and ***machine learning approaches*** in which a method is employed that allows computers to make inference about the language) that is used to tackle a given NLP problem, the aim is to achieve a good ability to process text or speech in either analysis, recognition or generation. However, to be able to reach such a performance a

[1] http://tagcrowd.com

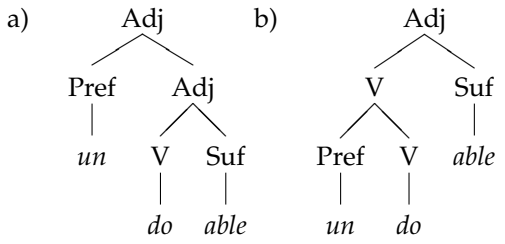

Figure 1.2: Possible morphological ambiguity for the adjective *undoable* where the structure given in a) derives the meaning of *"cannot be done"* and the structure shown in b) elicits the meaning of *"can be undone"*.

machine needs to have an in-depth representation of the input: its discourse, its deeper semantics and syntactic structure, all its ambiguities and various pitfalls. Such a representation can only be reached by exhaustive language and discourse models as well as the use of an in-depth world knowledge. Modelling world knowledge is one of the biggest challenges of computational linguistics and is a task that all subfields of CL desperately try to solve. For example, let us have a look at the ambiguity in example (1):

(1) If the dog leaves the cage, lock it.

In order to be able to find the correct referent for the pronoun *it* in example (1), one needs to have world knowledge about its predicate. This means that without knowing that a cage can be locked and that the dog needs to be in the cage in order to be locked, it is hard for a machine to pick the correct referent in this particular case. Both nouns *dog* and *cage* are equally good candidates. In fact, ambiguity may occur in all linguistic subfields and levels:

- ***Phonetics and Phonology*** - the field of study that systematically analyzes and organizes the various sounds in natural language. Possible ambiguities in phonetics and phonology consist of different NL words/phrases/fragments that sound the same but have different semantics, such as the following pairs for example: there/their, here/hear, plane/plain, sea/see, ice cream/I scream, etc. *phonetics and phonology*

- ***Morphology*** - the field of study that identifies, analyzes and describes the structure of language morphemes and other linguistic units, such as, affixes, parts of speech, etc. Ambiguity in the field of morphology occurs mainly for morphologically complex words for which there exists more than one way to combine their building morphemes and therefore there is more than one meaning of the resulting word. For example, consider the word *undoable* for which two different meanings and thus two different structures are possible. The first one, structurally represented *morphology*

in figure 1.2 a), attaches first the suffix *able* to the verb *do* that forms a word, namely *doable*, with the semantic meaning *"able to be done"*. Then the prefix *un* is added and the full meaning is changed to *"not able to be done"*. On the other hand, the tree in figure 1.2 b) attaches first the prefix *un* to the verb *do* that forms the word *undo* with the meaning *"reverse the process of doing"* and when the suffix *able* is added a different meaning is formed: *"able to reverse the process of doing"* or in other words *"can be undone"*.

syntax

- ***Syntax*** - is the field of study that examines the rules and principles according to which sentences in natural language are formed. Similar to the fields of phonetics and phonology as well as of morphology, computational linguistics can be challenged by ambiguities in syntax as well. In general, these ambiguities are similar to the ones we just presented for morphology. One very typical example is the problem of coordinated phrases. Let us take, for instance, the sentence *She met the old women and men.* Two possible syntactic structures for this sentence are given in figure 1.3 a) and b). The meaning that can be derived from part a) indicates that *She* has met multiple old women and some men (here only the women are old since there is no restriction to the relative age of the men), while the possibility in b) indicates that both groups, the women and the men, were old.

semantics

- ***Semantics*** - is the study of meaning of various linguistic structures, such as words, phrases, sentences, etc. A simple example of semantic ambiguity is lexical ambiguity. For instance, the noun *wedding* may have three different senses according to WordNet[2]: 1) the social event at which the ceremony of marriage is performed; 2) the act of marrying; the nuptial ceremony; 3) a party of people at a wedding. Such ambiguity may also pose difficulties for many NLP approaches.

pragmatics

- ***Pragmatics*** - is the subfield of linguistics concerned with the ways in which context can contribute to meaning, it is knowledge of the relationship of meaning to the actual goals and intentions of the speaker/writer. In order to decode the meaning of the sentences presented in figure 1.3, pragmatics will rely on the context provided around them in the discourse, the speaker/writer to the target part and his/her intent, but not only on the syntactic structure of the sentences.

discourse

- ***Discourse*** - is a subfield of linguistics that aims to derive knowledge about larger linguistic units than a single sentence. The actual objects of discourse analysis are mainly defined in terms of coherent structures or sequences of sentences, propositions, speech acts, or dialogue turns. For example, consider the dialogue in example (2). The referent of the

[2] http://wordnetweb.princeton.edu/perl/webwn

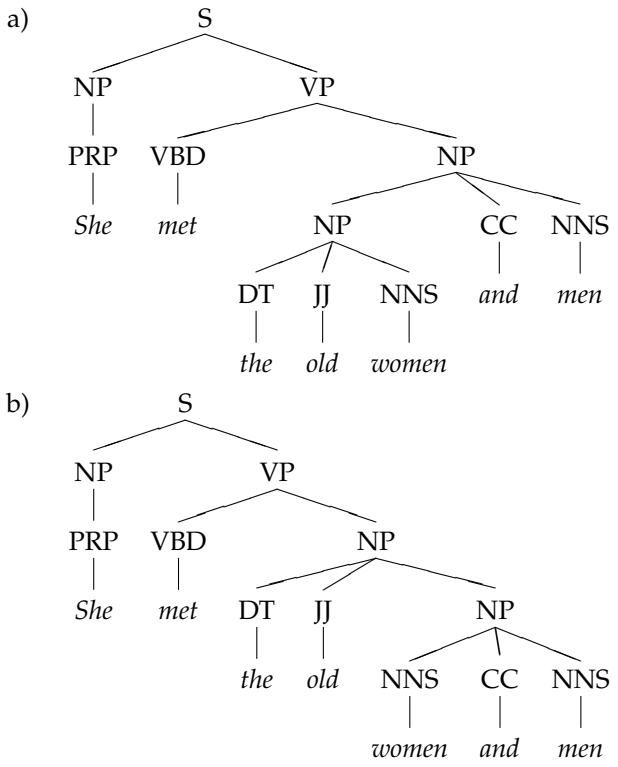

Figure 1.3: Possible syntactic ambiguity for the sentence *She met the old women and men.* where the structure given in a) means that the group of women *She* met consisted only of old women and of some men of undefined relative age, while the structure shown in b) means that both groups, women and men, consisted only of old individuals.

pronoun *it* in the last two utterances is different: *Mary* refers to the apple, while *John* refers to the shelf.

(2) *John:* The apple fell off the shelf.
Mary: It must have rolled down the pile.
John: It is a tiled one.

This last type of ambiguity, namely discourse ambiguity, is a building block of the current thesis. Discourse ambiguity is present in all natural language occurrences, both spoken and written. It is not tied to a specific language, culture, social background or age of the speaker. Indeed, natural language can be so ambiguous that it is almost impossible for any human to make use of only unambiguous discourse. Yet, people have the capability to easily and correctly find the intended meaning of the given discourse without consciously

realizing there was a difficulty. With respect to ambiguity, machines are far more inadequate than humans, because the former cannot easily derive the actual meaning of words, sentences, larger linguistic fragments, etc.

Our aim in the current thesis is to target one specific problem of discourse ambiguity, namely Coreference Resolution (CR) (see chapter 2). In the last decade the field of CR has been widely investigated and great advances have been reached, especially when one specific language, and even more, a given domain was the target of the resolution process. However, there are multiple reasons why this task should be extended to a multilingual endeavor, which is the main objective and investigation of our work in the field. The current work will provide a detailed discussion of the problems that arise when coreference resolution is applied to more than one language at a time. We present the issues that a CR system that uses the mention-pair coreference model and memory-based learning is confronted with when multilinguality is approached. Our investigation delineates various solutions for the most important subtasks in the Multilingual Coreference Resolution (MCR) field. We show that multilinguality and even language independence of the pipeline can be achieved when information-poor approaches that rely on machine learning techniques are used. Before we continue with our exploration, we need to delineate in more detail why multilinguality with respect to CR should be devoted more attention from the CL community.

1.2 MOTIVATION

In a very recent study [Rehm and Uszkoreit, 2012], META-NET[3], a Network of Excellence funded by the European Commission, has published in the form of white papers[4] a large-scale analysis of the resources provided for 23 official, national and regional languages in Europe. The analysis was targeted with respect to four different aspects connected to natural language processing: Machine Translation, Speech Processing, Text Analysis and Speech and Text Resources. The overall results of this thorough investigation indicated that there is an exceptionally high number of deficits in language technology and language support as well as significant research gaps for each of the researched languages. The study issued information for each separate language: Basque [Hernáez et al., 2012], Bulgarian [Blagoeva et al., 2012], Catalan [Moreno et al., 2012], Croatian [Tadić et al., 2012], Czech [Bojar et al., 2012], Danish [Pedersen et al., 2012], Dutch [Odijk, 2012], English [Ananiadou et al., 2012], Estonian [Liin et al., 2012], Finnish [Koskenniemi et al., 2012], French[Mariani et al., 2012], Galician [García-Mateo and Arza Rodríguez, 2012], Greek [Gavrilidou et al., 2012], German [Burchardt et al., 2012], Hungarian [Eszter et al., 2012], Icelandic [Rögnvaldsson et al., 2012], Irish [Judge et al., 2012], Italian [Calzolari et al.,

[3] http://www.meta-net.eu
[4] http://www.meta-net.eu/whitepapers

2012], Latvian [Skadiņa et al., 2012], Lithuanian [Vaišniene and Zabarskaite, 2012], Maltese [Rosner and Joachimsen, 2012], Norwegian (bokmål) [De Smedt et al., 2012a], Norwegian (nynorsk) [De Smedt et al., 2012b], Polish [Miłkowski, 2012], Portuguese [Branco et al., 2012], Romanian [Trandabăț et al., 2012], Serbian [Vitas et al., 2012], Slovak [Šimková et al., 2012], Slovene [Krek, 2012], Spanish [Melero et al., 2012] and Swedish [Borin et al., 2012].

The general outcome of the investigation on the language support across Europe was well systematized by Rehm and Uszkoreit [2012] as shown in table 1.1. The information in the table shows that across four different target areas (Machine Translation, Speech Processing, Text Analysis as well as Speech and Text Resources) the distribution of datasets, support, research and integration across the languages is highly divergent and significantly insufficient. The only regularities and general conclusions confirmed by the achieved results are the outlines that excellent support is not provided for any of the 23 targeted languages. This, in a way, is a fact that is strikingly expressive as well as worrying. In general, it means that language support and technology "must urgently" be enhanced and further developed across all European languages. Additionally, the results show, that there is only one language for which the support can be classified as good, namely English. This is a well known fact, since with the beginning of the digital age, English has been the prevailing language across all technical innovations. On the one hand, this tendency provides the ease and possibility for an in-depth research and development of new approaches, support and language-related novelties, but on the other hand, languages other than English are often neglected and correspondingly, resources are seldom developed and collected for them. The latter well explains the distribution of the languages in the last three columns of table 1.1.

The META-NET study is only one example of the tremendous and pressing need for enhanced research and language-related development for languages other than English. In the last decades, most of the state-of-the-art approaches to various natural language processing tasks have targeted largely language-specific and even more domain-specific solutions to diverse computational linguistic subtasks and problems. Such approaches have flourished, because only they could achieve an acceptable overall performance. However, the development of separate software, concepts, algorithms, resources and support for each domain and that in each language is a highly costly endeavor. In the last years, the computational linguistics community has invested massive assets in the development of such language-specific, linguistically prepared and annotated (meaning labeled or analysed) data. The latter has lead to a significant improvement in various NLP areas, but has mostly been concentrating on well resourced languages, such as English, German, French and Spanish, which is also confirmed by the META-NET study.

With this work, we claim that the digital era has pushed the limits of language-specific approaches and that nowadays much more effort

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
Machine Translation				
	English	French, Spanish	Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian	Basque, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian (Bokmål, Nynorsk), Portuguese, Serbian, Slovak, Slovene, Swedish
Speech Processing				
	English	Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish	Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian (Bokmål, Nynorsk), Polish, Serbian, Slovak, Slovene, Swedish	Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian
Text Analysis				
	English	Dutch, French, German, Italian, Spanish	Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian (Bokmål, Nynorsk), Polish, Portuguese, Romanian, Slovak, Slovene, Swedish	Croatian, Estonian, Icelandic, Irish, Latvian, Lithuanian, Maltese, Serbian
Speech and Text Resources				
	English	Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish	Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian (Bokmål, Nynorsk), Portuguese, Romanian, Serbian, Slovak, Slovene	Icelandic, Irish, Latvian, Lithuanian, Maltese

Table 1.1: A summary of the outcome of the META-NET white paper series as presented in [Rehm and Uszkoreit, 2012] in which the four areas (Machine Translation, Speech Processing, Text Analysis as well as Speech and Text Resources) were evaluated.

and research should be invested into multilingual or even better language-independent algorithms, methods, systems, concepts and models. We believe that such a direction of advancement is more appropriate to the development of society and especially to the changes in all *living* languages. For this reason, our investigation focuses mainly on solving the issues that raise when multilinguality is targeted within the coreference resolution task. Our approaches suggest various ways to tackle the problems, which we compare, discuss and evaluate in the context of this complex task. The following section (section 1.3) lists a detailed outline of the work.

1.3 OUTLINE

In order to be able to present the problems with multilingual coreference resolution, we first need to properly delineate the dimensions of the CR task. For this reason, chapter 2 serves as a detailed introduction to the field, while chapter 3 exhibits the problems and issues that multilinguality poses to this already exceedingly complex task. In chapter 4, we isolate the issues to the context of a specific framework, namely one that uses the mention-pair coreference model and memory-based learning for the resolution process. The framework is set by the MCR system in use: UBIU [Zhekova and Kübler, 2010, 2011, Zhekova et al., 2012]. Chapter 5 discusses the first subtask, namely mention detection, within the MCR pipeline that is challenged by the multilinguality aspect. We present various heuristic, rule-based and machine learning methods to tackle the problem and compare, discuss and evaluate all proposed approaches. Another subtask of MCR that also needs additional vigilance is mention head detection. Similar to our work on mention detection, chapter 6 reviews different heuristic, rule-based and machine learning methods that can be applied to multilingual mention head detection. We present their evaluation and discuss to what extend they can be employed on a language independent level. Finally, the last important subarea of MCR that we need to examine is feature selection. In chapter 7, we describe the difficulty of this task and the ways in which it can be altered in order for language independent approaches to be possible. In chapter 7, a thorough evaluation and discussion of the system results with all employed methods is also presented and a language dependent optimization based on the achieved output is carried out. Chapter 8 summarizes our findings, proposes new directions and concludes our work.

Part II

BACKBONE

CHAPTER

2

FUNDAMENTALS OF COREFERENCE RESOLUTION

Natural Language (NL) is inherently ambiguous, which is one of the reasons why its automatic processing is so immensely difficult. It is not always easy for a human to decipher all hidden pitfalls of ambiguity, let alone a machine that generally does not have a very profound knowledge of the world that we live in. One of the requirements for NLP to be perceived as a manageable task is the ability to easily and precisely identify and classify the real world entities that we refer to on a daily basis. Yet, from the CL perspective, if we consider this classification task as a hard one, being able to perform it for multiple languages is, indeed, even harder.

Before we start discussing the issue and predicaments of multilinguality and their instances in Multilingual Coreference Resolution (see chapter 3), we first need to delineate the concepts of its components. Thus, in the current chapter we will introduce basic notions such as Reference Resolution (see section 2.1), then in section 2.2 we will present Anaphora Resolution and following in section 2.3 we will discuss Coreference Resolution. Section 2.4 provides a short summary and conclusive remarks for the discussion in this part.

2.1 REFERENCE RESOLUTION

Ambiguity has been attracting the attention of various linguistic researchers for several decades now [Hindle and Rooth, 1993, Kooij, 1971, Krovetz and Croft, 1992] and at present it is still one of the most significant and unresolved problems in NLP, for NL can be inherently ambiguous at multiple levels (e.g. the lexical level, the pragmatic level, the reference level, the structural level, etc.). For this reason innumerable state-of-the-art computer linguistic systems are generally restricted to the use of domain specific discourse as the latter reduces the uncertainty of meaning and its countless possibilities for interpretation. Yet, in order to be able to comprehend NL with respect to any given discourse one first needs to be capable of identifying the separate entities and the references or relations that are being used in that discourse [Webber, 1978].

(3)

John: *Mary baked a vanilla slice for the birthday party.*

Bob: *Really?*

For example, we can consider the first sentence in (3). There are three different entities that can be extracted just from this short utterance, as visualized in example (4) – *Mary*, *vanilla slice* and *the birthday party*.

(4)

John: *[Mary]$_1$ baked a [vanilla slice]$_2$ for [the birthday party]$_3$.*

Bob: *Really?*

reference
referring expression
referent
reference resolution

The relation between the surface form in use and the actual discourse entity is defined as ***reference***. Identifying the linguistic expressions that refer to a given real-world entity (also called ***referring expressions***) in a sentence and establishing the relations between them and the discourse entities (***referents***) they refer to is the main task of ***reference resolution***. However, in our toy example (4) all three linguistic expressions refer to distinct discourse entities. If we extend the example and add another sentence, as shown in (5), we acquire three further linguistic expressions (marked with indices 4, 5 and 6 in example (5)) that need to be identified and related to their corresponding real-world entities.

(5)

John: *[Mary]$_1$ baked a [vanilla slice]$_2$ for [the birthday party]$_3$. Unfortunately, [she]$_4$ forgot [the cake]$_5$ in [the oven]$_6$.*

Bob: *Really?*

#	Referring Expression	Referent
1	Mary, she	Mary
2	vanilla slice, the cake	the vanilla slice cake
3	the birthday party	the birthday party
4	the oven	the oven

Table 2.1: The referring expressions and their corresponding referents extracted from the toy sentences given in example (5).

Yet, as can be seen in table 2.1 in which all referring expressions are listed with their corresponding referents, not all new phrases refer to entities that are new to the discourse. In (5) the personal pronoun *she* refers to the proper noun *Mary* previously mentioned in the discourse and the Noun Phrase (NP) *the cake* refers to the NP *vanilla slice* that has been also previously introduced. This analogy can be further generalized by the terms ***antecedent*** (denoting the referring expression that appears previous to a referring expression to the same discourse entity) and ***anaphor*** (denoting the referring expression to an entity that has already been referred to previously in the discourse). For this reason, the anaphor and the antecedent are bound by an ***anaphoric relation*** and the actual reference to a given entity that has already been introduced in the discourse is called ***anaphora***. However, in example (5), *Mary* also refers to *Mary* as well as *vanilla slice* also refers to *vanilla slice*. Such referring expressions that share a single discourse entity as their referent are said to ***corefer***.

antecedent

anaphor

anaphoric relation

anaphora

corefer

The difference between anaphora and coreference is often misunderstood or equivocal, because the terms do describe similar phenomena. Before we continue with more in-depth presentation of both concepts we consider it essential to clarify the difference between them at this stage. A very precise separation of coreference and anaphora is provided by Deemter and Kibble [2000]. The authors define the former as an equivalence relation and as such it can be described as reflexive, symmetric and transitive. Even though anaphora is often considered to be a building block of coreference, it does not inherit any of the latter three relations, it is irreflexive, nonsymmetrical and nontransitive. One of the most characteristic differences between both phenomena is the necessity of context for an appropriate interpretation. Anaphora is context-sensitive, meaning that the resolution of the anaphor always depends on the context, or in other words, the interpretation of the anaphor depends on the interpretation of the antecedent. For coreferent phrases, however, the interpretation of the anaphor does not depend on the interpretation of the antecedent and can be achieved independently for each considered phrase.

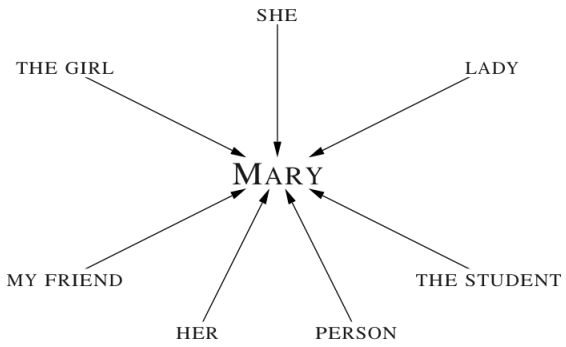

Figure 2.1: An example of potential anaphora relations with *Mary* being the antecedent and *she, lady, the student, person, her, my friend, the girl* potential anaphors.

2.2 ANAPHORA

anaphora resolution

Anaphora resolution [Barss, 2003, Branco, 2007, Lappin and Leass, 1994, Mitkov, 2002, Mitkov et al., 2001, Reuland, 2011] is the task that aims at the identification of the antecedent of a target word or phrase previously introduced to the discourse. If we refer back to example (5) on page 16, the task of anaphora resolution will be to detect that the first noun phrase *Mary* is the antecedent of the anaphor *she*.

Yet, anaphora resolution does not aim to detect all possible antecedents of the target anaphor in the discourse. A good representation of potential anaphoric relations is depicted in figure 2.1 – only the relations between the phrases that can be used to refer to *Mary* are sought and not further relations among those phrases. For this reason, as Mitkov [1999a] reports, the task of anaphora resolution is already completed when one of the antecedents of the target anaphor is identified (while coreference attempts to detect all phrases in the text that refer to the entity that the anaphor refers to). Thus, Anaphora Resolution (AR) is generally considered as being a subtask of Coreference Resolution (CR) (see section 2.3). In computational linguistic applications, anaphora most often occurs only as pronominal anaphora (section 2.2.1.1).

There are various types of anaphora, which are normally distinguished either by the form of the anaphor or the location and the type of relation that binds both. The following sections will only briefly introduce the basic types of anaphora for it is not our aim to give a full account of that phenomenon. A more detailed delineation of the concept can be found in [Barss, 2008, Fox, 1993, Kabadjov, 2010, Mitkov, 2002]. Yet, a potential AR and respectively

CR implementation needs to distinguish between the different instances of both phenomena. In our explanation, we will mainly follow Mitkov's [2002] description and extract only the instances most common and most relevant to our further discussion. We will first introduce the types of anaphora according to the form of the anaphor (see section 2.2.1), then we will present possible variations according to the locations of the anaphor and the antecedent (see section 2.2.2) and in section 2.2.3 we will include other specific types that do not fall in the first two categories.

2.2.1 *Types of Anaphora according to the Form of the Anaphor*

Each of the different types of anaphors can have a variety of antecedents: nouns, noun phrases, verbs, verb phrases, clauses, sentences and even a sequence of sentences. Yet, there is no defined subdivision of the type of the anaphora phenomenon depending on the type of the antecedent, but rather depending on the type of the anaphor.

In the following section we distinguish between four types of anaphora depending on the form of the anaphor: pronominal anaphora (section 2.2.1.1), lexical noun phrase anaphora (section 2.2.1.2), verb/adverb anaphora (section 2.2.1.3) and zero anaphora (section 2.2.1.4). Those four instances are among the most important and most often occurring types of anaphora and thus they are also highly important to our further discussion.

2.2.1.1 *Pronominal Anaphora*

Pronominal anaphora is the anaphoric relation between an antecedent and a pronoun anaphor. In table 2.2, we provide an example of an anaphora relation between each type of pronoun and a corresponding antecedent. We extracted these instances from the CoNLL-2012 English dataset [Pradhan et al., 2012]. We will present this data in more detail in section 3.1.2. As Gundel et al. [2003] present, pronouns are normally used to refer to entities introduced to the discourse by a nominal expression. Clausal or non-nominal constructions (such as situations, facts, acts, etc.), however, are mostly referred to by demonstratives.

pronominal anaphora

Pronouns are generally said to be anaphoric. Yet, one often seen exception to the anaphoric phenomenon is the ***pleonastic*** occurrence of the personal pronoun *it*. This is the occurrence in which the pronoun *it* does not refer to a specific entity, e.g. when it is used in temporal constructions, cleft constructions, etc. Those uses of *it* are not considered anaphoric. An example of a pleonastic use of *it* is provided in example (6).

pleonastic

(6) *It* seems that Mary forgot the vanilla slice in the oven.

Generic (a pronoun that does not refer to a specific referent, as in example (7)) and ***deictic*** (acquiring a meaning only in a given context, as in exam-

generic
deictic

personal	***Parks*** inspired the Civil Rights movement of course when ***she*** refused to give up her seat on a bus to a white man in nineteen fifty-five.
possessive	***Daisy*** Peters' feelings are understandable, ***her*** house in ruins.
reflexive	***Chang*** is ***herself*** an accomplished carver of seal knobs, and she knows the details and dates of all Wang's seals and calligraphies.
demonstrative	Lin Mei-lun's mother often ***comes*** to stay, and ***this*** has done much to lighten the burden of running her household
relative	Of ***all the ethnic tensions in America***, ***which*** is the most troublesome right now?

Table 2.2: Example instances of anaphoric pronouns (*personal, possessive, reflexive, demonstrative, relative*) extracted from the CoNLL-2012 English dataset [Pradhan et al., 2012].

ple (8)) uses of pronouns are also an exception to the anaphora phenomenon and thus generic and deictic pronouns are not considered anaphoric.

(7) *One* should never waste chocolate.

(8) *I* love chocolate.

In (7) there is no reference to a specific person, whereas in (8) there is, but only if that sentence is used in a context in which the referent behind the pronoun *I* has already been introduced.

2.2.1.2 *Lexical Noun Phrase Anaphora*

lexical noun phrase anaphora

definite expressions

Lexical noun phrase anaphora is the type of anaphora that considers only a subset of all ***definite expressions*** as an anaphor. In that regard, definite expressions are definite noun phrases, proper names, personal, reflexive, possessive and demonstrative pronouns. Lexical noun phrase anaphora should not be confused with noun anaphora (see section 2.2.3). Only definite NPs (e.g. *the only person* in (9)) and proper names (e.g. *Mary* in (10)) can be used as an anaphor in lexical noun phrase anaphora. ***Indefinite expressions*** (phrases that are not specific and not identifiable), as *a cake* in both example (9) and example (10), cannot be considered as anaphors, for they are new to the discourse and there is no specific antecedent preceding them in the context.

indefinite expressions

(9) *Mary* is *the person* that can make me bake a cake.

(10) *The person* that can make me bake a cake is *Mary*.

Similar to pronominal anaphora, lexical noun phrase anaphora can also make use of generic expressions that (as we described for pronominal anaphora) are not anaphoric. Those expressions can be both definite NPs (see example (11)) and proper names (see example (12)).

(11) Who invented *the telephone* is a well known fact – it was Alexander Graham Bell!

(12) We always have ice-cream on *Monday*.

2.2.1.3 *Verb/Adverb Anaphora*

Not only noun phrases and pronouns can be employed as anaphors. Verbs, defining ***verb anaphora*** can also serve that purpose as well as adverbs, representing ***adverb anaphora***. These two types are not as common as the ones we presented above. A verb can have both a verb (as in (13)) or a verb phrase (as in (14)) as a possible antecedent.

verb anaphora
adverb anaphora

(13) *Tell* me that you brought the cake, please *do*!

(14) John explicitly reminded Mary not *to forget* the cake when he called to tell her about the traffic jam on the way, yet, Mary *did*.

2.2.1.4 *Zero Anaphora*

Zero anaphora is the last type of anaphora that we will discuss in the current section. Already its name signals the way that this relation is realized in its surface form, namely the anaphor position does not contain anything, it is empty (denoted further as ∅). Depending on the types of the omitted anaphor, there are several types of zero anaphora: ***zero pronominal anaphora*** (in which the missing anaphor is a pronoun (see example (15)), ***zero noun anaphora*** (where the missing anaphor is a noun or only the head noun of a phrase, but not the whole phrase itself (see example (16)), ***zero verb anaphora*** (accordingly, this is the type of zero anaphora in which the missing anaphor is a verb (see example (17)) and ***zero verb phrase anaphora*** also known in the literature as ***ellipsis*** (here instead of missing only the verb, the whole verb phrase is excluded from the surface form (see example (18)).

zero anaphora
zero pronominal anaphora
zero noun anaphora
zero verb anaphora
zero verb phrase anaphora
ellipsis

(15) *She* left the birthday party and ∅ drove off. (∅ = she)

(16) A lot of *people* at the party asked her about the cake, but some ∅ didn't. (∅ = people)

(17) Mary *baked* a vanilla slice cake for the birthday party and ∅ some muffins for her grandmother. (∅ = baked)

(18) Mary has never *baked a cake*, but she was surprised to find out that no one else has ∅. (∅ = baked a cake)

pro-drop languages
pronoun-drop languages

Zero pronominal anaphora occurs in ***pro-drop languages*** also called ***pronoun-drop languages***. A pro-drop language is one in which given pronouns can be omitted, but only in the cases in which they are pragmatically inferable. For example, an elliptical subject can be excluded from the surface form of a sentence but introduced in a different ***annotation layer***, the syntactic annotation for instance, and still be considered coreferent. Annotation layers, also referred to as ***linguistic annotations*** or simply ***annotations***, are additional levels of information (both descriptive and analytic) provided for a given raw text. In example (19), the Spanish sentence includes an elliptical pronoun subject that is not included in the surface form. Yet, it is still considered coreferent with the proper name *Toni*. Such pro-drop features, conditions and annotations can be highly complex and differ considerably from language to language, which is a crucial issue when multilinguality is considered.

annotation layer
linguistic annotation
annotation

(19) Spanish: "Por ahora, [$\emptyset_1$] prefiero no hablar de [ese asunto $_2$]", concluya [Toni $_1$].
English: " For now, *I* would rather not discuss that issue," Toni completed.

2.2.2 *Types of Anaphora according to the Locations of the Anaphor and the Antecedent*

This section is devoted to the differences in anaphora depending on the location of both the anaphor and the antecedent. Most of the example sentences we have given by now represent ***intrasentential anaphora*** – this means that both the anaphor (*herself*) and the antecedent *Mary* are within the bounds of a single sentence as represented in example (20). Yet, if we extend example (20) with another sentence as in example (21) we can observe that the anaphoric relation between the anaphor *It* and the antecedent *the cake* spans over the sentence boundaries. This is an example of ***intersentential anaphora***.

intrasentential anaphora
intersentential anaphora

(20) *Mary* told *herself* that going without the cake is not a big deal.

(21) Mary told herself that going without *the cake* is not a big deal. *It* is not going to be eaten anyways.

In specific cases, the type of the pronoun used as an anaphor can already select the type of anaphora it falls in. For example, reflexive pronouns, as *herself* in example (20), always fall in the case of intrasentential anaphora.

In the last decade, there has been an exceptional effort to increase the available datasets containing anaphora annotations (see section 2.3.4.1). Such datasets consist of collections of documents and, often, a given story/thread/topic in that data is not represented only by a single document. For this reason, it is possible that anaphors refer to antecedents in the document previous to their own – ***interdocument anaphora***. Respectively, in the cases in

interdocument anaphora

which both phrases are in the same document we talk about ***intradocument anaphora***.

intradocument anaphora

2.2.3 *Further Specific Types*

In sections 2.2.1 and 2.2.2, we have introduced the types of anaphora that are most often used and of high importance to our forthcoming discussion. However, Mitkov [2002] extends his discussion on the topic and delineates several other shades of the phenomenon. One of them is ***indirect anaphora*** which arises when the relation between the anaphor and the antecedent requires world knowledge and inference. This is, because the anaphor is not referring to the exact same entity that the antecedent presents, but rather to a part of it, a member from it, a more specific subset or a superset, etc. Example (22) shows an instance of indirect reference between the antecedent *a piece* that is only a piece of the whole cake and the anaphor *The cake* which is the superset of the antecedent.

indirect anaphora

(22) Mary cut herself *a piece* and ate it right after she got back home. *The cake* was delicious.

The counterpart of indirect anaphora is called ***direct anaphora***. It occurs when both the anaphor and the antecedent share the exact same head of the phrase as example (23), in which the head *cake* is present in both the antecedent *vanilla slice cake* and the anaphor *that cake*.

direct anaphora

(23) Mary was now really sorry that she forgot the *vanilla slice cake* – she did not even suspect that she can bake properly when she made *that cake*.

In general, the anaphor refers to the exact same individual that the antecedent represents, which is when the term ***identity-of-reference anaphora*** is used. Yet, this is not always the case. Example (24) shows a relation between the pronoun *it* and *the vanilla slice* that is not one of identity-of-reference, because the grandmother did not think of baking that exact same cake but a cake of that type – vanilla slice. This type of anaphora is known as ***identity-of-sense anaphora***.

identity-of-reference anaphora

identity-of-sense anaphora

(24) Mary told her grandmother about the success of *the vanilla slice*. The old woman remembered often baking *it* and she smiled at her granddaughter.

A very special case of identity-of-sense anaphora is ***noun anaphora***. It represents the relation between the antecedent and an anaphor that is a non-lexical ***pro-form*** (a function word that replaces another word, phrase or even a clause in order to avoid repetition). In example (25) the pro-form *one* is used to replace a word or a phrase describing a piece of the cake, but, a piece different from the pieces Mary and her grandmother actually ate.

noun anaphora

pro-form

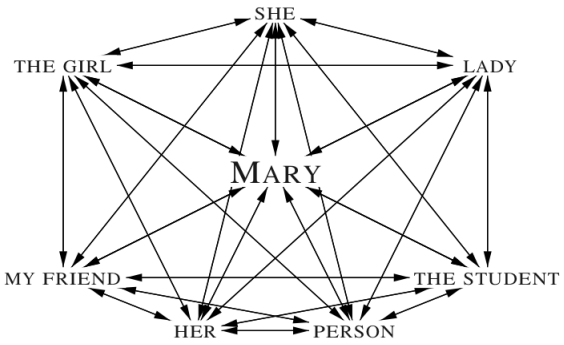

Figure 2.2: A lattice structure of potential coreference relations depicting the possibility for connecting mentions and the respective formation of a coreference chain.

(25) Mary and her grandmother ate *a piece* of the cake and gave *one* to the raccoon that came by through the garden fence.

As Mitkov [2002] further describes, in the cases most often seen in natural language, the antecedent precedes the anaphor. For pronominal anaphora that precedence is usually within the same sentence or the sentence before it. In case that the order in which the antecedent and the anaphor occur is reversed, the relation that binds both is ***cataphoric*** and the corresponding reference is consequently ***cataphora***. An instance of such an occurrence is given in example (26) in which the anaphor *Mary* occurs after its antecedent *she*.

cataphoric relation
cataphora

(26) When *she* arrived at the party, *Mary* realised her blooper.

2.3 COREFERENCE RESOLUTION

In general, ***coreference*** is tightly bound to anaphora. It extends anaphora in such a way that we talk about identifying not only one, but all expressions referring to one single discourse entity. The actual process of identification is known as ***coreference resolution***. As figure 2.2 depicts, not only the links from a given phrase to the target entity are sought (as was shown in figure 2.1), but the links between the referring phrases as well. Once all linguistic expressions that link to the same entity are identified, an ***equivalence class*** or ***coreference chain*** is created. An equivalence class may have an unlimited number of members as long as all refer to the same discourse entity. All phrases that can potentially be part of an equivalence class, or a coreference chain, are called ***mentions***. The process of identifying or extracting mentions from the data

coreference
coreference resolution
equivalence class
coreference chain
mention

is called ***mention detection***. Mentions can be found in the literature as well under the name of ***markables*** or even ***potentially anaphoric phrases***. Similar to anaphora, coreference also restricts the type of phrases that are accepted as mentions – these can be noun phrases, pronouns, named entities or verbs.

mention detection
markable
potentially anaphoric phrase

In order to exemplify better what equivalence classes and respectively coreference chains are, let us look at the text in example (27)[1] which is an excerpt from the CoNLL-2012 Shared Task English dataset [Pradhan et al., 2012]. Section 3.1.2 describes this data in more detail.

(27)

Larry King:	*Hello hello Jay Georgia hello.*
caller_3:	*Ah thank (you_1) ($Larry_1$). And ($Mike_2$) (I_3) loved (($your_2$) book $_4$). (It_4) was great. And toward the end of the ($book_4$) (you_2) said Secretary (Putin of $Russia_5$) had asked (you_2) to come over and ($interview_6$) (him_5). Had (you_2) done ($that_6$)? Uh and (I_3)'d like to know about (it_6). Thank (you_2) so much.*
Mike Wallace:	*Yeah.*
Larry King:	*(I_1) did interview ($Putin_5$) yes.*
Mike Wallace:	*on the sixtieth anniversary of the uh end of World War Two (he_5) asked (me_2) to come on over and ($interview_7$) (him_5). And (it_7) was carried uh in a lot of places. But (I_2) tell you something. ($Putin_5$) to (my_2) way of thinking who calls ($himself_5$) a democrat - (He_5)'s not our kind of democrat.*

In the given context in example (27), there are seven entities that were referred to more than once: 1) Lary; 2) Mike; 3) caller_3; 4) Mike's book; 5) Putin; 6) an interview; 7) the interview. For this reason, seven different coreference chains could be formed with all phrases used to refer to those seven entities. Once the mentions are considered to be a part of a specific coreference chain they are said to refer to the same real-world entity that is represented by that chain. In other words, all phrases are different descriptions or linguistic forms of the same entity. A graphical visualization of the resulting classes from example (27) is depicted in figure 2.3.

Resolving the equivalence classes determines the set of ***coreferent mentions***. This is the collection of all members of the equivalence classes or in other words the collection of all mentions that refer to an entity that has more than one referent in the text. The rest of the mentions are said to be ***singletons*** – mentions that refer to an entity in the text that no other mention refers to. In

coreferent mention
singleton

[1] The text is presented as it occurs in the data, no punctuation, grammar or other types of errors were corrected. The IDs assigned to the coreference chains were simplified with smaller numbers for better readability.

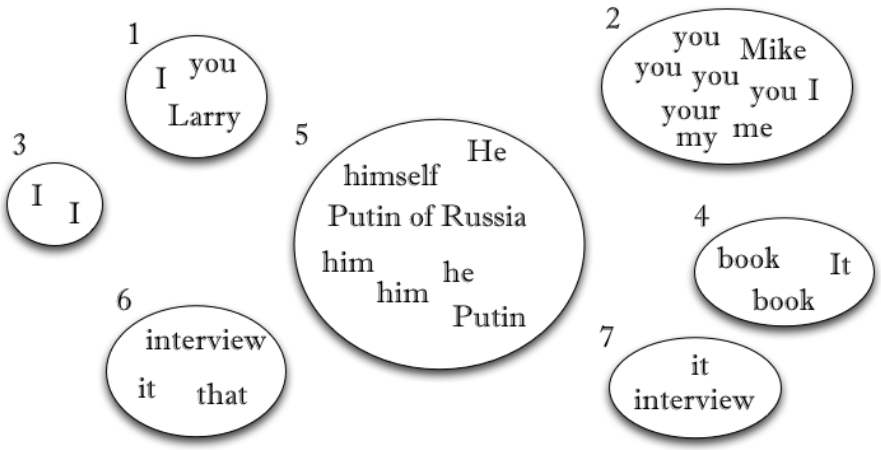

Figure 2.3: A graphical representation of the seven resulting equivalence classes (coreference chains) extracted from example (27).

terms of set theory the set of all mentions is the union of the set of singletons and the set of coreferent mentions graphically represented in figure 2.4 on page 27.

After we have clarified the difference between anaphora resolution and coreference resolution, from this point on we will use only the latter to refer to both or will explicitly note if we refer only to anaphora resolution.

Since coreference is extensively used in natural language, it is regarded as highly important with respect to the preservation of a coherent discourse nature.

cohesion

Cohesion represents the grammatical and lexical relations within a given discourse in which the interpretation of a linguistic expression is dependent on a previously used alternative variant of that expression. In other words, cohesion is often defined as the links that preserve the meaning and integrity of text [Halliday and Hasan, 1976, Morris and Hirst, 1991, Hobbs, 1978].

Knowing what coreference resolution is, however, does not give us much information about the computational linguistic approaches that have been explored for tackling the CR problem. As Wunsch [2010] presents in his comparative study on anaphora resolution, one can employ either rule-based approaches (see section 2.3.1) or machine learning approaches (see section 2.3.2) to solve the anaphora problem. The same is also possible for coreference resolution. For this reason, we devote the next two sections to these two variations. In section 2.3.3, we discuss additional improvements or advances in both and in section 2.3.4, we introduce the available resources and attempted evaluation approaches.

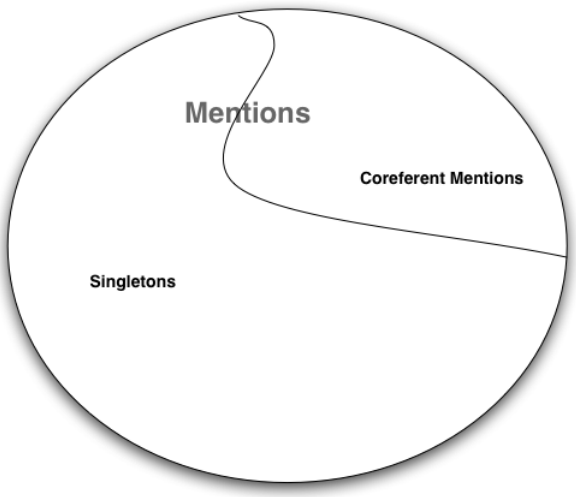

Figure 2.4: A graphical representation of the set of all mentions defined as the union of the set of singletons and the set of coreferent mentions.

2.3.1 *Rule-Based Approaches to Coreference Resolution*

rule-based approaches

Rule-based approaches to CR rely on the availability of lexical and encyclopedic knowledge and manually handcrafted rules. Earlier work of this type of approach is reviewed by Aone and McKee [1993], Harabagiu et al. [2001], Markert and Nissim [2005], Mitkov [1998], Poesio et al. [2002], Yang and Su [2007]. As Harabagiu et al. [2001] notes, such rules are designed to capture both grammatical and lexical cohesion by ensuring agreement of gender, number, semantic class, etc. The examples of coreference rules that Harabagiu et al. [2001] show (duplicated in figure 2.5 for more convenience), demonstrate how such an agreement can be forced over the pairs. However, rule-based methods' reliance on handcrafted patterns has several drawbacks. Although the latter are generally easy to design because of their simplicity and self-explanatory individual structure, the lack of organization across entire rule sets leads to difficulties in observing the impact of each distinct rule on the whole process. This makes such rule sets difficult to validate and, since there is a direct correlation between the number of rules included in the system and its efficiency, also renders the processes using them progressively more inefficient as the rule sets grow. However, the biggest disadvantages of rule-based approaches are, on the one hand, the need to individually implement and regularly revise rules for each separate phenomenon under consideration and, on the other

```
RULE-1-Filter-1-Pronoun (R1F1Pron)
  If (( Syntactic_Category(Anaphor)== Pronoun) AND  Repetition (Anaphor, Antecedent) )
     then  Cast_in_Chain(Antecedent, Anaphor)
RULE-1-Filter-1-Nominal (R1F1Nom)
  If (( Syntactic_Category(Anaphor)== Common Noun) AND  (Anaphor == Apposition(Antecedent) )
     then  Cast_in_Chain(Antecedent, Anaphor)
RULE-2-Filter-1-Nominal (R2F1Nom)
  If (( Syntactic_Category(Anaphor)==  Syntactic_Category(Antecedent)==Proper Noun) AND  Same-Category(Antecedent,Anaphor) )
     If  ( Category(Anaphor) == PERSON) AND (  Last_Name(Antecedent)==Last_Name(Anaphor) )  AND
                 AND (Gender(Antecedent) = Gender(Anaphor) AND  Surface_Distance(Anaphor,Antecedent)=min)
          then  Cast_in_Chain(Antecedent, Anaphor)
     If  ( Category(Anaphor) == ORGANIZATION) AND  Acronym(Anaphor,Antecedent))
          then  Cast_in_Chain(Antecedent, Anaphor)
```

Figure 2.5: An example of multiple manually developed rules for coreference resolution as presented in [Harabagiu et al., 2001]. This set has been developed for English and would need to be revized and adapted if applied to other languages.

hand, an inherent inability to learn and re-use experience gained during rule set building.

As Klaussner and Zhekova [2011] show, manually designed patterns can be highly accurate when they are formulated for not so complex phenomena such as hyponymy for example. Yet, the introduction of new examples, exceptions, special cases and variation will inevitably lead to imperfection and thus the need for revision and adaptation of the set of rules. This can be highly labor intensive and therefore it is not a reasonably achievable goal. Recognizing the immense effort that is needed for such an enterprise, Aone and Bennett [1995] evaluate the trade off between employing a manual vs. automated approach to anaphora resolution and conclude that rule-based methods as in [Aone and McKee, 1993], even if exceptionally robust and extensible, should nevertheless be geared towards a truly automated approach.

2.3.2 *Machine-Learning Approaches to Coreference Resolution*

machine learning approaches

machine learning artificial intelligence

In the search for other possibilities for addressing the CR task, it has been natural to consider ***machine learning approaches*** as in [Aone and Bennett, 1995, Luo et al., 2004, McCarthy and Lehnert, 1995, Ng and Cardie, 2002a, Ponzetto and Strube, 2006, Soon et al., 2001, Versley et al., 2008a, Yang et al., 2003]. Machine Learning (ML) is a branch of Artificial Intelligence (AI) that considers various methods or algorithms for the analysis of data based on empirical evidence. One of the principal advantages of machine learning approaches is their concentration on the automatic extraction, analysis and evaluation of patterns varying in complexity which can subsequently be used for intelligent decision-making based on the information uncovered. A natural hurdle to machine-based CR, however, is that it is necessarily *data driven* and

thus dependent on the resources available for the targeted language. This is typical not only for machine-based CR, but as well for all other machine learning NLP approaches. Yet, the resources that those methods require are still in short demand (see section 2.3.4).

Instead of manually handcrafted rules, machine learning for CR most often recasts the problem as a binary classification task. The latter is also known as the ***mention-pair model***. The mention-pair model is the most widely used and understood model for attempting the coreference task. Rahman and Ng [2011] summarize three further approaches: ***mention-ranking model***, ***entity-mention model*** and their own ***cluster-ranking model***. Further on in their discussion, Rahman and Ng [2011] comment that all the latter models attempt to account for the two major weaknesses of the mention-pair model: its inability to represent transitivity within the CR phenomenon and the limitations in its expressiveness caused by the consideration of information representing only two mentions at a time. However, neither the mention-ranking model nor the entity-mention model manage to resolve both problems simultaneously. Furthermore, these models, as well as the cluster-ranking model have a highly increased complexity and thus cannot be employed as a simple baseline approach. Thus, most state-of-the-art CR systems, as we show in sections 3.1 and 7.1.1, make use of the mention-pair model. For the latter reason, we employ this model in the investigation presented further in this work.

mention-pair model

mention-ranking model

entity-mention model

cluster-ranking model

The first step in the employment of a mention-pair model (see chapter 4 for further details) is to identify all potential mentions in the data. Then, all detected phrases are paired up and ***features*** (e.g. information as number, gender, semantic class, syntactic dependents, etc.) for the mention pairs and their context are extracted and stored as ***feature vectors*** (a collection of features, or more precisely, a collection of feature values). Then, an appropriate machine learning algorithm that represents a distinct ***predictive model*** should be selected. A predictive model defines the way in which the outcome is computed.

feature

feature vector

predictive model

In case ***supervised learning*** is employed [Soon et al., 2001, Ng and Cardie, 2002a, Bengtson and Roth, 2008], the correct labels, or in other words, answers for the feature vectors of the mention pairs are provided in the initial (training) dataset. Otherwise ***unsupervised learning*** [Ng, 2008, Haghighi and Klein, 2007] is made use of. A decision on the new, unseen instances known in the literature as ***test data*** is made, based on the information extracted from the initial corpus/corpora also called ***training data***. Once the classification task is finished the resulting positively linked pairs are clustered into equivalence classes as in figure 2.3. A graphic representation of a machine learning process is shown in figure 2.6.

supervised learning

unsupervised learning

test data

training data

Solving the problem of CR means that a complex combination of constraints based on the correctness and coherence of salience, syntax, semantics and discourse needs to be fulfilled. For example, in order to be coherent a given text needs to fulfill the constraint of including successive linguistic expressions

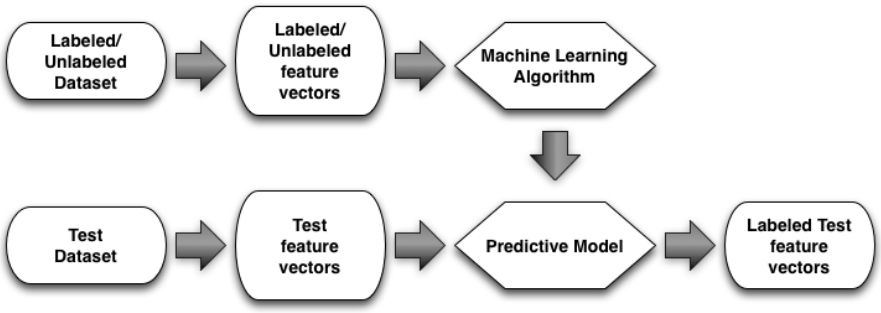

Figure 2.6: The backbone scheme of a machine learning process using labeled (for supervized) or unlabeled (for unsupervized) training data for the learning process and classifying the test data according to a previously selected predictive model.

that refer to the same entity. Hence, assuming that a text is coherent, one can traverse the latter constraint in use for coreference resolution, meaning that only phrases found in a close distance can be rendered coreferent. Another such restriction can be based on the syntactic analysis of the context: reflexive pronouns are most often coreferent with the closest preceding subject, thus syntactic knowledge can aid a constraint that ties such pronouns to the last seen subject.

Machine learning approaches, however, cannot always account for such specific restrictions that are based on precise observations. For this reason more complex models combining both machine learning and rule-based methods are often used. The latter are known as ***hybrid approaches*** [Hartrumpf, 2001, Lalitha Devi et al., 2011]. Hybrid approaches can represent a diverse combination of both methods throughout all subtasks of CR. One such combination for example is using machine learning for the resolution process and rules in a postprocessing step to enhance the output.

hybrid approaches

2.3.3 *Various Improvements to Coreference Resolution*

In the attempt to advance and improve CR system performance, research has considered a variety of further aspects of the overall problem in the hope that systems developed will reach higher accuracies and better, closer to natural performance. For example, some approaches have attempted to employ different modalities (***modality*** can be described as a manner of communication in a human-computer interaction) to improve accuracy, as Eisenstein and Davis [2006] who explored features of hand gestures combined with a traditional textual model. Results reported for this method suggest a statistically significant

modality

improvement of system performance. Alternatively, Luo et al. [2009] made use of the speaker and turn-taking metadata provided in documents, which led to improvements in system performance. Explorations of this kind, although valuable, are nevertheless limited in their application and often too domain specific or inapplicable to a wider variety of systems. Thus, such advances can in certain respects be described as avoiding a more thorough consideration of the 'core linguistic' contributions to CR.

There are also approaches that attempt to improve the coreference resolution process by exploring new sets of linguistic features. Such a large-scale expansion of feature sets allows the inclusion of more sophisticated linguistic knowledge [Mayfield et al., 2009, Ng and Cardie, 2002a]. Sasano et al. [2007], who work on Japanese for example, report that a good knowledge of language synonyms is required for effective CR approaches for that language.

All these approaches target different aspects of the coreference resolution process and, as a consequence, there are still only a few full CR systems publicly available. The development of such systems is a considerable engineering effort, which itself results in systems that either are applicable to only one language following the methodology proposed by Soon et al. [2001] or Steinberger et al. [2007], or which can only perform partial resolution (e.g., pronoun resolution), as in JAVARAP described by Qiu et al. [2004].

Only recently has a highly modular toolkit, BART by Versley et al. [2008b], become available for the purpose of developing coreference applications. This tool provides the possibility to develop CR systems for integration in further applications. This is particularly valuable in that it makes it possible for researchers with main interests in other areas to still use CR solutions. Such publicly available general toolkits have also been developed for other CL areas (such as Word Sense Disambiguation for example) marking a significant maturation of the field. In the area of CR, however, there are still no publicly available systems that are able to carry out the whole pipeline of necessary procedures, starting with raw text and producing the final semantic interpretation, and all this for multiple languages. Such systems demand considerable effort and also a combination of the various approaches and techniques developed previously in order to improve different aspects of coreference resolution in combination.

2.3.4 *Resources and Evaluation for Coreference Resolution*

As the definition of machine learning itself indicates, collections of data, and more specifically linguistically annotated data, are needed in order for those automated approaches to be achievable. Moreover, supervised methods require those collections to be of a considerable size in order to accomplish competitive performance. Once trained and evaluated, two distinct systems can only be objectively compared if they are trained and tested on the exact same dataset, which poses the additional requirement that those datasets be freely available.

2.3.4.1 *Available Resources*

As Elango [2005] reports, there were only a few standard datasets available at that time, which were distributed during the Message Understanding Competition (MUC) evaluation exercises (MUC-6 [Grishman and Sundheim, 1995] and MUC-7 [Hirschman and Chinchor, 1997]). Yet the author also expressed the need for a wider variety of annotated data because the available datasets represented only narrow domains[2] which makes the development of general-purpose CR systems difficult.

Later evaluation exercises such as ACE [Doddington et al., 2004] and ARE [Orăsan et al., 2008] were undertaken. Additionally, multiple domain specific corpora such as the GENIA corpus on the biomedical domain were annotated with coreference information [Ohta et al., 2002]. Advancing the available datasets allows for more detailed analysis of the CR phenomenon and its issues as for example in the analysis presented by Stoyanov et al. [2009] on various CR subtasks and their overall effect on the CR process for both MUC and ACE datasets. One of the valuable contributions of Stoyanov et al. [2009] enabled by the additional data is the evaluation of state-of-the-art resolvers of different types of anaphora. The gained knowledge was further used to design a measure that can provide a good estimate of the performance of a given resolver for a new dataset.

Only a few years later a new collection of data was released that aimed to provide datasets for more than the English language. The SemEval-2010 task 1: Coreference Resolution in Multiple Languages[3] [Recasens et al., 2010] included various domains for six languages – Catalan, Dutch, English, German, Italian and Spanish. Covering this range of languages, the SemEval-2010 task 1 was the first to enable easier development and objective evaluation of multilingual CR systems. Since it is the purpose of this work to discuss the multilingual issues in CR, we devote more attention to the description of this task in section 3.1.1.

Up to 2011, all datasets were concentrating on noun phrase CR. It was the CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes[4] [Pradhan et al., 2011] that provided a dataset for English not restricted to noun phrases or a given set of entity types. A year later the enterprise was extended with two further languages (Arabic and Chinese) during the CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes [Pradhan et al., 2012]. Again, for the multilingual purpose of our work, we include further details on this task in section 3.1.2.

[2] MUC-6 was assembled from news reports in the domain of "Negotiation of Labor Disputes and Corporate Management Succession" and MUC-7 on news reports in the domain of "Airplane Crashes, and Rocket/Missile Launches".

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

[4] http://conll.cemantix.org/2012

2.3.4.2 *Evaluation Metrics*

Evaluating CR system performance has proven to be highly challenging [Pradhan et al., 2011, Recasens et al., 2010]. In general, evaluation of system performance is represented in terms of recall, precision and their harmonic mean, the F-measure/F-score. ***Precision***, or ***P***, (see equation (2.1)) represents the ratio of the correct answers given by the system to the number of all given answers. ***Recall***, or ***R***, (see equation (2.2)) is the percentage of instances from the test set for which the systems gives an answer. Their harmonic mean, the ***F-score***/***F-measure***/***F*** (see equation (2.3)) represents the final system score.

precision
P
recall
R
F-score
F-measure
F

$$\text{precision} = \frac{|\text{correct answer} \cap \text{given answer}|}{|\text{given answer}|} \tag{2.1}$$

$$\text{recall} = \frac{|\text{correct answer} \cap \text{given answer}|}{|\text{correct answer}|} \tag{2.2}$$

$$\text{F} = \frac{2 \cdot (\text{precision} \cdot \text{recall})}{\text{precision} + \text{recall}} \tag{2.3}$$

There are four commonly used evaluation metrics that provide distinct implementations of precision and recall that have been employed in the last decade: Message Understanding Competition (MUC) metric [Vilain et al., 1995], B^3 [Bagga and Baldwin, 1998], Constrained Entity Alignment F-Measure (CEAF) [Luo, 2005], and BiLateral Assessment of Noun-Phrase Coreference (BLANC) [Recasens and Hovy, 2011]. Diverse variations of those metrics, as the improvement of B^3 and CEAF proposed by Cai and Strube [2010] were also explored. As Cai and Strube [2010] note, the comparison between various systems evaluated with those metrics has been a hard task on its own, for these systems were not necessarily evaluated on the exact same metric or on different versions of one metric. Another problem for this comparison to be objective, as the authors show, is the fact that the different metrics consider different ways of computing the correctness of the resulting coreference links.

In order to avoid the above mentioned drawbacks Recasens et al. [2010] used a combination of MUC, B^3, CEAF and BLANC for the SemEval-2010 evaluation, while [Pradhan et al., 2011] used the same set of metrics apart from BLANC for the CoNLL-2011 and CoNLL-2012 evaluations. The evaluations of the methods presented in this work will be consistent with the evaluations used for the aforementioned shared tasks and datasets on which we employ the methods so that further comparability can be achieved.

MUC [Vilain et al., 1995] is one of the first metrics used for evaluating coreference within the MUC-6 and MUC-7 evaluation tasks. It is a link-based metric. It

key data
key
response data

compares the links between the equivalence classes in the ***key data***, also known as ***key*** (the data containing the answers) and ***response data*** (the system output) by estimating the minimal set of actions needed to transform the response classes to their corresponding key classes. MUC calculates the number of links that are present in both key and response – obtained by the difference between the set of existing links in the equivalence classes in the key S_i and the links existing in the partitions relative to those classes $p(S_i)$ in the response. To compute recall (as shown in equation (2.4)), MUC divides this number by the minimum number of correct links that are needed to form the equivalence classes in the key data.

$$R = \frac{\sum(|S_i| - |p(S_i)|)}{\sum(|S_i| - 1)} \tag{2.4}$$

For further clarification, the relative partition $p(S)$ of a given equivalence set S is derived by the unification of the sets gained by the intersection of S with the equivalence sets included in the response that overlap with S.

In contrast, precision considers the difference between the links in the equivalence classes in the response S'_i and the links existing in the partitions relative to them $p(S'_i)$ in the key, divided by the minimum number of correct links required to gain the equivalence classes in the response data. The complete formula for calculating precision within the initial MUC metric is given in equation (2.5).

$$P = \frac{\sum(|S'_i| - |p'(S'_i)|)}{\sum(|S'_i| - 1)} \tag{2.5}$$

All following explorations of evaluation metrics [Bagga and Baldwin, 1998, Luo, 2005, Recasens and Hovy, 2011] report that MUC has two major drawbacks that need to be resolved. First, the authors address the tendency of MUC to be more compliant with overmerged equivalence classes resulting from the consideration of the minimal number of needed links. This is a consequence of the definition of MUC to penalize each error with an equal precision point. The second disadvantage of MUC, as indicated by [Bagga and Baldwin, 1998, Luo, 2005, Recasens and Hovy, 2011], is the fact that MUC is completely insensitive to the presence of singletons in the response data.

These two issues lay the grounds for the subsequent attempts to overcome the shortcomings by accounting for them in appropriate ways. One such approach is the development of the B^3 metric.

B-CUBED [Bagga and Baldwin, 1998], or B^3, is a mention-based metric that addresses the favorable behaviour of MUC towards overmerged entities by calculating precision and recall separately for each distinct mention. Unlike

MUC, B^3 also includes singleton mentions in its computation. In the latter, recall (see equation (2.6)) is obtained by calculating the weighted average of the separate equivalence class recalls. According to Bagga and Baldwin [1998], weights are defined by the number of entities in a class. The authors define precision analogically to MUC by reversing the key and answer sets in the formula, as presented in equation (2.7).

$$R = \sum_{i=1}^{n} \frac{|S_i|}{\sum_{j=1}^{n} |S_j|} R_i \tag{2.6}$$

$$P = \sum_{i=1}^{n} \frac{|S'_i|}{\sum_{j=1}^{n} |S'_j|} R'_i \tag{2.7}$$

In an attempt to overcome the problem that MUC faces with singletons, B^3 shows a countereffect, namely exceptionally increased overall scores when singletons are present in the response data. Both MUC and B^3 thus fail to perform evaluation in a completely intuitive way, which led to further investigations of the matter.

CEAF [Luo, 2005] is an enhancement of the B^3 approach. According to Recasens and Hovy [2011], B^3 also provides counterintuitive evaluation of the coreference phenomenon because every entity may be used more than once when the alignment between the key and response is achieved. Luo [2005] then defines a similarity function Φ which is the sum of similarities of all aligned entity pairs. Depending on the exact definition of Φ, CEAF can be used as either a mention-based or an entity-based metric. Nevertheless, recall is computed by acquiring the similarity value for the optimal alignment divided by the sum of the entity self-similarites in the key set (see equation (2.8)), while precision considers the entity self-similarities in the response set (see equation (2.9)).

$$R = \frac{\Phi(g^*)}{\sum_i \phi(R_i, R_i)} \tag{2.8}$$

$$P = \frac{\Phi(g^*)}{\sum_i \phi(S_i, S_i)} \tag{2.9}$$

As Recasens and Hovy [2011] further report, CEAF still fails to provide an acceptable solution to the singleton problem raised by both MUC and B^3, which also leads to boosted scores when singletons are present in the response data. Moreover, if an entity from the response is not aligned properly with

a key entity, a potential correctly identified coreference link will be ignored by CEAF. The evaluation presented by Stoyanov et al. [2009] also remarks that entities are weighted equally, independently of their size in terms of the number of mentions they contain.

BLANC [Recasens and Hovy, 2011] employs an implementation of the Rand index presented in [Rand, 1971]. BLANC also aims at targeting the known shortcomings of the previous metrics. It calculates both precision and recall based on both coreference and non-coreference links. The overall recall, defined as BLANC-R (see equation (2.12)) is defined by the arithmetic mean between the recall of coreference R_c (presented in equation (2.10)) and non-coreference links R_n (see equation (2.11)). R_c is obtained by the consideration of all right coreference links *rc* towards their sum with all wrong non-coreference links *wn*. R_n, by analogy, calculates the right non-coreference links *rn* divided by their sum with all wrong coreference links *wc*.

$$R_c = \frac{rc}{rc + wn} \tag{2.10}$$

$$R_n = \frac{rn}{rn + wc} \tag{2.11}$$

$$BLANC - R = \frac{R_c + R_n}{2} \tag{2.12}$$

The overall precision, also indicated as BLANC-P (see equation (2.15)), is as well the arithmetic mean of precision considering coreference links P_c (equation (2.13)) and precision gained from non-coreference links P_n (see equation (2.14)). P_c is the right coreference links from all coreference links, while P_n is gained by the same combination but considering non-coreference links.

$$P_c = \frac{rc}{rc + wc} \tag{2.13}$$

$$P_n = \frac{rn}{rn + wn} \tag{2.14}$$

$$BLANC - P = \frac{P_c + P_n}{2} \tag{2.15}$$

Unlike MUC, B^3 and CEAF, BLANC does not calculate precision as shown in equation (2.3) but uses the latter to obtain F-features for both F-score for

coreference links F_c (see equation (2.16)) and F-score for non-coreference links F_n (see equation (2.17)) that consider the respective precision and recall figures. Then, the overall F-score, defined as BLANC (see equation (2.18)), is the arithmetic mean of F_c and F_n.

$$F_c = \frac{2P_cR_c}{P_c + R_c} \tag{2.16}$$

$$F_n = \frac{2P_nR_n}{P_n + R_n} \tag{2.17}$$

$$\text{BLANC} = \frac{F_c + F_n}{2} \tag{2.18}$$

The difference in the calculation of the BLANC metric also leads to the fact that the final score can be lower than both BLANC-P and BLANC-R, unless they are both considerably high. As Recasens and Hovy [2011] also reports, BLANC manages to account well for the singleton problem of earlier metrics. Yet, they also note that in specific cases (e.g. when the key contains multiple non-coreference links and only one coreference link and the response includes only non-coreferent ones) the BLANC metric will put an equal weight on the one coreferent link in comparison to all non-coreferent ones leading to an overall score not higher than 50%. This is a major drawback of the BLANC metric, which can be seen by the results we report further on in chapter 5, chapter 6 and chapter 7, because the figures that the metric reports are seldom far from 50%, which renders the scores rather uninformative.

2.3.4.3 *Exemplifying the Problems*

Presenting the definitions and all known drawbacks of all four evaluation metrics in section 2.3.4.2 points out two basic flaws that the majority of the metrics cannot appropriately overcome. The first, is their evaluation of equivalence classes that consist of multiple real-world entities (overmerged entities) or its opposing case in which all classes consist of one single member representing one single entity. The second consideration is the behaviour of the metrics when singleton mentions are present in the response. In the current section, we will show examples of those issues and list the scores that each of the metrics achieves in both cases in order to provide a better understanding of their behaviour.

OVERMERGED ENTITIES As presented in Kübler and Zhekova [2011] there are two baselines that can be used in order to expose the deficiencies of the evaluation metrics when overmerging of equivalence classes is concerned. The

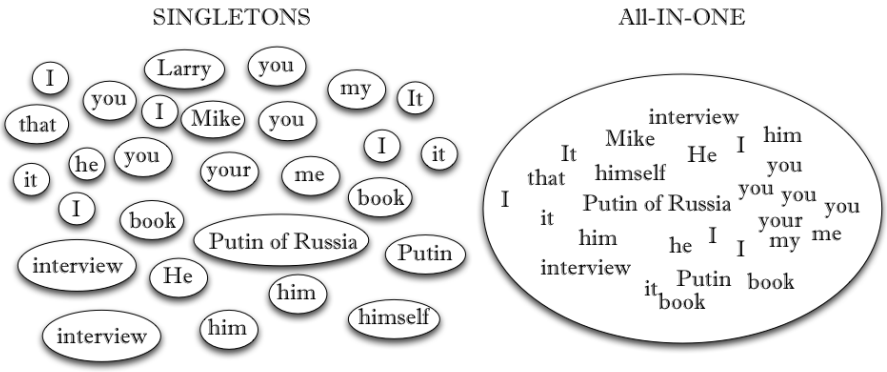

Figure 2.7: All mentions from example (27) on page 25 represented as separate entities (when the *singletons* baseline is used) and building a single entity (when the *all-in-one* baseline is employed).

first considers all mentions to represent a class on their own (all mentions being singletons), referred to as *singletons* baseline, and the second merges all mentions in one equivalence class, used further as *all-in-one* baseline. Let us return again to the text in example (27). As figure 2.3 shows, there are altogether seven different equivalence classes or in other words seven different entities that are marked in the key dataset. A representation of those classes within the *singletons* and *all-in-one* baselines is offered in figure 2.7.

Considering the two baselines, Recasens et al. [2010] assessed the performance of the four evaluation metrics for the SemEval-2010 task 1 (see section 3.1.1) English dataset. The results that the authors reported are listed in table 2.3.

What the figures in table 2.3 show is that there is an exceedingly high variation in the performance of all metrics for both baselines. When all mentions are marked as singletons, or in other words, when there is absolutely no coreference information present in the response set, the MUC metric does not reward any points while CEAF and B^3 report an exceedingly boosted

	MUC			CEAF			B^3			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	BLANC
SINGLETONS	0.0	0.0	0.0	71.2	71.2	71.2	71.2	100	83.2	50.0	49.2	49.6
ALL-IN-ONE	100	29.2	45.2	10.5	10.5	10.5	100	3.5	6.7	50.0	0.8	1.6

Table 2.3: Baseline scores according to the two baselines (*singletons* and *all-in-one*) for the English data set in the SemEval-2010 task 1 evaluated by the *MUC*, *CEAF*, *BCUB* and *BLANC* metrics.

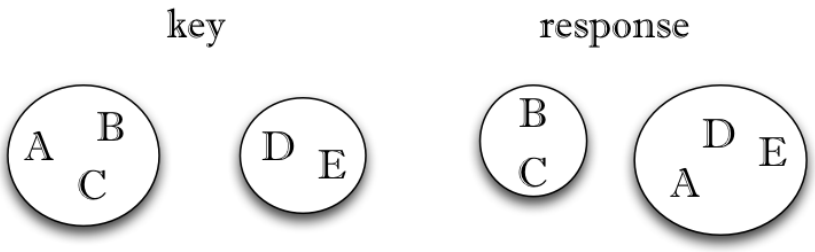

Figure 2.8: Setting 1 from our toy example representing two equivalence classes in the key set *{A, B, C}* and *{D, E}*, and two in the response *{B, C}* and *{A, D, E}*.

performance of 71.2% and 83.2% respectively. The BLANC metric keeps the performance in the range of 50% reporting a BLANC score of 49.6%. On the other hand, merging all mentions in one single class, the *all-in-one* baseline, leads to a relatively boosted performance of MUC – 45.2%, while the rest of the metrics do not increase beyond 10.5%. These results show that evaluating a system via any of these metrics separately cannot lead to objective results, because systems that are well-adapted at identifying all the mentions, but not the coreference links between them, will be performing exceedingly well judged by the CEAF and B^3, while their performance according to MUC will be close to 0. Furthermore, systems that have the tendency to overmerge entities will receive higher scores by MUC, but not by the rest of the evaluation metrics.

SINGLETONS The second major drawback of MUC, B^3, CEAF and BLANC is their behaviour against singleton mentions in the response, as shown in the toy example in Kübler and Zhekova [2011]. In the base setting (referred to as setting 1), two equivalence classes are included in the key set – $S_{1k} = \{A, B, C\}$ and $S_{2k} = \{D, E\}$. For the same setting, one link error is introduced in the response that attaches mention *A* erroneously to the wrong class, resulting in the following equivalence classes: $S_{1r} = \{B, C\}$ and $S_{2r} = \{A, D, E\}$. For clarification, these sets are presented in figure 2.8.

Further, we included an additional error in the response by introducing a new class with a single member $S_{3r} = \{Y\}$ (see figure 2.9), referred to as setting 2. The last setting that we use as an example, setting 3, contains a singleton in the key $S_{3k} = \{X\}$ (see figure 2.10).

In order to assess setting 1, 2 and 3 we use the scoring software presented by the SemEval-2010 task 1 (see section 3.1.1). The results from this evaluation are listed in table 2.4. The figures show that within the SemEval-2010 task 1 all metrics, apart from MUC, were sensitive to the presence of singletons in both the key and the response. We note that the difference between setting 1 and 2

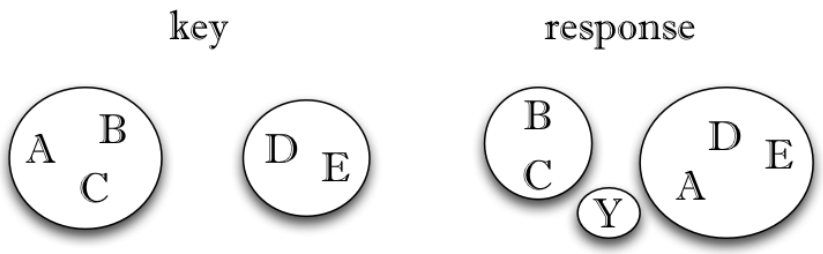

Figure 2.9: Setting 2 from our toy example representing two equivalence classes in the key set *{A, B, C}* and *{D, E}*, and three classes in the response *{B, C}*, *{Y}* and *{A, D, E}*.

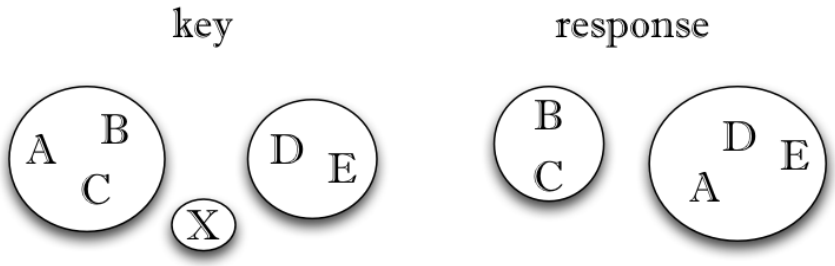

Figure 2.10: Setting 3 from our toy example representing three equivalence classes in the key set *{A, B, C}*, *{X}* and *{D, E}*, and two classes in the response *{B, C}* and *{A, D, E}*.

is the presence of the singleton Y in response; the actual coreference links are the same in both variants. However, B^3, CEAF and BLANC report decreases in scores. When a singleton is introduced in the key (setting 3), MUC and BLANC do not report variation in scores with respect to setting 1. However, B^3 indicates a small decrease, while for CEAF there is no difference between having an additional singleton in the key or in the response.

This exploratory investigation shows that singletons are exceedingly important for the objective evaluation of coreference. They should be taken into account in both key and response in order to achieve a better alignment of the detected mentions and the links between them. However, the datasets provided by the CoNLL-2012 shared task (see section 3.1.2) did not include singleton mentions, for the latter are not contained in the annotation schemes of the targeted languages. This fact poses a problem for the overall evaluation of the participating systems, which led to the adaptation of the scoring software by

	MUC			B^3			CEAF			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
1	66.7	66.7	66.7	73.3	73.3	73.3	80.0	80.0	80.0	58.3	58.3	58.3
2	66.7	66.7	66.7	73.3	61.1	66.7	80.0	53.3	64.0	51.8	52.3	49.8
3	66.7	66.7	66.7	61.1	73.3	66.6	53.3	80.0	64.0	58.3	58.3	58.3

Table 2.4: Coreference scores on Setting 1, 2 and 3 from our toy example achieved by the scoring software provided in SemEval-2010 task 1 evaluated by the *MUC*, *CEAF*, *BCUB* and *BLANC* metrics.

	MUC			B^3			CEAF			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
1	66.7	66.7	66.7	73.3	73.3	73.3	80.0	80.0	80.0	58.3	58.3	58.3
2	66.7	66.7	66.7	73.3	73.3	73.3	80.0	80.0	80.0	58.3	58.3	58.3
3	66.7	66.7	66.7	77.8	77.8	77.8	86.7	86.7	86.7	65.9	65.9	65.9

Table 2.5: Coreference scores on Setting 1, 2 and 3 from our toy example achieved by the scoring software provided in CoNLL-2012 shared task evaluated by the *MUC*, *CEAF*, *BCUB* and *BLANC* metrics.

the SemEval-2010 task 1 to account better for the presence of singletons in the response. A reevaluation of the three toy settings with the scoring software (see section 3.1.2) from the CoNLL-2012 shared task, the scores from which are listed in table 2.5, shows that all metrics are rendered insensitive to singletons in the response. Yet, additional singletons in the key boost the scores for all metrics, except for MUC. This fact is not problematic when data from the CoNLL-2012 shared task is used, since no singletons were included in the key, yet, it should be taken into consideration if that version of the software is used on different datasets.

2.4 SUMMARY AND CONCLUSION

In the current chapter, chapter 2, we presented coreference resolution by introducing various aspects of the concept. We discussed both rule-based and machine learning approaches to CR and outlined their advantages and disadvantages in their employment for the task. The chapter highlighted numerous new and innovative attempts for the improvement of coreference resolution and provided an overview of existing resources and evaluation standards for this task.

The depicted information will serve as the basis for our further discussion in the thesis on redefining this task into a more complex enterprise. Yet, our aim was to provide only the relevant to our investigation details. If more

elaborate description and delineation of the concept is needed, we advocate the exploration of the referenced relevant literature.

CHAPTER

3

MULTILINGUAL COREFERENCE RESOLUTION

In the increasingly complex and rapidly changing world, the need for robust and efficient methods for Natural Language Processing (NLP) applications that are flexible and that lead to good and stable system performance is rapidly growing. With the advances of science and technological development as well as the boosted access to information, software and ever growing communication, the demand for multilingual applications is more than ample. Modern multilingual systems build a bridge between the already widely available knowledge and the monolingual end-user. One well known multilingual project for example is Wikipedia[1] – a multilingual, web-based, free-content encyclopedia. This easily accessible resource allows for textual content to be entered and used across language boundaries due to its hyperlinked nature. Yet, there is no guarantee for the user that the content he or she is searching for will be available in a language that the user can actually speak or understand. Further multilingual assistants as Google Translate[2] for example can be made use of in order for that content to be understandable. Yet, multilingual approaches often

[1] http://www.wikipedia.org
[2] http://translate.google.com

carry an immense engineering and implementation effort with them. For this reason, it is necessary to shed more light on the problem of multilinguality for the subject of our interest – coreference resolution.

Thus, in the current chapter we will continue beyond the notion of simple CR and revise the advances of that field into more than one targeted language and in this way we will delineate the complex task of Multilingual Coreference Resolution (MCR). We will first review all initial approaches to MCR (see section 3.1) and then discuss the basic necessities as well as pressing issues with respect to MCR-based approaches (see section 3.2). Section 3.3 offers concluding remarks.

multilingual coreference resolution

3.1 CONTEMPORARY MULTILINGUAL COREFERENCE RESOLUTION

Multilingual Coreference Resolution has been gaining a great amount of interest in the CL community for almost two decades now. It was first Aone and McKee [1993] who presented a data-driven architecture for language-independent anaphora resolution that was capable of functioning on any language and still was robust, easily extendable and trainable. Mitkov [1999b] proposed a knowledge-poor approach to AR that was initially developed and tested for English and then further extended to Polish and Arabic as well as Finnish, Russian and French. Yet, as the author notes, by that time there were already several approaches on various languages such as: French [Popescu-Belis and Robba, 1997, Rolbert, 1989], German [Dunker and Umbach, 1993, Fischer et al., 1995, Leass and Schwall, 1991, Stuckardt, 1997], Japanese [Mori et al., 1997, Nakaiwa and Ikehara, 1992, 1995], Portuguese [Abraços and Lopes, 1994], Swedish [Fraurud, 1988] and Turkish [Tin and Akman, 1994]. Later on numerous other languages were added to that list: Bulgarian [Grigorova, 2011, Tanev and Mitkov, 2002], Catalan [Mayol, 2006, Potau, 2008], Dutch [Hendrickx et al., 2008, Hoste, 2005], Italian [Poesio et al., 2010, Sorace and Filiaci, 2006], Spanish [Palomar and Martínez-Barco, 2001, Potau, 2008], etc.

However, the cases given above were and still are only a very small portion of the AR and CR research, because it is on English that the most effort from the CL community is concentrated. This is explained by the fact that linguistic information, annotations and analysis tools are easily available for English, but not for less resourced languages such as Bulgarian and Portuguese, for example (see section 1.2). A multilingual approach dependent on deeper semantic and syntactic analysis will inevitably prove to be inapplicable when that information is not accessible for every targeted language. Yet, Mitkov [1999b] also points out that the endeavour of concentrating on a multilingual approach is bound to be directed towards circumventing more complex syntactic, semantic and discourse analysis. After the two multilingual approaches [Aone and McKee, 1993, Mitkov, 1999b], there were only a few other methods concentrating on more than one language at a time: [Harabagiu and Maiorano, 2000, Luo

and Zitouni, 2005]. It was not until the introduction of two highly important events for multilingual approaches that further methods and systems featuring multiple languages simultaneously were presented:

1. SemEval-2 task 1: Coreference Resolution for Multiple Languages, further referred to as SemEval-2[3] (see section 3.1.1)
2. CoNLL 2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, further referred to as CoNLL 2012[4] (see section 3.1.2)

Both events were organized as shared tasks and targeted the development of coreference resolution systems that can be applied to the languages addressed by the competitions. Both tasks lay foundational ground in MCR and play a central role for our further discussion. Thus, we devote the following sections to their introduction and the main aspects of their proceedings. Further, in section 3.2 we will delineate the key problems to multilingual CR, because these are the issues that serve as basis to the research in this work. Pradhan et al. [2012] as well as Recasens et al. [2010] provide more detailed information about the proceedings of both tasks.

3.1.1 SemEval-2 *task 1: Coreference Resolution for Multiple Languages*

The first multilingual endeavour was approached in 2010 by the SemEval-2 task 1: Coreference Resolution for Multiple Languages Recasens et al. [2010]. This was the first opportunity for MCR systems to be objectively reviewed and comparatively evaluated. A new and highly innovative pursuit, as this aimed at answering various questions (with respect to CR applied on multiple languages) that were still open to the research community. Because of the fact that there were hardly any systems able to work on more than one language, SemEval-2 planned to estimate the effort needed to transform a monolingual system to a multilingual one. As Recasens et al. [2010] report, it was unclear how much language specific modifications would be needed for a competitive performance as well as how important general linguistic annotations as morphology, syntactic and semantic layers are to that performance. Since manually annotated data, also called ***gold data*** or ***gold standard***, is exceptionally hard and expensive to obtain, it was necessary to investigate the difference between the system performance on gold data vs. auto data. ***Auto data*** is noisier and inferior to gold data, because it is collected by the use of various computational tools. As we presented in section 2.3.4.2, evaluation of CR systems is still highly difficult. Thus, another question Recasens et al. [2010] were interested in was the overall effect of the various evaluation metrics (MUC, CEAF, B^3, BLANC) on the

gold data
gold standard
auto data

[3] http://stel.ub.edu/semeval2010-coref
[4] http://conll.cemantix.org/2012

ranking, comparison and altogether the representation of the performance of the participating systems. It is those and many other questions with respect to multilinguality that we focus on in the context of our work. We will look into the full coreference resolution pipeline within the SEMEVAL-2 and the CONLL 2012 shared tasks and analyze the results from the approach we make use of (see chapter 4).

3.1.1.1 *Data*

The SEMEVAL-2 shared task targeted six different languages: Catalan, Dutch, English, German, Italian and Spanish. The six languages cover two language families – the Romance language family (with representatives: Catalan, Italian and Spanish) and the Germanic language family (with representatives: Dutch, English and German). As Recasens et al. [2010] present, the datasets were assembled based on the availability of distinct corpora and annotation tools for the six approached languages that we summarize in the following paragraphs.

CATALAN AND SPANISH The Catalan and Spanish data was extracted from the AnCora corpora [Recasens and Martí, 2010], which mainly contains newswire texts annotated manually for arguments and thematic roles, predicate and semantic classes, named entities, WordNet[5] nominal senses as well as coreference. A ***Named Entity (NE)*** can be categorized as atomic element in text according to a predefined list of categories. NEs can be of various different types: proper names, locations, expressions of times or quantities, monetary values, percentages, etc. Additionally, automatic annotations for lemmas and Part of Speech (POS) information were acquired via the FreeLing[6] open source suit of language analyzers [Padró and Stanilovsky, 2012]. The dependency structure and predicate semantic roles were achieved via the syntactic-semantic JointParser[7] [Lluís et al., 2009]. An example sentence for each of the two languages, Catalan and Spanish, is provided in table A.1 on page 248 and table A.6 on page 253 respectively.

named entity

DUTCH The dataset for the Dutch language was assembled from the KNACK-2002 corpus [Hoste and Pauw, 2006], which also contains newswire texts. The annotations in the texts include manually identified coreference relations and semi-automatically annotated POS, phrase chunks and named entities. The automatic part of the annotation of lemmas, POS, and named entities was acquired by the memory-based shallow parser for Dutch, presented in [Daelemans et al., 1999]. The parser was developed by the Induction of Linguistic Knowledge Research Group and is available from their website[8]. The dependency informa-

[5] http://wordnet.princeton.edu
[6] http://nlp.lsi.upc.edu/freeling
[7] http://nlp.lsi.upc.edu/jointparser/demo
[8] http://ilk.uvt.nl

tion was labeled by the Alpino[9] parser introduced in [Van Noord et al., 2006]. An example sentence for Dutch is given in table A.2 on page 249.

ENGLISH The English part of the SEMEVAL-2 shared task dataset was taken from the OntoNotes Release 2.0 corpus [Pradhan et al., 2007]. This release consists of newswire and broadcast news annotated with Penn Treebank[10] syntactic annotations, Penn Propbank[11] predicate argument structures, named entities, word senses and coreference information. Automatic annotations for lemmas and POS information were generated using the SVMTagger[12] presented in [Giménez and Màrquez, 2004]. The syntactic-semantic JointParser[7] parser [Lluís et al., 2009] was again used for the dependency structure and predicate semantic roles. An example sentence for English can be found in table A.3 on page 250.

GERMAN For German the data was extracted from the Tüba-D/Z corpus [Hinrichs et al., 2005], which is a treebank of newswire texts with syntactic and coreference annotations. Lemmas, POS, morphological and dependency information were also automatically annotated. Lemmas were labeled by the TreeTagger[13] [Schmid, 1995]. POS tags and morphological information were predicted by the RFTagger[14] introduced in [Schmid and Laws, 2008], while the dependency layer was constructed by the MaltParser[15] presented in [Hall and Nivre, 2008]. A German excerpt from the data is shown in table A.4 on page 251.

ITALIAN The collection for Italian was acquired from the LiveMemories corpus [Rodríguez et al., 2010] built up of Wikipedia, blogs, newswire and dialogues. The data is annotated for coreference, agreement and named entities on the basis of automatic parses. The TextPro[16] suit of modular NLP tools was used for the lemmas and POS annotations and the MaltParser[15] [Hall and Nivre, 2008] was employed for the acquisition of the dependency information. An example sentence from the Italian dataset can be found in table A.5 on page 252.

A complete summary of the size of the used datasets per language, as given in [Recasens et al., 2010], is shown in table 3.1. The figures are separated for the training, development and test parts of the datasets and counts are listed for the number of documents, sentences and tokens within each part.

[9] http://www.let.rug.nl/vannoord/alp/Alpino
[10] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99T42
[11] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2004T14
[12] http://www.lsi.upc.edu/~nlp/SVMTool
[13] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger
[14] http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger
[15] http://www.maltparser.org
[16] http://textpro.fbk.eu

	training			development			test		
	docs	sents	tokens	docs	sents	tokens	docs	sents	tokens
Catalan	829	8,709	253,513	142	1,445	42,072	167	1,698	49,260
Dutch	145	2,544	46,894	23	496	9,165	72	2,410	48,007
English	229	3,648	79,060	39	741	17,044	85	1,141	24,206
German	900	19,233	331,614	199	4,129	73,145	136	2,736	50,287
Italian	80	2,951	81,400	17	551	16,904	46	1,494	41,586
Spanish	875	9,022	284,179	140	1,419	44,460	168	1,705	51,040

Table 3.1: A full summary of the size of the datasets for all six languages within the SemEval-2 shared task. The numbers are separated for the training, development and test sets and counts are provided for the number of documents (docs), sentences (sents) and tokens (tokens).

As can be seen, the datasets differed to a great extent in length, with German having the largest set and Italian the smallest. The size of the provided data is important, because a machine learning approach, as the one that we will use in our investigation (see chapter 4), needs a large number of examples to train on.

3.1.1.2 *Task Definition*

Unlike previous evaluation exercises, such as ACE [Doddington et al., 2004] and ARE [Orăsan et al., 2008], the task description of the SemEval-2 shared task given in [Recasens et al., 2010] included the identification of mentions in its definition. The competing systems needed to extract all types of noun phrases (apart from NPs that cannot be referential, such as appositives, expletive NPs, attributive NPs, etc.) and possessive determiners which were regarded as mentions. Singletons are also considered entities and included in the set of *gold* mentions. Both *auto* and *gold* annotation layers were provided for the majority of languages and annotations: No *gold* layers were given for Italian and Dutch, apart from named entities for Italian; German did not include *gold* NEs; None of the datasets but the one for the Dutch language provided *auto* NEs.

The task aimed at the identification of intra-document coreference relations across the identified mentions and their proper clustering into coreference classes. Each class represents a distinct discourse entity.

3.1.1.3 *Data Format*

The format of the data was prepared in a simplified and uniform column-based format. The dataset for each separate language consisted of one single file – one file for the training, one for the development and one for the test data. Since intra-document coreference was the target of the task, the files were divided

```
#begin document <document ID>
<sentence>

<sentence>
...
<sentence>

#end document <document ID>
...
#begin document <document ID>
<sentence>

<sentence>
...
<sentence>

#end document <document ID>
```

Figure 3.1: The structure of the *train/devel/test* files provided for each of the six languages in the SEMEVAL-2 shared task. The information listed within < > is a placeholder for the actual data.

into documents. This structure is visualized in figure 3.1. Specific examples including one sentence for each of the task languages are provided in A.1.

Each document consists of *n* sentences separated by empty lines. The sentences were represented by their tokens listed each on a distinct line. The latter is shown in figure 3.2. The various columns contained the diverse layers of linguistic annotations made available by the task. The actual information listed in the columns is given in table 3.2 on page 51. The two types of annotations, *auto* and *gold*, were appended in an alternating order which is also made visible by the descriptions provided in table 3.2[17]. In case the information

```
<token#1 column#1>   <token#1 column#2>   <token#1 column#3> ...
<token#2 column#1>   <token#2 column#2>   <token#2 column#3> ...
<token#3 column#1>   <token#3 column#2>   <token#3 column#3> ...
...
```

Figure 3.2: The structure of the sentences building the documents provided for all six languages in the SEMEVAL-2 shared task. The information listed within < > is a placeholder for the actual data.

[17] http://stel.ub.edu/semeval2010-coref/datasets

in the column is not made available or it is irrelevant to the given token, an underscore was used as a placeholder.

The coreference annotation was represented in a bracketed notation, the so called open-close notation, which uses "(<entityID>" to signify that the token is the beginning of a mention that refers to the entity identified by the <entityID>. The "<entityID>)", respectively, denotes the end of that mention. Mentions that are marked by the same <entityID> are coreferent, because they refer to the same entity. Yet, this is only true for mentions that are situated in the same document. Mentions across documents that share identical <entityID> are not coreferent. The same is also true for mentions across languages that share the same <entityID>.

3.1.1.4 *Evaluation*

The SemEval-2 shared task included four different evaluation settings: *gold-closed*, *auto-closed*, *gold-open* and *auto-open*. Those variations regulated the use of *gold* vs. *auto* annotations and external tools and resources for preprocessing. The groups are to be read as follows:

gold-closed – *gold* linguistic annotations must be used by the systems and no external tools and resources are allowed for additional preprocessing.

auto-closed – *auto* linguistic annotations must be used by the systems and no external tools and resources are allowed for additional preprocessing.

gold-open – *gold* linguistic annotations must be used by the systems and external tools and resources are allowed for additional preprocessing.

auto-open – *auto* linguistic annotations must be used by the systems and external tools and resources are allowed for additional preprocessing.

The SemEval-2 shared task did not release system rankings according to the results submitted by all participating teams. Furthermore, as Pradhan et al. [2012] report, because of the low number of contributors, the organizers of the task were not able to achieve any strong conclusions. Appendix B.1 lists the full system scores as reported by the SemEval-2 shared task.

3.1.2 CoNLL 2012 *Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes*

The second multilingual task that aimed at resolving coreference relations for more than one language at a time was the CoNLL 2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes [Pradhan et al., 2012]. The task closely followed the framework established by the SemEval-2 shared task. For this reason, similar to the presentation in section 3.1.1, in the

#	type	description
1	ID	word identifiers in the sentence
2	TOKEN	word forms
3	LEMMA	word lemmas (gold standard manual annotation)
4	PLEMMA	word lemmas predicted by an automatic analyzer
5	POS	coarse part of speech
6	PPOS	same as 5 but predicted by an automatic analyzer
7	FEAT	morphological features (part of speech type, number, gender, case, tense, aspect, degree of comparison, etc., separated by the character "|")
8	PFEAT	same as 7 but predicted by an automatic analyzer
9	HEAD	for each word, the ID of the syntactic head ('0' if the word is the root of the tree)
10	PHEAD	same as 9 but predicted by an automatic analyzer
11	DEPREL	dependency relation labels corresponding to the dependencies described in 9
12	PDEPREL	same as 11 but predicted by an automatic analyzer
13	NE	named entities
14	PNE	same as 13 but predicted by a named entity recognizer
15	PRED	predicates are marked and annotated with a semantic class label
16	PPRED	Same as 13 but predicted by an automatic analyzer
*	APREDs	N columns, one for each predicate in 15, containing the semantic roles/dependencies of each particular predicate
*	PAPREDs	M columns, one for each predicate in 16, with the same information as APREDs but predicted with an automatic analyzer.
*	COREF	coreference annotation in open-close notation, using "|" to separate multiple annotations (see more details below)

Table 3.2: The types of linguistic annotations provided for all six languages in the SemEval-2 shared task.

following sections we describe the data sets used for the CoNLL 2012 shared task (see section 3.1.2.1), then we delineate the exact definition of the task (section 3.1.2.2), in section 3.1.2.3 we describe the format of the data in more detail and in section 3.1.2.4 we report on the evaluation procedure.

3.1.2.1 *Data*

As Pradhan et al. [2012] present, the datasets for the task were assembled from the OntoNotes[18] corpus [Hovy et al., 2006] available from the Linguistic Data Consortium[19]. The task provided data for three languages (Arabic, English and Chinese) within three distinct language families: Semitic (Arabic), Germanic (English), Sino-Tibetan (Chinese). The following paragraphs describe the datasets for the separate languages in more detail.

ARABIC The syntactic annotations for the Arabic dataset were acquired based on the guidelines from the Arabic Treebank [Maamouri and Bies, 2004]. Verb proposition was labeled according to the Arabic Proposition Bank [Palmer et al., 2008, Zaghouani et al., 2010] guidelines. Word senses were manually marked without interresource mappings, such as the mapping for English word senses to WordNet. Named entities as well as coreference were also included in the annotation. An example sentence for Arabic can be found under table A.7 on page 254.

ENGLISH For the English part of the data, the syntactic structure was labeled according to a revised version of the English Penn Treebank [Marcus et al., 1993, Babko-Malaya et al., 2006] guidelines. Propositions were labeled via the English PropBank [Palmer et al., 2005, Babko-Malaya et al., 2006] guidelines. Word senses were manually annotated and mapped to the WordNet semantic structure. NEs and coreference information were also added to the annotation. An excerpt from the English dataset is given in table A.8 on page 255.

CHINESE The dataset for Chinese was assembled similarly to those for Arabic and English. The syntactic structure for this language was achieved by the use of the Chinese version of the Penn Treebank [Xue et al., 2005] guidelines. Propositions were extracted according to the Chinese Proposition Bank [Xue and Palmer, 2009] annotation guidelines. Word senses were also manually marked without interresource mappings. Additionally, named entities and coreference were included. An example sentence for Chinese is listed in table A.9 on page 256.

A full summary of the size of the provided datasets for each of the targeted languages is listed in table 3.3. The information is separated for the training,

[18] http://www.bbn.com/nlp/ontonotes
[19] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T03

	training			development			test		
	docs	sents	tokens	docs	sents	tokens	docs	sents	tokens
Arabic	359	7,422	242,702	44	950	28,327	44	1,003	28,371
English	2,802	75,187	1 299,312	343	9,603	163,104	348	9,479	169,579
Chinese	1,810	36,487	756,063	252	6,083	110,034	218	4472	92,308

Table 3.3: A full summary of the size of the datasets for all three languages within the CoNLL 2012 shared task. The numbers are separated for the training, development and test sets and counts are provided for the number of documents (docs), sentences (sents) and tokens (tokens).

development and test subsets of the data and counts are listed for the number of documents, sentences and tokens within each part.

3.1.2.2 *Task Definition*

Although the CoNLL 2012 shared task was highly similar to its multilingual predecessor, the SemEval-2 shared task, it extended the task to the identification not only of intradocument coreference relations between entities, but rather between intradocument coreference relations among entities and events. Consequently, the expected outcome was not in the form of entity chains as before, but in the form of entity/event chains. The introduction of event coreference increased the overall complexity of the CR task that is already challenging for state-of-the-art systems. As a result, most participating teams (e.g. [Björkelund and Farkas, 2012, Martschat et al., 2012]) did not aim at resolving event coreference. Furthermore, the task also included the identification of mentions.

As Pradhan et al. [2012] discuss, providing *auto* and *gold* annotations across all languages and linguistic information types proved more challenging than expected. The lack of resources and availability of NLP tools as well as the time constraint that the organizers faced within the frameset of the shared task led to the fact that some of the annotation layers could not be provided for all targeted languages. As reported in [Pradhan et al., 2012], the following annotations were addressed for the Arabic, English and Chinese languages in this respective order as follows:

SEGMENTATION – *gold*, *gold*, *gold*

LEMMA – *gold*, not applicable, *auto*

PARSE – *auto*, *auto*, *auto*

PROPOSITION – not provided, *auto*, *auto*

PREDICATE FRAME – not provided, not provided, *auto*

WORD SENSE – *auto*, *auto*, *auto*

NAMED ENTITIES – not provided, not provided, *auto*

SPEAKER – not applicable, *gold*, *gold*

3.1.2.3 *Data Format*

The format of the provided datasets again followed closely the guidelines used by the Recasens et al. [2010] (see section 3.1.1.3).

One difference between the format of the data between both tasks is with respect to the distribution/collection of documents within one/multiple files. For the SEMEVAL-2 shared task, a file was used to collect all documents for the training/development/testing data. The CONLL 2012 task included one file for each of the documents.

Another difference concerns the number of columns and type of information provided per column. A complete list of the annotation types and their arrangement into the column system for the CONLL 2012 shared task is given in table 3.4[20].

As described in section 3.1.1.3, the coreference annotation layer is represented in the open-close notation. Again, mentions that share the same `<entityID>` are coreferent, because they refer to the same entity. Still, this is only true for the mentions that are in the same document. Mentions across documents/languages that share identical `<entityID>` are not coreferent. The latter facts should be noted, since throughout our further work we often demonstrate various issues with examples from the data where these peculiarities of the coreference annotation are often to be seen and might cause confusion if not clarified. Thus, in all following examples, only mentions within the same document, same language and sharing the same `<entityID>` should be considered coreferent.

3.1.2.4 *Evaluation*

With respect to the evaluation procedure, the CONLL 2012 task used both *open* and *close* tracks that were also employed for the SEMEVAL-2 task. Additionally, the CONLL 2012 task also provided a supplementary evaluation track. The latter included the use of *gold* mention boundaries (the correct boundaries for each of the mentions in the data, both singletons and coreferent, were provided), *gold* mentions (only the non-singleton mention boundaries were given), *gold* parse (manually annotated linguistic information could be used). Appendix B.3 provides all system scores as reported by the CONLL 2012 shared task.

[20] http://conll.cemantix.org/2012/data.html

#	type	description
1	Document ID	This is a variation on the document filename
2	Part number	Some files are divided into multiple parts numbered as 000, 001, 002, ... etc.
3	Word number	This is the word index in the sentence
4	Word itself	This is the token as segmented/tokenized in the Treebank. Initially the *$_s$kel file contain the placeholder [WORD] which gets replaced by the actual token from the Treebank which is part of the OntoNotes release.
5	Part-of-Speech	Part of Speech of the word
6	Parse bit	This is the bracketed structure broken before the first open parenthesis in the parse, and the word/part-of-speech leaf replaced with a *. The full parse can be created by substituting the asterisk with the "([pos] [word])" string (or leaf) and concatenating the items in the rows of that column.
7	Predicate lemma	The predicate lemma is mentioned for the rows for which we have semantic role information. All other rows are marked with a "-"
8	Predicate Frameset ID	This is the PropBank frameset ID of the predicate in Column 7.
9	Word sense	This is the word sense of the word in Column 3.
10	Speaker/Author	This is the speaker or author name where available. Mostly in Broadcast Conversation and Web Log data.
11	Named Entities	These columns identifies the spans representing various named entities.
12:N	Predicate Arguments	There is one column each of predicate argument structure information for the predicate mentioned in Column 7.
N	Coreference	Coreference chain information encoded in a parenthesis structure.

Table 3.4: The types of linguistic annotations provided for all three languages in the CoNLL 2012 shared task.

3.2 PREDICAMENTS OF MULTILINGUAL COREFERENCE RESOLUTION

In section 3.1 we addressed the remark by Mitkov [1999b] that the endeavour of concentrating on a multilingual approach is bound to be directed to circumventing more complex syntactic, semantic and discourse analysis. Another constraint that multilinguality imposes on the task of CR is the expectation and reliance of the CR pipeline on consistent, uniformly formatted and coordinated data across all languages that the pipeline is working with.

Our multilingual approach is also, to a great extent, based on the assumption that various standard layers of linguistic annotation are provided for each of the languages we target. Working on a single language does not pose many issues with respect to the format and irregularities/errors in the provided annotations. However, this is not the case within a multilingual system, such as the one we use.

Keeping this in mind, the current section discusses the matters with respect to the availability of corpora annotations (see section 3.2.1) and introduces the annotation schemes used for the different languages (section 3.2.2).

3.2.1 *Availability of Corpora Annotations*

The availability of different types of linguistic annotations that can be used for the resolution of coreference as well as corpora annotated for coreference itself are a crucial point in the research on multilingual coreference resolution. The constant work in the direction also makes it exceedingly hard to achieve a good and complete account for all existing tools and resources as well as to have an objective evaluation on their quality and coverage.

For example, let us consider the thirty European languages listed in table 3.5. The table gives an overview of the availability of two basic types of linguistic annotations (POS and syntax) together with the availability of corpora annotated for coreference. The constant development of NLP tools and annotation of corpora, however, makes it very hard to give a complete and objective overview with a comparison of the stage of development, quality, coverage and financial value of the tool/corpus. This is an important issue in research, the optimal solution to which is still not found. META-NET[21] is one of the organizations that put a considerable effort in this direction, which led to the designation and implementation of META-SHARE[22]. META-SHARE is a sustainable network containing information about different repositories of corpora, tools and web services. Yet, even though the information META-SHARE can provide is highly valuable its coverage is still limited. For this reason, the information provided in table 3.5 can be regarded only as an outline of the availability of resources/tools.

[21] http://www.meta-net.eu
[22] http://www.meta-share.eu

language	POS	syntax	coreference
Basque	√	√	-
Bulgarian	√	√	√
Catalan	√	√	√
Croatian	√	√	-
Czech	√	√	-
Danish	√	√	-
Dutch	√	√	√
English	√	√	√
Estonian	√	√	-
Finnish	√	√	-
French	√	√	√
Galician	√	√	-
German	√	√	√
Greek	√	√	-
Hungarian	√	√	√
Icelandic	√	√	-
Irish	√	√	-
Italian	√	√	√
Latvian	√	√	-
Lithuanian	√	-	-
Maltese	√	-	-
Norwegian	√	√	√
Polish	√	√	√
Portuguese	√	√	√
Romanian	√	√	√
Serbian	√	-	-
Slovak	√	√	-
Slovene	√	√	-
Spanish	√	√	√
Swedish	√	√	√

Table 3.5: A list of the availability of POS, syntactic and coreference annotations across the thirty European languages, independent of their stage of development, quality, coverage and financial value. √ indicates that there are existing resources/tools to achieve this layer of annotation, while - denotes its lack.

From the information in table 3.5, we can see that POS information is available across all thirty European languages. This indicates, that POS is a layer of linguistic annotations that has a high chance of being provided within standard data distributions. Syntactic information also adheres to this tendency. However, for three out of all thirty languages (Maltese, Serbian, Slovak) there are no tools or corpora that can provide syntactic analysis. As we show further in our work, syntactic information is highly important and beneficial to the process of coreference resolution. Although it is widely available, we cannot completely rely on the guaranteed presence of syntactic annotation within various datasets across languages.

One of the most crucial annotation layers to the CR task is the actual coreference annotation of the data. Without corpora already annotated for coreference, statistical systems would not be comparable, because they cannot be properly evaluated. As table 3.5 shows, coreference is hardly the most widely distributed and available annotation layer. Because coreference resolution is a significantly complex phenomenon (see chapter 2), often enough there are annotations for coreference that follow different annotation schemes (see section 3.2.2) or only cover the relations partially. For example, the dataset available for Bulgarian is derived from the BulTreeBank[23] [Simov et al., 2002] where coreference is annotated on the sentence level (comprising about 15000 sentences). The relations marked are identity, member-of, subset-of. What this tells us, is that in order for Bulgarian to be included in a shared task, such as SemEval-2 or CoNLL 2012, additional coreference annotations across sentences need also be enclosed.

The size of the context window in which the coreference links are annotated, however, is not the only important issue within standard annotation guidelines. Even the few languages within the SemEval-2 shared task were annotated according to divergent annotation schemes which can pose multiple difficulties in the development of a coreference resolution system that should be able to work equally well with all languages. For this reason, we devote section 3.2.2 on this issue.

3.2.2 *Differences in Annotation Schemes*

One of the reasons that makes working on one single language with respect to the task of multilingual coreference resolution easier is the fact that only one annotation scheme needs to be considered during the development of the coreference model and resolution pipeline. Typologically close languages, such as Catalan and Spanish, may have similar annotation guidelines, which simplifies the problem, but still does not present a general solution. The datasets provided by the two shared tasks used different annotation guidelines across the distinct languages. In order to first identify the mentions that are

[23] http://www.bultreebank.org

potentially coreferent and then detect the appropriate coreference links between those mentions, multilingual systems are based around the assumption that the underlying mentions are defined in a similar way.

Different annotation schemes pose substantial difficulties to the development of heuristic or rule-based approaches to mention detection. Even though machine learning approaches are a lot more flexible, they are also highly affected by the variation in the annotation guidelines. This is so, because ML approaches make use of informative features, yet, the latter can differ depending on the annotation scheme for the given language. For this reason, the current section gives an overview of those schemes pointing out the relations and definitions most important to our work.

3.2.2.1 SemEval-2 *shared task*

Most datasets for the languages targeted by the SemEval-2 shared task were extracted from different corpora as discussed in section 3.1.1. The annotation guidelines for the distinct corpora focused on the following aspects:

CATALAN AND SPANISH For Catalan and Spanish identity, predicative and discourse deixis relations are marked [Recasens and Martí, 2010]. Part-of and set-member relations are excluded from the set of labeled relations.

With respect to the identity relation, NPs and proper nouns may be linked to other proper pronouns, full NPs, 3rd person pronouns, 1st/2nd person pronouns in quoted speech, clitics, demonstratives, relative and zero pronouns. The full span of possessive phrases is marked coreferent. Embedded mentions may refer to entities different than the referents of their larger NPs. In case both coincide, the larger span is selected. Relative pronouns are not linked when the phrase is nominalized and their maximal span is also preferred. Full coordinated phrases or their singular parts may also be coreferent. Generic nominal phrases are also marked when used referentially. As Recasens and Martí [2010] report, even phrases that do not agree on gender or number may be also linked. In the cases that a multiword expressions consist of multiple embedded mentions and only one of these embedded mentions is coreferent, the latter cannot be separated from the expression and thus cannot be annotated for coreference.

DUTCH For Dutch [Hoste, 2005], identity (both identity of sense and identity of reference) and bound relations are marked. Appositions may also be considered coreferent. All types of nominal phrases, named entities as well as personal, demonstrative and indefinite pronouns can enter a coreference relation. In the cases in which reflexives denote a world entity and are not lexicalized, they may be coreferent. Null pronouns are also not marked for coreference. Noun phrases with non-restrictive relative clauses may not be coreferent. The guidelines for the Dutch corpus also allow the phrases that

have a syntactic head other than a noun as well as metonyms to be marked as coreferent.

ENGLISH Not all relations annotated in the OntoNotes corpus [Hovy et al., 2006] were used for the dataset for the English language. This means that the English dataset within the SEMEVAL-2 and the one from the CONLL 2012 shared task constitute different collections of annotations. OntoNotes does not annotate singletons, thus, these were marked additionally via heuristic methods. Both maximal span as well as single constituents within coordinated phrases may also be coreferent. See section 3.2.2.2 for a description of the annotations in OntoNotes.

GERMAN The coreference annotation scheme for German [Naumann, 2006] included definite noun phrases, personal pronouns, relative, reflexive, and reciprocal pronouns, demonstrative, indefinite and possessive pronouns in the set of mentions thus labeling only entity and not event coreference. Identity of sense (only bound anaphora) and identity of reference relations are considered. In contrast, part-whole and holonymy-metonymy relations as well as zero anaphora are not marked. Maximal NPs are considered for complex cases as coordination, apposition, etc. Coreference relations for indefinite noun phrases are not labeled.

ITALIAN The dataset for Italian basically follows the guidelines provided by the MATE meta-scheme [Poesio et al., 1999]. Noun phrases and possessives are considered markables, however, singletons are excluded from the set. Additionally, all anaphoric types of relations are marked. Predicative NPs are not considered coreferent. Multiple antecedents referred to by a singular NP are marked separately even if they are coordinated. The maximal span of the coorrdination itself is not labeled as coreferent. Discourse deixis is also not included.

3.2.2.2 CoNLL 2012 *shared task*

As Pradhan et al. [2012] report, the annotations provided in the CoNLL 2012 shared task were more compliant to a language independent annotation scheme since they were acquired from one single corpus – the OntoNotes corpus. All languages were labeled with identity and appositive coreference relations between entities and events which were not restricted to predefined subset of entity types. Bare verbs are also marked for coreference if they can refer to a NP or another verb. All pronouns for English and Chinese (apart from expletive or pleonastic for English) and demonstratives are linked to their respective referents. The English generic *you* as well as Chinese generic pronouns were excluded from the annotation. For Arabic, nominative personal pronouns and demonstrative pronouns were marked.

Generic NPs may refer to pronouns and definite NPs but not to other generic noun phrases. Furthermore, bare plural nouns are always regarded as being generic. Pre-modifiers and pre-modifier acronyms (unless the latter refer to a nationality) that are not in a morphologically adjectival form may be also marked as coreferent. For Chinese, adjectival and nominal forms of the geographical and political type of named entities (GPE) are not morphologically distinct – thus the distinction in the usage is decided on by the annotators. Named entities of type nationality, other, religion, political (NORP), subject complements of copular verbs and small clauses are in general not regarded as coreferent unlike NEs such as dates (DATE), monetary values (MONEY), temporal expressions (TIME).

Additionally, Pradhan et al. [2012] list several special cases which we summarize below:

- An organization and its members may not be marked as coreferent.
- GPEs refer to their governments.
- Metonymic mentions may be linked when high confidence about their coreference is present.
- Verbal inflections are regarded not coreferent for Arabic.

3.2.2.3 *Consequences of the Use of Divergent Annotation Schemes*

In order to demonstrate how the differences in the various annotation schemes might be harmful to a multilingual approach as ours or as well what they may be helpful with, we would like to discuss several issues with respect to the SemEval-2 datasets that directly affect the first and highly important subtask of CR, mention detection (see chapter 5).

MULTIWORD EXPRESSIONS There is a significant difference in the use of multiword expressions across the six languages of the task. For both Catalan and Spanish (see table 3.6), for example, names, dates and complex numbers are represented in the form of a multiword expression and not as separate tokens as is the case for the rest of the languages. In general, such expressions may be regarded as good indicators for mentions, especially if the annotation scheme includes singletons in the set of labeled mentions. Additionally, as described in section 3.2.2.1, in case the maximal span of the multiword expression is not coreferent to another mention, but only part of it is, this link cannot be included in the annotations.

CORRESPONDENCE TO THE SYNTACTIC STRUCTURE One of the logical ways to detect mentions within the datasets of various languages is to rely on the syntactic annotations provided for them. Yet, it is also important that the

language	example
Catalan	[la Gran_Bretanya]
	[el 3,7_per_cent]
	[el 30_de_novembre]
	per_sobre_de
Dutch	[de Europese Unie]
	5,2 procent
	maandag 18 februari
	heeft kennis nodig
English	Texas Commerce Bank
	[March 17, 2003]
	in spite of
German	[Komitee Cap Anamur]
	am [4. Juni 1989]
	[gut elf Prozent]
	in [Gang] bringen
Italian	[Rai Uno]
	[Il 1 gennaio [2000]]
	non [aveva] bisogno de
Spanish	[el japonés Mitsubishi_Corporation]
	71_por_ciento
	[el 5_de_julio_de_1993]
	poner_en_marcha

Table 3.6: Examples from the annotation across the six languages of the SemEval-2 shared task. The tokens connected with underscores represent one single multiword expression. The actual mentions are marked in square brackets.

annotation schemes do agree on the level of overlap of the actual mentions and the phrases that can be extracted from the syntactic layer. This is so because a mention that does not have the exact same boundaries as its corresponding mention in the key set will be completely discarded by the evaluation software that is currently available. While mention boundaries generally correspond to NPs, there are differences in whether non-referring phrases, including indefinite NPs, and reflexive and relative pronouns are regarded as mentions. For example, the English and German data sets consider all of these noun phrases as mentions. Italian marks all non-pronominal NPs but does not mark reflexive or relative pronouns. Catalan and Spanish mark relative pronouns, but neither reflexives nor non-referring NPs. Dutch marks neither of these categories. For Catalan and Spanish, the situation is further complicated by the fact that null subjects are explicitly encoded in the data and are annotated as mentions: *[[*$\_{\text{subj}}$*] Reconoce que [la deuda exterior] es el principal obstáculo para [el crecimiento] . . . .]*. In Italian, in contrast, the verb is annotated as mention in such cases. The fact that 3 out of the 6 languages do not mark singletons has severe ramifications for the evaluation of mention detection since only after coreference is resolved a system decides which mentions to discard.

CORRESPONDENCE TO NAMED ENTITIES Named entities are often good indicators for mentions (see section 5.1.1). Thus, additional information can be extracted, when annotations for named entities are provided. Yet, named entities are not always included in standard annotation distributions. Such an example can be found in the NE annotations provided in the SEMEVAL-2 shared task. In table 3.7 we list excerpt examples from the NE annotations from all datasets. No information is included for German. For the other languages, the named entities generally correspond to mentions, but to different degrees: since Italian also annotates pronouns and abstract NPs (such as *Il suo ecumenismo*, in English: "his ecumenism") as named entities, the percentage of mentions that correspond to named entities is considerably higher than for Catalan and Spanish. Note that languages differ in whether determiners are regarded as being part of the named entity. In English, they are normally not included while Spanish does contain them. In the former case, this means that mentions based on named entities must be processed to exclude determiners.

Another peculiarity of the overlap of noun phrases with named entities can be observed in the Italian dataset of the SEMEVAL-2 shared task. The example listed in table 3.8 shows the named entities within two different sentences. What is interesting here, is the fact that the NEs may span over sentence boundary, which is not included in the definition of the annotation scheme for the mentions. Thus, considering NEs in this particular case can be misleading to the mention detection process.

language	token	NE
Catalan	Jordi_Virallonga	(person
	,	(person
	director	_
	de	_
	l'	(org
	Aula	org)
	,	person)|person)
Dutch	Frans	PER
	Ferdinand	PER
English	Denise	(person
	Dillon	person)
German	Pedros	_
	Frau	_
	Mari-Gaila	_
Italian	Mao	(person
	Asada	person)
Spanish	la	(person
	pareja	_
	Sandon_Stolle-Mark_Woodforde	(person)|person)

Table 3.7: Examples from the named entity annotations across all six languages of the SemEval-2 shared task with included examples of multiword expressions.

LEVEL OF MENTION EMBEDDING The data sets of the SemEval-2 shared task differ to a great extent in the maximal level of mention embedding: The highest level of 13 embedded mentions is reached in English followed by 10 embedded mentions in Catalan and Italian, closely followed by Spanish with 9. German has a maximum of 5 and Dutch a very moderate embedding with a maximum of 3. This is partly due to decisions whether non-referring mentions are annotated. Thus, Dutch, which has the most restricted definition also has the lowest number of embeddings. However, Catalan and Spanish have a tendency to use definite NPs for non-referring expressions, thus having high levels of embedding. Additionally, the wide definition of mention phenomena leads to situations where a mention contains a relative pronoun belonging to the same coreference chain: *[$_6$ a los militantes de muchos años [$_6$ que] hoy tienen una emoción especial]*. Examples for the highest level of embedding in Catalan and Dutch are shown in Table 3.9.

#	**token**	**NE**	**mentionID**
1	Torna	(person)	(5078)
2	nuovamente	_	_
3	a	_	_
4	Milano	(gsp)	(5079
5	,	_	5079)
6	per	_	_
7	recarsi	_	(5080)
8	,	_	_
9	nel	(time	_
10	1675	time)	_
11	,	_	_
12	a	_	_
13	Torino	(gsp)	(5082)
14	su	_	_
15	invito	(abstract	_
16	della	(person	_
17	corte	person) | abstract)	_
18	,	_	_
19	per	_	_
20	le	(concrete	(5084
21	decorazioni	_	_
22	della	(facility	_
23	chiesa	_	_
24	gesuita	_	_
25	dei	(person	_
26	Ss	_	5084)
27	.	_	_
1	Martiri	person) | facility) | concrete)	_
2	;	_	_

Table 3.8: An example of the named entity annotations in the Italian dataset from the SemEval-2 shared task that includes named entities spanning over sentence boundaries. Column *NE* lists the entity annotations in the two different sentences and column *mentionID* the set of *gold* mentions.

language	example
Catalan	[Joves_Agricultors_i_Ramaders_de_Catalunya (JARC)] vol denunciar [l'actitud d'[alguns escorxadors catalans [que] es neguen a [la creació de [la llotja única per establir [els preus del [conill]] i evitar així [['les dents de serra' en [aquest mercat]] i [les especulacions de [majoristes [que] juguen amb [els diferents preus [que] es fixen ara a [les llotges de [[Silleda] , [Madrid], [Saragossa] i [Bellpuig]]]]]]]]].
Dutch	[De zakenman] werd vooral bekend als [voorzitter van [voetbalclub KV [Mechelen]]].

Table 3.9: Example sentences for Catalan and Dutch from the SemEval-2 shared task datasets including maximal level of embedding of mentions. The separate mentions are marked by square brackets.

The level of embedding does not necessarily result from the typology of the language at hand. For example, the level of embedding for English in the CoNLL 2012 shared task was drastically decreased in comparison to the level of embedding in the English SemEval-2 shared task dataset, namely it was 5. Arabic also shared the same level of embedded phrases, while for Chinese the maximum number of embedded phrases was 4. The latter fact shows that it is not only hard to develop a coreference pipeline that can work efficiently across languages. On the contrary, training a coreference model or a CR system does not guarantee optimal performance even if applied on the same language, if the annotation scheme has such drastic differences, such as the change of embedding for English in this case.

3.3 SUMMARY AND CONCLUSION

The current chapter offered a detailed introduction to the state-of-the art approaches to multilingual coreference resolution introducing the two shared tasks, which are important for this enterprise: the SemEval-2[24] task 1: Coreference Resolution for Multiple Languages (see section 3.1.1) and the CoNLL 2012[25] Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes (see section 3.1.2).

In the second part of the chapter, we also discussed the availability of various types of linguistic annotations as well as the fact that even across the thirty European languages there are many cases for which corpora annotated for coreference still do not exist. Moreover, the languages that do provide such

[24] http://stel.ub.edu/semeval2010-coref
[25] http://conll.cemantix.org/2012

layers of annotation, follow highly divergent annotation schemes that additionally hinder the development of efficient and well performing approaches. We consider it important that, whenever possible, alignment of the various annotation schemes should be attempted as well as their standardization across the languages for which data is already annotated so that future annotation schemes can be easily made more uniform to that standard.

CHAPTER

4

PREREQUISITES FOR A PAIRWISE ML APPROACH TO CR

As we showed in section 2.3.2 and section 2.3.1 there are two ways to address the coreference resolution problem – rule-based and machine learning approaches, as well as various combinations of both, known as hybrid approaches. Since we cannot discuss and analyze our results without specifying the framework within which they were achieved, we devote the following chapter to that technicality. Nevertheless, in our work we aim to gain a deeper insight into the problems occurring within multilingual coreference resolution and not just to achieve optimal system performance. For this reason, we do not target system optimization and select the most widely used methods for the resolution of this task.

While designing and applying rules is, as a method, relatively straightforward, the selection, design and execution of machine learning approaches is far more complex. Apart from the decision on using either supervised, or unsupervised, or semi-supervised, etc. learning algorithms (see section 2.3.2), which mainly differ in the form of input they require in terms of answer labels, there is as well a diverse variety of machine learning approaches that can be chosen from. These vary based on the learning method they use for the resolution

process. In both, the SemEval-2 and the CoNLL 2012 shared tasks, the majority of the approaches to the multilingual coreference resolution problem were implemented within a machine learning framework. The concentration on one type of system can be attributed to multiple reasons, the most important of which are the following:

1. the incommensurately increasing complexity and required in-depth linguistic knowledge for the design of rule-based systems when more than one language is targeted,
2. the availability of a wide range of annotated corpora for various languages on which supervised machine learning systems can be trained and tested,
3. the realization that only machine learning can provide the flexibility and multilinguality that is needed for the use of an approach that targets more than one language simultaneously,
4. the constant growth, advancement and improvement of statistical approaches to the coreference resolution task that can be easily applied and extended for further languages.

These reasons also motivate our choice of selecting a machine learning approach for our investigation with respect to the problems and peculiarities that can arise within a multilingual coreference resolution pipeline. Since we aim at investigating CR across multiple languages simultaneously, we consider a rule-based approach to the task as unmanageable in a reasonable time frame, especially since the methodology must be extendable and applicable to new languages. More importantly, an in-depth knowledge of the peculiarities of the coreference phenomenon across all languages needs to be present in order for efficient and robust rule sets to be assembled, which is not an easy task considering the number of targeted languages. Furthermore, language specific adjustments defeat the purpose of an efficient and easily adaptable multilingual approach, which we regard as a starting point for our research and a reasonable objective.

Based on the general methodology and implementation (in both machine learning and rule-based approaches as well as using diverse representation models), the systems participating in the SemEval-2 and the CoNLL 2012 shared tasks were divided into the following broad categories [Recasens et al., 2010, Pradhan et al., 2012]:

decision trees
leaf

1. ***decision trees*** – employ a tree-like structure in which the ***leaves*** (the nodes without child nodes), represent the actual classes in the data and the branches or conditions leading to those leaves are the combinations of considered features that result in the given classes. Used in: [Broscheit et al., 2010], [Kobdani and Schütze, 2010], [Sapena et al., 2010], [Xu et al., 2012]

2. ***maximum entropy*** – is used to determine probability distributions of the classes in the training data given the combinations of observed features for these labels and employing these probabilities for the observations in the test data. Used in: [Broscheit et al., 2010], [Kobdani and Schütze, 2010], [Attardi et al., 2010], [Björkelund and Farkas, 2012], [Li, 2012], [Li et al., 2012], [Yang et al., 2011]

   maximum entropy

3. ***memory-based learning*** – memory-based learning does not abstract away from the data. All training instances are stored in memory and new decisions are made on the basis of previously seen examples. Used in: [Zhekova and Kübler, 2010], [Zhekova et al., 2012]

   memory-based learning

4. ***naïve Bayes*** – the naïve Bayes algorithm, also known as ***independent feature model***, employs the ***Bayes' theorem*** and considers all included features to have an independent contribution to the overall probability of the class. Used in: [Kobdani and Schütze, 2010]

   naïve Bayes
   independent feature model
   Bayes' theorem

5. ***support vector machines*** – also known as ***SVMs***, or ***support vector networks***, are geared towards the recognition of patterns in the training data that contribute to the accomplishment of a binary decision on the class of the new, test example. Used in: [Uryupina, 2010]

   support vector machines
   SVMs
   support vector networks

6. ***sieve-based*** – sieve-based or ***multi-sieve-based*** is an approach that splits the classification in multiple stages by attempting to make the easiest decision in the first stage and increasing the difficulty for each further level. Used in: [Fernandes et al., 2012], [Chen and Ng, 2012], [Zhang et al., 2012], [Shou and Zhao, 2012], [Xiong and Liu, 2012]

   sieve-based
   multi-sieve-based

7. ***logistic regression*** – logistic regressions are also binary classification algorithms that provide the positive or respectively negative outcome in terms of probability. In fact, logistic regression modelling is often defined as being equivalent to maximum entropy modelling [Klein and Manning, 2003]. Used in: [Stamborg et al., 2012]

   logistic regression

8. ***directed multigraph representation*** – is a complex graph structure used to represent the mentions and the relations between them allowing for more than one relation between two mentions. The final class is induced by observations on the clusters that the nodes build within the graph. Used in: [Martschat et al., 2012]

   directed multigraph representation

9. ***latent structure*** – is based around the assumption that the coreference trees in a document are latent structures and thus the probabilities are calculated to represent the chance of a given classification depending on predefined values. Used in: [Chang et al., 2012], [Fernandes et al., 2012]

   latent structure

10. ***BART-based*** – using the BART toolkit [Versley et al., 2008b] for resolution. BART employes a variety of machine learning approaches and/or toolkits

    BART-based

(e.g. WEKA and MaxEnt). Used in: [Martschat et al., 2012], [Uryupina et al., 2012]

C4.5

11. ***C4.5*** – C4.5 is a subtype of the decision tree category, because the C4.5 algorithm generates a decision tree for the classification. It is an extension of the previously used ***ID3*** algorithm. Used in: [Yuan et al., 2012]

ID3

deterministic rules

12. ***deterministic rules*** – using language specific rules for each targeted language. Used in: [Yuan et al., 2012], [Xu et al., 2012], [Zhang et al., 2012]

None of the systems participating in either of the shared tasks that submitted results for all targeted languages were solely rule-based. Yet, the diversity of machine learning methods that were employed in the system implementations shows that it is still unclear which approach can provide the best solution to this exceedingly complex task.

In order to tackle MCR and to be able to identify the problems that arise within that complex enterprise, we employ memory-based learning, which we present in more detail in the following section, section 4.1. In section 4.2, the specific memory-based learning software that we embed in the overall system architecture (see section 4.3) is described. Section 4.4 then provides a summary and concluding remarks for the chapter.

4.1 MEMORY-BASED LEARNING FOR NLP

memory-based learning

Memory-Based Learning (MBL) methods [Daelemans and Van Den Bosch, 2005] are suitable for a wide range of NLP tasks and are thus often used as solutions for various NLP problems. MBL has been used for decades in the research community: [Aha, 1997, Aha et al., 1991, Cost and Salzberg, 1993, Daelemans et al., 2007, Kolodner, 1993, Stanfill, 1987, Stanfill and Waltz, 1986], leading to efficient and robust performance across various natural language processing tasks. As Roth [1999] reports, MBL is one of the most successful learning approaches used in empirical NLP. Its flexibility and capability to adapt to various applications renders memory-based learning appropriate for our multilingual approach and provides a reasonable motivation for us to employ this technique within our research. Before we continue, we provide a short introduction to MBL (section 4.1.1) after which we delineate in more detail the manner of application of the method (see section 4.1.2).

4.1.1 *Overview*

similarity metric

MBL is a representative of supervised learning methods, meaning that the instances used for training include the correct answers. In memory-based learning, every new example is classified by a ***similarity metric*** that compares that example to previously seen instances and assigns a class to it – normally,

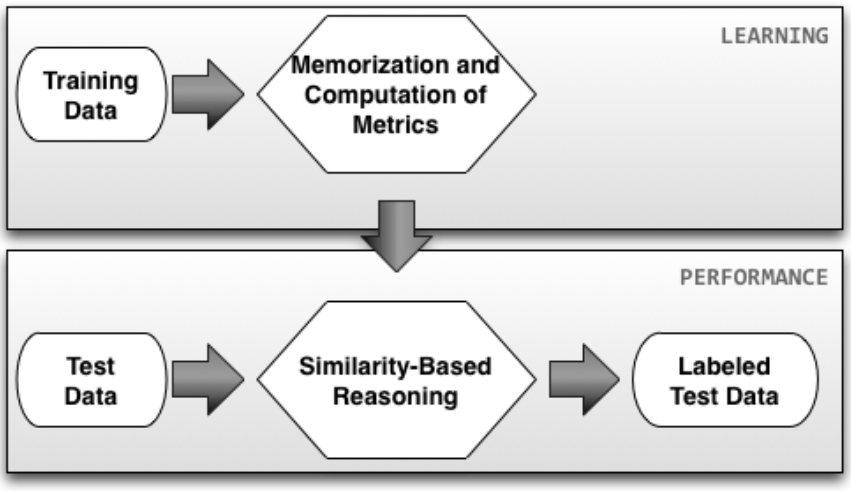

Figure 4.1: The general architecture of a memory-based learning system presenting the learning (the upper part of the figure) and the performance (the lower part of the figure) modules.

this is the most frequently seen class in a collected pool of the most similar instances. Over the years the approach has gained a wide variety of names that put the focus on various aspects of its nature: ***instance-based***, ***case-based***, ***similarity-based***, ***example-based***, ***memory-based***, ***exemplar-based*** as well as ***analogical***. Because of the fact that the instances are stored directly in memory, without any restructuring, reformulating or abstraction, this method is also often called a ***lazy learning*** method. As Daelemans et al. [2007] report, a MBL system is essentially built up of two components – a memory-based learning component and a similarity-based one that is also known as the performance component. Roth [1999] also points out that one of the key features of MBL, as a result of memorizing instances and basing a decision on their type and nature, is its closeness to rule-like behaviour. Due to the robust performance of MBL we consider it a good solution for the coreference resolution task and its multilingual setting. In the following section we present the application of that approach to the CR task and provide examples for clarification. A more detailed account of the methodology is provided in [Daelemans and Van Den Bosch, 2005].

instance-based learning
case-based learning
similarity-based learning
example-based learning
exemplar-based learning
analogical learning
lazy learning

4.1.2 *Application*

learning component

performance component

The fundamental architecture of a system based on similarity of examples is visualized in figure 4.1. The ***learning component*** (to be found in the upper part of figure 4.1) has an easy and exceptionally straightforward task, namely adding training examples directly into memory without any processing. Additionally, it also performs the computation of the metrics. The ***performance component*** (found in the lower part of figure 4.1) uses the examples stored in memory by the learning component to classify the new cases by selecting the class that is represented with the highest frequency among the most similar previously seen instances.

Usually, the data, or the examples that are used in a MBL approach, are in the form of feature vectors. They represent a collection of features that provide information about a given instance and its context in either the training or test data.

In section 2.3.2, we noted that we consider mention pairs as instances for the resolution process. So, let us take as an example the two sentences in (28) where all NPs are marked as mentions and no coreference information is added to the provided annotation.

(28) [Mary$_1$] had [a good idea$_2$]. [She$_3$] wanted to tell [John$_4$].

In order to create the feature vectors for these two sentences, a pairwise approach, following the mention-pair model [Rahman and Ng, 2011], combines each mention with all the mentions occurring previously to it in the part of the text that is examined. In this way, for every mention for which an antecedent needs to be found, all potential applicants for antecedents are observed. The process creates the foundation of the feature vectors that are later extended. As a result, we get the collection of mention-pairs listed in (29).

(29)
```
[a good idea] [Mary]
[She] [a good idea]
[She] [Mary]
[John] [She]
[John] [a good idea]
[John] [Mary]
```

However, in a pairwise approach to CR, feature vectors include only the syntactic heads of the actual mentions and thus the pairs acquire the respective forms in (30) below.

(30)
```
idea Mary
She idea
She Mary
John She
```

```
John idea
John Mary
```

Once a decision on the features that need to be used is met (we provide a detailed discussion on the topic in chapter 7), the information representing those features is added to the vectors. For example, if we want to include a feature distinguishing the ***lexical category***, also called ***part-of-speech*** or ***POS***, we can either choose a binary value (being true or false for a given assumption) or the part-of-speech information itself. Adding the latter for each of the members of the pair to our toy example results in the short toy vectors listed in (31). For more clarity, the first instance in (31) indicates that the mention head *idea* is a common noun (specified by the POS tag *NN*), while the mention head *Mary* is a proper noun (designated by the POS tag *NNP*).

lexical category
part-of-speech
POS

(31)
```
idea Mary NN NNP
She idea NNP NN
She Mary PRP NNP
John She NNP PRP
John idea NNP NN
John Mary NNP NNP
```

The vectors within the training and test data need to have the same number of features and the order in which the values for those features are ordered also needs to be kept. Additionally, there is no limitation to the number of features that can be added to a feature vector. An important one for supervised approaches, however, is the feature representing the answer – i.e. the classes to be assigned to all new instances. In CR that class can be represented by a binary value (T(rue) if the pair represents two coreferent mentions or F(alse) if the mentions are not coreferent). The only coreferent pair in (28) is *She Mary* as shown in (32).

(32) [Mary$_1$] had a good idea. [She$_1$] wanted to tell John.

Thus, adding that information to our feature vectors leads to a complete, but still rather minimalistic, representation as in (33).

(33)
```
idea Mary NN NNP F
She idea NNP NN F
She Mary PRP NNP T
John She NNP PRP F
John idea NNP NN F
John Mary NNP NNP F
```

The new/test instances acquire the exact same form without the last feature – the answer. On the basis of all training instances from the training set that have been stored in memory, the test instances can be labeled by a similarity-based reasoning method.

The similarity between the instances can be determined by various distance metrics (such as the Overlap metric for example). As Daelemans et al. [2010] report, the Overlap metric (as given in equation 4.1 and equation 4.2) denotes the distance between example X and example Y ($\Delta(X, Y)$). The distance overall is represented by n number of features and is then defined by the sum of the distances (δ) between the separate features.

$$\Delta(X, Y) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{4.1}$$

where:

$$\delta(x_i, y_i) \begin{cases} abs(\frac{x_i - y_i}{max_i - min_i}) & \text{if numeric, else} \\ 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases} \tag{4.2}$$

The overlap metric is just one of the possibilities to measure the distance between two given examples. Another possibility would be the Levenstein metric for instance. The latter does not simply calculate the difference overall between the examples but also examines the total number of needed actions (insertions, deletions and substitutions) to transform instance X to instance Y. Since our work does not aim at a comprehensive evaluation of the distance metrics for the given problem but rather at the assembling of a basic framework for coreference resolution that employs widely used approaches, we do not assess all available distance metrics for MBL. We exploit mainly the default settings and approaches (see section 4.2) and for that we make use of the Overlap metric. For a wider overview of the possibilities in that area, we recommend [Daelemans and Van Den Bosch, 2005] for more comprehensive and detailed information on the distance metrics that can be used within a complex MBL architecture.

Independently of the distance metric in use, one very well known drawback of memory-based learning is the fact that features carrying less informative or misleading information have a highly detrimental effect on the performance of the learner [Aha, 1998]. In order to account for this deficiency, feature weights can be included in the instance representation. These can be used to determine which features are more relevant and descriptive for the phenomenon and to put more emphasis on them.

For any objective voting system, it is impossible to make a decision if there is a choice between two equally weighted options. Thus, in MBL, in the case that two or more possible answers receive the same ranking (the so called ***tie***) a tie-breaking resolution method is used. This can either involve incrementing the set of nearest neighbors until the tie is resolved or randomly choosing between the possibilities.

tie

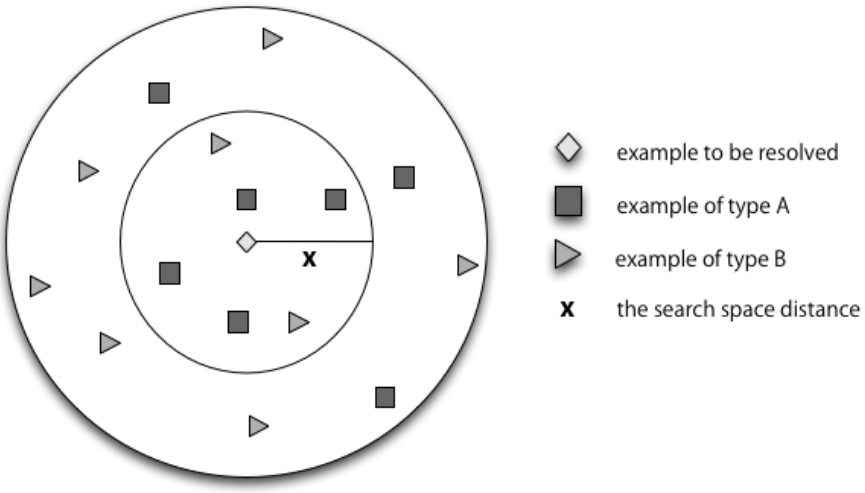

Figure 4.2: A graphical representation of the distribution of examples across the search space of a *k*-nearest neighbor classification procedure.

4.2 TILBURG MEMORY-BASED-LEARNER

In section 4.1 we introduced memory-based learning and motivated our decision for choosing it, among all machine learning methods, for our investigation of the issues of multilinguality in CR. As an implementation of that algorithm, we chose the Tilburg Memory-Based Learner (TiMBL). The following section provides information on the structure and usage of this learner.

TiMBL[1] is a robust representation of a combination of various different MBL approaches. It is a highly efficient and discrete implementation of the ***k*-nearest neighbor (k-NN)**. The k-NN approach is an algorithm that makes a decision on the classification of a new, unseen instance derived from the class of the most often seen closest example or examples within all training instances. As illustrated in figure 4.2, the example to be resolved is labelled with the class of the most often seen example in the search space (with x defining the radius of the search space). Within that distance there are four examples of type A and two examples of type B. Thus the type of the target instance is classified as A. In our coreference resolution context, this classification procedure means that a feature vector from the test set will acquire the label that represented the highest number of closest vectors within the examined distance. A more detailed description of the k-NN classification algorithm is presented in Daelemans et al. [2007].

k-nearest neighbor
k-NN

[1] http://ilk.uvt.nl/timbl

TiMBL is a highly important and useful NLP tool because it is created around the belief that intelligent behavior can be accomplished by analogical reasoning rather than the use of abstract mental rules. The complete C++ source code is released under the GNU General Public License (GNU GPL)[2] as published by the Free Software Foundation[3].

Although originally TiMBL was designed to be an efficient solution for the linguistic classification task, it can be exploited for any alternative categorization task with appropriate (symbolic or numeric) features and discrete (non-continuous) classes for which a sufficient amount of training data per class is available. The latter directs us to the previously discussed topic of the acute shortage of labeled data within the context of coreference resolution. In fact, the shortage of labeled data is twofold. First, there is the acute need for more corpora annotated for coreference altogether. Second, working on already provided datasets, the coreference phenomenon in the context of a pairwise resolution approach leads to a highly skewed distribution of positive (pairs that are coreferent) vs. negative (pairs that are not coreferent) instances. For example, let us look into the excerpt shown in (34), taken from [Recasens et al., 2009]. The actual coreference chains in (34) are the following: 1-5-6-30-36, 9-11 and 7-18. However, before these chains are acquired the pairwise approach first assembles a set of all potentially coreferent pairs (e.g. 3-4, 2-4, 1-4, etc.). Altogether, this results in an exceptionally high number of pairs for such a short example (more than 500). Yet the actually coreferent pairs are only 12, which leads to a ratio of approximately 1:42 positive vs. negative instances.

(34) [The beneficiaries of [[spouse's]$_3$ pensions]$_2$]$_1$ will be able to keep [the payment]$_4$ even if [they]$_5$ remarry provided that [they]$_6$ fulfill [a series of [conditions]$_8$]$_7$, according to [the royal decree approved yesterday by [the Council of Ministers]$_{10}$]$_9$.
[The new rule]$_{11}$ affects [the recipients of [a [spouse's]$_{13}$ pension]$_{12}$ [that]$_{14}$ get married after [January_1_,_2002]$_{16}$]$_{17}$.
[The first of [the conditions]$_{18}$]$_{19}$ is being older [than 61 years old]$_{20}$ or having [an officially recognized permanent disability [that]$_{22}$ makes one disabled for [any [profession]$_{24}$ or [job]$_{25}$]$_{23}$]$_{21}$.
[The second one]$_{26}$ requires that [the pension]$_{27}$ be [the main or only source of [the [pensioner's]$_{30}$ income]$_{29}$]$_{28}$, and provided that [the annual amount of [the pension]$_{32}$]$_{31}$ represents, at least, [75% of [the total [yearly income of [the pensioner]$_{36}$]$_{35}$]$_{34}$]$_{33}$.

At first sight, such a highly skewed distribution raises many questions about the validity and importance of this ratio. Even though we observe it only in a small, toy example, such similar cases have also been reported for large scale corpora. Uryupina [2004], for instance, showed that in the MUC-7

[2] http://www.gnu.org/copyleft/gpl.html
[3] http://www.fsf.org

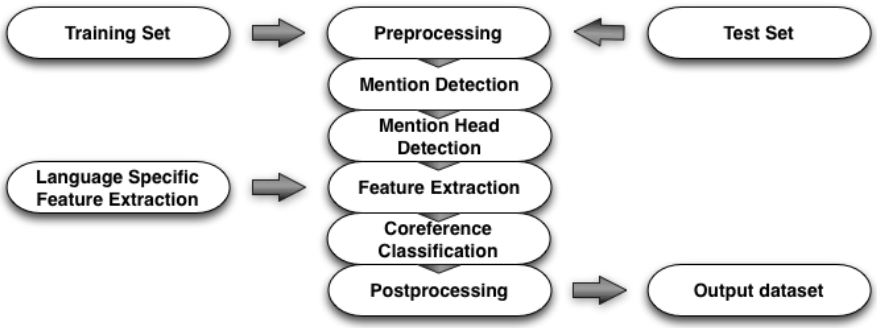

Figure 4.3: A general overview of the workflow of the multilingual coreference resolution system that is employed in our work showing its most important modules: Mention Detection, Mention Head Detection, Feature Extraction and Coreference Classification.

corpus presented in [Hirschman and Chinchor, 1997], there was an even bigger discrepancy between the two types of pairs with an approximate ratio of 1:48 (in other words only 1-2% of the instances are actually coreferent). Similar observations were announced by Ng and Cardie [2002b] for the MUC-6 dataset as well. These findings confirm that the skewed distribution of examples is typical for the pairwise approach. This fact is important for us, since as Hoste [2005] points out, standard classification algorithms tend to lead to poor performance when unbalanced datasets are used. In such cases, the minority classes in the data (e.g. the positive/coreferent class in our toy instances) tend to be highly or even completely ignored by various algorithms. This results in a tendency of the learner to output only instances of the majority classes (negative examples in our case). In the search for solutions to the problem various **instance sampling** techniques have been explored that aim at the intentional increase of positive or/and decrease of negative examples according to a predefined schema [Ng and Cardie, 2002b, Uryupina, 2004, Zhao and Ng, 2007, Wunsch et al., 2009, Recasens and Hovy, 2009, Zhekova, 2011].

instance sampling

4.3 UBIU – A MULTILINGUAL COREFERENCE RESOLUTION SYSTEM

The full coreference pipeline that we are going to employ for our experimental research is integrated within the UBIU coreference resolution system [Zhekova and Kübler, 2010, 2011, Zhekova et al., 2012]. Since the focus of the thesis is to propose and discuss solutions to the problems that occur within that framework, a general introduction of the system is required and presented in the following

section. A modular overview of the system is also presented in figure 4.3. The figure shows the most important components of UBIU and their overall interaction. Further description of the system and its participation within both shared tasks (SEMEVAL-2 and CONLL 2012) can be found in [Zhekova and Kübler, 2010, 2011, Zhekova et al., 2012]. Consequently, we will mainly focus on the pipeline in order to provide more clarity to circumstances in which we observe the multilingual issues for coreference resolution.

In section 4.3.1 we review the preprocessing techniques that we undertake, while section 4.3.2 discusses the mention detection subtask of CR. Another important subtask, namely mention head detection, is presented in section 4.3.3. The initial set of features that we use is listed in section 4.3.4. The last, but clearly not least important, part of the CR pipeline, coreference classification, is delineated in section 4.3.5. All postprocessing procedures are accounted for in section 4.3.6.

4.3.1 *Preprocessing*

As for any other natural language processing task that uses large collections of data, a preprocessing step is an important and needed building block of the pipeline. It ensures the integrity and mainly the consistency of the bits and pieces that the following processes rely on. Additional steps such as data reformatting, restructuring and enhancement can also be carried out within this step. In UBIU, the preprocessing module guarantees that the data used for the distinct languages avoids such inconsistencies and reformats the data in the structure that the pipeline expects. For example, let us consider the predicate arguments information provided in the annotations of the CoNLL 2012 datasets. The original format of those annotations is as shown in column *ParseBit_BPP* in table 4.1. Yet, it is more helpful to have that information on a per-token basis and thus during preprocessing the module distributes this information for each of the tokens as shown in column *ParseBitAPP* in table 4.1.

4.3.2 *Mention Detection*

MD in UBIU strongly relies on the annotations provided for the various targeted languages. The MD module identifies the mentions depending on the definition of the CR task at hand. This means that if the system is supposed to tackle only noun-phrase coreference, it will aim at extracting only nominal phrases as mentions. Alternatively, if noun-phrase and verb coreference is the goal, then verbs will be added to the set of mentions that the system has to identify as well. This increases the difficulty of the CR task not only with respect to MD. Adding verb coreference means that the modules following mention detection need to handle two different resolution procedures – one for noun phrases and one for verbs. This increased complexity led to the fact that most participating

#	token	POS	ParseBit_BPP	ParseBit_APP
0	On	IN	(ARGM-LOC*	ARGM-LOC
1	a	DT	*	ARGM-LOC
2	wall	NN	*	ARGM-LOC
3	outside	IN	*	ARGM-LOC
4	the	DT	*	ARGM-LOC
5	headquarters	NN	*)	ARGM-LOC
6	we	PRP	(ARG0*)	ARG0
7	found	VBD	(V*)	V
8	a	DT	(ARG1*	ARG1
9	map	NN	*)	ARG1
10	.	.	*	*

Table 4.1: Example of data-reformatting.

systems in the CoNLL 2012 shared task completely ignored verb coreference [Pradhan et al., 2012].

Mention detection is a highly important and relatively complicated subtask of multilingual CR. In order for a system to be able to decide whether phrases are coreferent or not, those phrases first need to be correctly identified. However, **chunking**, also known as **shallow parsing** or **light parsing**, which is the task of identifying constituents of a specific type in text (e.g. noun phrases) is already a challenge on its own when a single language is targeted. In a multilingual setting, the phrases, or **chunks**, to be extracted may differ considerably in their characteristics. Thus we devote chapter 5 to various issues concerned with the problems arising within the multilingual coreference subtask – multilingual mention detection.

chunking
shallow parsing
light parsing
chunks

4.3.3 *Mention Head Detection*

The next module in the system pipeline is still partially concerned with the mentions in the data. This is so, because once the mentions are identified, the system needs to extract their syntactic heads in order to be able to assemble feature vectors, as already described in section 4.1.2.

In a monolingual setting, mention head detection can be successfully performed by simple heuristics or rule-based methods and thus, in state-of-the-art approaches rather than being a proper subtask of coreference resolution, it was mainly considered as a subtask of mention detection. In chapter 6, we address that issue and advocate its revision. Furthermore, similar to mention detection itself, mention head detection can also face numerous additional

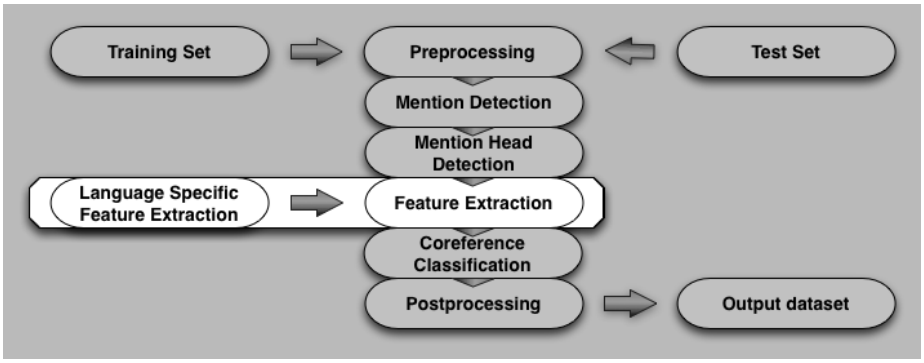

Figure 4.4: The language specific part of the multilingual coreference resolution system that is employed in our work, which is part of the Feature Extraction module of the system.

problems in a multilingual setting. One difficulty, for example, is the fact that languages tend to place the syntactic heads of their mentions in different positions. Such problems as well as suggestions for their solutions are also presented in chapter 6.

4.3.4 *Feature Extraction*

The feature extraction (also called feature selection) module in UBIU provides highly important functionality to the performance of the full coreference pipeline. The features that this module collects carry the information on the basis of which the final decision by the coreference resolver can be made. As emphasized in figure 4.4, it is generally considered to be a language-dependent component that extracts language specific features for the targeted languages, dependent on the provided annotations in the data. Feature extraction within a multilingual context raises various significant questions concerning the capability of the module to represent language specific knowledge in a flexible and easily adaptable way. One of the main aims of our work is to propose acceptable and efficient multilingual solutions to this very complex and elaborate task. For this reason, we devote chapter 7 to that topic where, depending on the selected features, we show that a multilingual and even language independent module can be assembled.

4.3.5 *Coreference Classification*

The actual coreference classification module represents a wrapper component for the TiMBL memory-based learner. It carries out the classification by labeling

the test instances. Those labels are used to form the coreference clusters in the postprocessing step. This part of the coreference pipeline is fully language independent and thus does not present an issue interesting for our discussion. The only linguistically motivated change that could be attempted in this module is concerned with various parameter optimizations of the coreference learner that are motivated by the distinct nature of the coreference phenomenon, which can differ depending on the approached language.

However, since our main aim is not to reach the most optimal system performance, we do not target an exhaustive parameter optimization for any of our reported experiments. We use *IB1* which is the default instance-based learning algorithm. It is generally known to lead to higher accuracy at the cost of more memory and slower computation [Daelemans et al., 2010]. Similarity is computed in UBIU based on weighted overlap, while gain ratio is considered for the relevance weights. These are default TiMBL parameter settings. The number of nearest neighbors that are included in the search space is set to 3 ($k = 3$). With respect to the number of nearest neighbors, $k = 1$ is the default, yet in [Zhekova and Kübler, 2010] we reported that using 3 instances leads to better overall performance. For more details on the numerous options for parameter optimization that TiMBL provides, the reader is referred to [Daelemans et al., 2010].

4.3.6 *Postprocessing*

The postprocessing module in UBIU constructs the clusters that represent the coreference chains based on the links identified by the coreference classification module. Its task is simply to include the achieved results in the data. Furthermore, depending on the existence of singletons in the key set (the mentions that constitute a distinct class on their own, or in other words, all singleton mentions), all mentions that are not identified as members of any of the clusters may either be removed (when the key does not contain singletons) or left in the system output (when the key contains singletons). Noun phrases do not always require an antecedent, since they can be new to the discourse. Yet, pronouns most often refer to already introduced entities (apart from exceptional cases such as the pleonastic use of the pronoun *it* for example). Thus, pronouns that were not linked to an antecedent are superficially bound to the last seen subject or, in the cases in which no subject is present, to the last seen mention in that step. We note that all mention pairs that were positively classified as coreferent by the previous module are involved in the formulation of the cluster.

The postprocessing step is also not dependent and affected by the variation in multilinguality approached by the overall coreference resolution system. For this reason, this part of the pipeline does not pose questions interesting to our investigation and thus we do not return to it in our further discussion.

4.4 SUMMARY AND CONCLUSION

In the current chapter, chapter 4, we presented the framework in which we situate our research on the peculiarities and problems connected with recasting the problem of coreference resolution as a multilingual enterprise.

As introduced in section 4.3.2, section 4.3.3 and section 4.3.4, mention detection, mention head detection and feature selection are the processes with a complex nature that are directly affected by the multilinguality aspect. For this reason, our investigation will concentrate only on these three subtasks in chapter 5, chapter 6 and chapter 7 respectively.

Part III

TOWARDS MULTILINGUAL COREFERENCE RESOLUTION

CHAPTER

5

MENTION DETECTION

In section 2.3, we introduced ***mention detection*** as the process of identifying the phrases that can potentially be rendered coreferent to others. This is one of the most essential subtasks of coreference resolution and, respectively, of its multilingual setting. One of the ways to extract such constituents is by using the syntactic annotations provided in the data – all included noun phrases are then marked as mentions. Yet, syntactic annotations are not always available, especially when underresourced languages are targeted by the coreference system. Among all European languages, for example, there are several cases such as Lithuanian, Maltese, Serbian, Slovak for which syntactic annotations are not easily available or existent. Therefore, a truly multilingual or even language independent system cannot always rely on this annotation layer.

mention detection

Furthermore, most of the state-of-the-art CR systems in the last decade concentrated only on the proper resolution process, because mention boundaries were included in the annotations provided for CR. Systems that use *gold* standard mentions (including singletons) [Denis and Baldridge, 2008, Haghighi and Klein, 2009, Ng and Cardie, 2002a] may be differently evaluated by the current evaluation metrics: MUC [Vilain et al., 1995], for instance, is completely insensitive to singletons in the system output, while B^3 [Bagga and Baldwin, 1998], CEAF [Luo, 2005], and BLANC Recasens and Hovy [2011], assume that there are no singletons in either training or test datasets. As Kübler and Zhekova [2011] show, the presence of singleton mentions in both the key and response sets can have a significant influence on the evaluation of coreference

resolution systems. We note that singletons in the key set can lead to boosted system performance, while singletons in the response affect less the overall system scores. Yet, including singletons in the key datasets is a necessary step that allows more realistic evaluation, because it properly rewards the system's mention detection performance. Moreover, in the case that a mention is removed by the system (assuming that it is a singleton, while that mention is present in the key set) it does not lead to artificially reduced scores. However, newer CR enterprises, as both SemEval-2 and CoNLL 2012[1] shared tasks, do not always include mention boundaries for singletons within the standard annotation layers provided for training and testing the various systems. This constitutes a problem, because any coreference resolution pipeline needs a set of mentions to work with. For this reason, mention detection has to be considered as a proper subtask of state-of-the-art CR systems. Altogether, this increases the complexity of the task, but it as well situates it in a more realistic scenario.

In general, mention detection can be achieved with high system performance when consistent and reliable linguistic annotation layers are available (such as POS tags, dependency or phrase structure, etc.). However, multilinguality can increase the difficulty within this subtask as well as annotations and annotation schemes may vary across languages or certain types of annotations may not be available for all of the targeted languages (see chapter 3). In these cases, information-poor approaches might be more easily applicable within the MCR pipeline and thus present an interesting exploration goal for the research community.

The increased difficulty, however, is not always concerned with the availability and type of annotations. It is also important to know which approaches for mention detection can be employed so that easier adaptation to data variations can be achieved. Both machine learning and rule-based approaches can be considered for this purpose. For example, Chen et al. [2011] and Zhou et al. [2011] employ ML, however, Zhou et al. [2011] report that the machine learning method that they used leads to low mention extraction recall. The latter generally results in low system performance of the coreference resolution pipeline. For this reason, the authors developed and applied a rule-based approach to the problem.

The present chapter investigates various methods for automatic mention detection in a multilingual coreference resolution setting (see section 5.1). We focus on methods that are applicable to different languages and investigate which linguistic annotations are most beneficial across the different datasets and annotation schemes. We present a thorough evaluation of the investigated methods (see section 5.2) for both SemEval-2 and CoNLL 2012 datasets, covering the following languages: Arabic, Catalan, Chinese, Dutch, English,

[1] However, CoNLL 2012 provided *gold* mentions and *gold* mention boundaries within extra evaluation settings.

German, Italian, and Spanish. Since we employ evidence from the datasets of both shared tasks, SemEval-2 and CoNLL 2012, for which different types and format of annotations were provided, we report experimental results for each task separately.

We note that the results from the full coreference resolution pipeline for SemEval-2 and CoNLL 2012 are not fully comparable, not only because of the fact that different datasets were employed, but as well because different feature sets and system settings were considered within the coreference resolver. This is so because there are too many variables or, in other words, differences between the pipelines employed on both datasets and a potential change in performance could not be unambiguously predicated on one of the components of these pipelines.

Via this detailed investigation, we are aiming to gain more insight into multilingual mention detection and attempt to find an answer to the question: *Can mention detection be performed in a close to language independent manner?* Along the way we examine which annotation layers are sufficient for the development of a multilingual approach and respectively which are necessary for a reliable and robust one. It is also important to ask which layer provides most indicative information and under what circumstances can this information be employed. Another issue interesting to us is the quality of the annotation layers and if they are equally reliable across the various languages. We are also interested in what problems can occur, when MD is approached for more than one language and what the prerequisites for objective evaluation of mention detection methods would be.

5.1 METHODS FOR MULTILINGUAL MENTION DETECTION

As Uryupina [2010] reports, the majority of the state-of-the-art systems at this time require mentions to be already identified within the data. The author points the lack of more extensive studies on the topic of mention detection, especially when the targeted noun phrases are not ***base NPs*** (simple, not recursive noun phrases), but rather complex structures without any restriction to their semantic type.

base NPs

Section 5.1 presents six different approaches to the problem of mention detection in a multilingual setting when semantically unrestricted and structurally complex phrases are targeted. We describe the methodology behind the approaches and discuss their data necessities and dependencies. Our main goal is to identify the prerequisites for a robust and most importantly multilingual approach by examining a variation of rule-based and machine learning techniques within both shared tasks and across all eight languages included in them.

Potential methods for mention detection that can identify complex phrases are not easy to create, as they are always dependent on the data and annota-

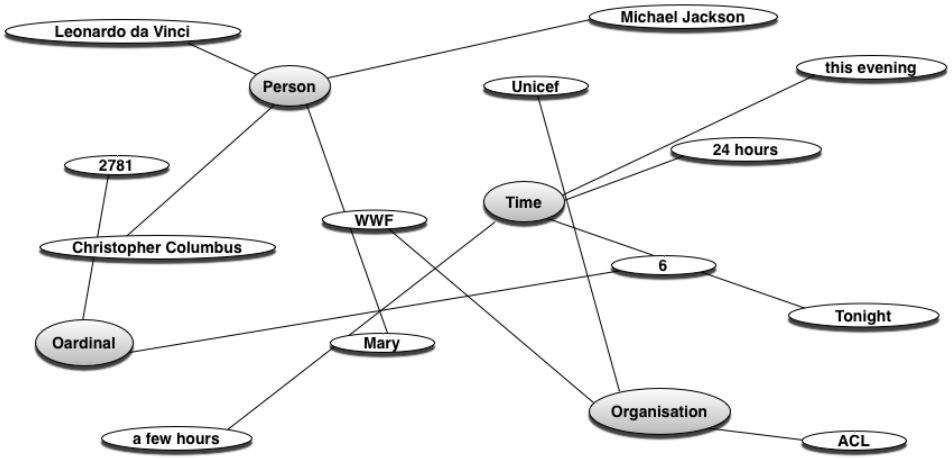

Figure 5.1: A graphical representation of a toy named entity network focusing on four different named entity types: ordinal, person, time and organisation.

tions provided for that task. Thus, not all methods are directly applicable to the datasets of both shared tasks, SemEval-2 and CoNLL 2012; section 5.1.7 provides a description of this issue and an overview of all methods and the data they can be applied on.

5.1.1 *Mention Detection based on Named Entity Structure*

Named entity annotations, or simply named entities, are phrases that can easily be discovered by both grammar-based and statistical approaches. They can be exceedingly helpful in coreference resolution, because mentions of the same entity type have a higher chance of being coreferent than mentions of different semantic types. For this reason, CR corpora often contain this type of annotations. As can be seen in figure 5.1 and in table 3.7 on page 64, where we presented examples of NE annotations across the languages in the SemEval-2 shared task, NEs can be instances of diverse entity types (e.g. person, time, ordinal, organization), which makes those phrases exceptionally good candidates for potentially coreferent mentions. They represent predefined categories such as proper names, locations, quantities, expressions of time, monetary values, etc, which are the phrases to which one most often refers to in a given discourse.

For this reason, we propose the investigation of a mention detection procedure based on named entity annotations (further called Mention Detection based on Named Entity Structure (mdNES)). mdNES is a rather simple, rule-based method that relies solely on entity annotations in the data set. It defines a

language	token	NE	mentionID	mdNES
Catalan	Jordi_Virallonga	(person	(8	(1
	,	(person	_	(2
	director	_	_	_
	de	_	_	_
	l'	(org	(1	(3
	Aula	org)	1)	3)
	,	person)|person)	8)	2)|1)
Dutch	Frans	PER	(2076	(1
	Ferdinand	PER	2076)|2114)	1)
English	Denise	(person	(5	(1
	Dillon	person)	5)	1)
German	Pedros	_	(506|(504)	_
	Frau	_	_	_
	Mari-Gaila	_	506)	_
Italian	Mao	(person	(67	(1
	Asada	person)	67)	1)
Spanish	la	(person	(28	(1
	pareja	_	_	_
	Sandon_Stolle-Mark_Woodforde	(person)|person)	28)|7)	(2)|1)

Table 5.1: Examples from the NE annotations within the SemEval-2 shared task datasets for all six languages with added mdNES annotations. Column *mentionID* lists the boundaries for the gold mentions in the data.

mention for each existing named entity. In order to visualize how mdNES would detect mentions for the excerpts presented in table 3.7 on page 64, we add another column to this table, as shown in table 5.1. Comparing columns *mdNES* (containing the output of the mdNES method) and *NE* (listing the named entity annotations in the data), there are several peculiarities that we should note. Because named entities can be embedded according to the annotation of some languages (e.g. Catalan, Spanish, etc.) and not according to the annotations of others (e.g. Dutch, English, etc.), we allow for the identification of a separate mention for each NE independently of its level of embedding. In case no NE information is provided the mdNES method is not capable of detecting any mentions. Therefore, we do not expect this method to lead to exceptionally good results in a multilingual context, because there is no guarantee that NE information will always be available. Additionally, for the simple reason that in most languages pronouns are not marked as named entities, and therefore

#	token	POS	mentionID	mdPOSP train	mdPOSP test
0	Edgar	NNP	(19	NNP NNP	(1|(2)
1	Medina	NNP	19)		(3)|1)
2	read	VBD	-		-
3	the	DT	(171	DT NNS	(4
4	books	NNS	171)		(5)|4)
5	and	CC	-		-
6	says	VBZ	-		-
7	Williams	NNP	(137)	NNP	(6)
8	convinced	VBD	-		-
9	him	PRP	(19)	PRP	(7)
10	to	TO	-		-
11	stay	VB	-		-
12	in	IN	-		-
13	school	NN	-		(8)
14	/.	.	-		-

Table 5.2: An example sentence from the SemEval-2 shared task English dataset. Column *mentionID* lists the *gold* mentions; *mdPOSP train* – the POS patterns that the mdPOSP method extracts and column *mdPOSP test* shows the corresponding output of the mdPOSP method.

they will not be labeled as mentions, the mdNES will lead to low recall which is undesirable for the mention detection subtask in the context of the CR task.

5.1.2 *Mention Detection based on Part of Speech Patterns*

Mention Detection based on Part of Speech Patterns (mdPOSP) is the second method that we propose in our investigation. It is a heuristic method designed to identify and extract patterns based on the part-of-speech tags in the data. This method is comparatively simple, has a straightforward implementation and does not contain any language specific designations; for this reason we consider it a baseline. The fact that the only required type of linguistic annotation for it to function is POS information, which is available for a wide range of languages, renders mdPOSP highly applicable in a multilingual environment. In general, mdPOSP uses the training data and for each *gold* mention in this data, mdPOSP extracts and memorizes a pattern of POS tags of the words building the given mention (in other words a POS pattern constitutes a concatenation of the POS tags of each of the tokens in the mention).

In order to exemplify mdPOSP's functionality, let us have a look at table 5.2. In column *mdPOSP train*, we list the POS patterns that the method extracts from the labels of the training *gold* mention (listed in column *mentionID*). Column *mdPOSP test* shows the mention boundaries that the method would assign to that sentence if it was a test instance. All patterns seen in the training dataset are stored in memory and used for detecting mentions from the test data. Analogically to the mdNES approach, if mentions are embedded, mdPOSP will extract a pattern for each embedded layer. This fact can be seen to have both positive and negative aspects, because mdPOSP will certainly extract all possible observed POS combinations, yet, the accumulated patterns will lead to excess overgeneration. The latter leads to high recall, which is desirable for CR, because only the mentions that were discovered by the MD module can be used further by the coreference pipeline (see further section 5.2.1 for more detailed clarification). Yet, we assume that the induced mentions will be beyond an acceptable threshold, meaning that generating too many mentions will lead to an increased difficulty for the coreference resolver, since it will have more mention pairs to work with. Moreover, this method will also detect low frequency patterns that correspond either to idiosyncratic phrases or to annotation errors. In order to filter such unsuccessful and thus undesirable patterns, manually designed rules will need to be created individually for each language, which will drastically decrease the multilingual flexibility of that approach. Another, multilingual-friendly strategy that we propose to tackle the problem is filtering. Filtering can be used in order to decrease the overall overgeneration by excluding infrequent patterns and keeping only the ones that occur at least n number of times within the set of patterns extracted from the training set. We test both, the unmodified and the filtered versions of this method. In our experiments, we vary n between 5, 10, and 20. We also increased n to 30, but this filter showed a drastically detrimental performance than n=20 and thus we do not report those scores. Altogether, mdPOSP requires only POS information and since it does not make any additional assumptions and abstractions over the data, it can be used with any language for which part-of-speech annotation is available. Additionally, we note that the type of the POS tagset in use and the granularity of its tags are also important for this method to a great extent. For example, tagsets that consist of morphologically rich tags may lead to more overspecified patterns with a different distribution and frequencies across the dataset than patterns built from underspecified tags. In such cases, if needed, stripping the morphological information can be used to reach the expected performance.

5.1.3 *Mention Detection based on Dependency Structure*

Mention Detection based on Dependency Structure (mdDS) is the next method that we would like to include in our comparison. Similar to the mdPOSP ap-

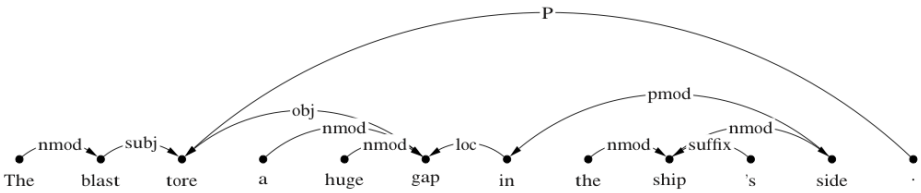

Figure 5.2: The dependency structure and relations for the sentence *"The blast tore a huge gap in the ship's side."* from the SemEval-2010 English test dataset.

dependency structure
dependency relation

proach, mdDS is a rule-based method that employs syntactic ***dependency structures*** as the basis for mention detection. In general, dependency structures determine the ***dependency relations*** (the labels within a dependency structure that describe the relation between the words in a sentence) between each syntactic head in a sentence and all its dependents. An example of a dependency structure can be seen in figure 5.2. While most mention detection methods first extract the whole phrase as a mention and then identify its syntactic head, the mdDS approach first selects the syntactic heads of NPs and then collects all the dependents via the labels of the dependency annotations. This process is not as straightforward and trivial as using constituency annotations, for example, since noun phrases are not directly represented in this annotation layer. The syntactic heads are identified within the concept of the mdDS method by the use of a predefined list of dependency relations. Possible relations to consider are: subject, direct object, or prepositional modifier (the noun phrase inside a prepositional phrase), etc, because noun phrases most often function as the subject, object, complement of pre-/postposition of the verb in a sentence. Once the set of relation labels used to identify the heads is assembled, the mdDS method uses it during mention detection to extract first those heads and then all their dependents in the test data. The dependents are extracted by following all relations pointing to the syntactic heads and building an ordered set of dependents for each of them. The successful identification of all dependents results in the detection of a mention representing the sequence of members of the ordered set for the given head.

In order to exemplify this procedure, let us look at table 5.3. Within that toy sentence we search for all tokens (listed in column *token*) that are either nouns or pronouns. In order to identify them as such, the POS tags provided by the annotations (in column *POS* in table 5.3) are used. Further, only the nouns and pronouns that have a label which is a member of our predefined set of labels (the labels are shown in column *DepRel*) are selected, because only those can be heads of noun phrases. As predefined set of target labels, let us assume the toy set $S_t = \{\text{SUBJ}, \text{OBJ}, \text{NMOD}, \text{PMOD}\}$.

#	token	POS	Head	DepRel	mentionID
1	Betsy	NNP	2	NAME	(184
2	Rogers	NNP	3	SBJ	184)
3	teaches	VBZ	0	sentence	_
4	first	JJ	7	NMOD	(349
5	and	CC	4	COORD	_
6	second	JJ	5	CONJ	_
7	grade	NN	8	NMOD	_
8	students	NNS	3	OBJ	349)
9	and	CC	3	COORD	_
10	leads	VBZ	9	CONJ	_
11	Alabama	NNP	10	OBJ	(371)
12	.	.	3	P	_

Table 5.3: An example sentence from the SemEval-2 shared task English dataset. The column *Head* lists the IDs of the heads for each token, *DepRel* includes the dependency labels and column *mentionID* shows the set of *gold* mentions for the sentence.

Following, the heads that we can identify from table 5.3 are listed in table 5.4 (where # is the column containing the token IDs and column *Head* indicates the token ID of the head of that current token). For each of these heads we collect all dependents. This is done by following the IDs in column *Head* (i.e. if we want to find all dependents of the head *Alabama*, we search for all tokens that contain the ID 11 in their *Head* column, since the token ID of *Alabama* is 11). Since this process should be repeated ***recursively*** (repeated over and over again, given a condition is true, until the process cannot be repeated further (until the base case is reached)) not only the direct dependents are gathered,

recursive

#	token	POS	Head	DepRel	mentionID
2	Rogers	NNP	3	SBJ	184)
7	grade	NN	8	NMOD	_
8	students	NNS	3	OBJ	349)
11	Alabama	NNP	10	OBJ	(371)

Table 5.4: A list of all extracted heads from the example sentence in table 5.3, which serve as a starting point of the identification of mentions for this excerpt.

#	token	POS	Head	DepRel	mentionID	mdDS
1	Betsy	NNP	2	NAME	(184	(1
2	Rogers	NNP	3	SBJ	184)	1)
3	teaches	VBZ	0	sentence	_	-
4	first	JJ	7	NMOD	(349	(2|(3
5	and	CC	4	COORD	_	-
6	second	JJ	5	CONJ	_	-
7	grade	NN	8	NMOD	_	3)
8	students	NNS	3	OBJ	349)	2)
9	and	CC	3	COORD	_	-
10	leads	VBZ	9	CONJ	_	-
11	Alabama	NNP	10	OBJ	(371)	(4)
12	.	.	3	P	_	-

Table 5.5: The output of the method mdDS for the example sentence in table 5.3, listed in column *mdDS*.

but all their dependents too. The complete spans of the mentions identified by the mdDS approach for the heads in table 5.4 are listed in the *mdDS* column in table 5.5.

Since, mdDS is not an approach based on statistics, it does not require training data and can directly be applied on the test set. Yet, it does require dependency annotation in the test sentences as well as previous knowledge about the format of that annotation so that the set of rules can be assembled.

mdDS is a highly efficient and easy to implement approach, yet, there are also some drawbacks that can be attributed to it. One disadvantage is the fact that mdDS extracts only one mention per head – the longest matching span. Additionally, dependency structure annotations are not always available for all languages which is inconvenient within the multilingual context. Another such complexity is the fact that the set of initial relation labels for the identification of the syntactic heads, must be manually defined for each separate language, defeating the purpose of easy adaptation in a multilingual context. Still, on the positive side, mdDS is much more trustworthy than mdPOSP for it relies on the dependency structure and not on POS information. For example, let us examine the phrase *Washington State* listed in table 5.6. mdPOSP is not able to distinguish between both tokens, because they are both proper nouns according to the part-of-speech tags in column *POS*. Yet, mdDS uses the labels in the *DepRel* column, which contains different labels for both tokens, depicting the difference between the head of the phrase and its dependent.

#	token	POS	Head	DepRel	mentionID
17	Washington	NNP	18	NAME	(1
18	State	NNP	16	CONJ	1)

Table 5.6: The dependency structure annotations (columns *Head* and *DepRel*) provided for the noun phrase *Washington State* in the SemEval-2 English dataset.

5.1.4 *Mention Detection based on Constituent Parse*

Dependency information is an efficient but not the only and surely not the most intuitive way to represent the syntactic structure of a sentence and correspondingly its nominal phrases. In figure 5.2 on page 94, we gave an example of the dependency structure for the sentence *"The blast tore a huge gap in the ship's side."*. In order to extract nominal phrases from such a structure mdDS relies on a predefined set of dependency labels and employs a search to recursively collect all tokens dependent on the ones having those labels. This is so, because dependency parses are ***syntactic parses*** that do not explicitly show the different parts of the sentence. ***Syntax*** in general provides various rules that allow the combination of words into different sentence components and as well rules to combine those components into complete sentences.

syntactic parse
syntax

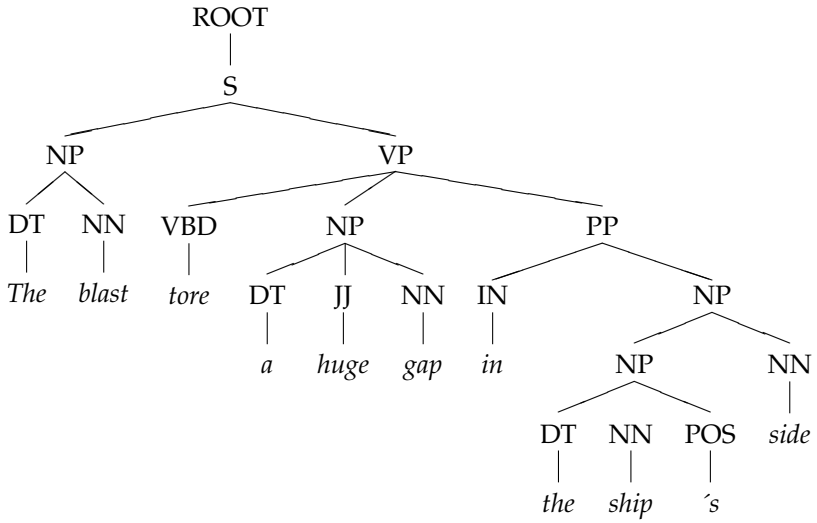

Figure 5.3: Syntactic parse in the form of a constituency-based parse tree for the sentence *"The blast tore a huge gap in the ship's side."* from figure 5.2 on page 94.

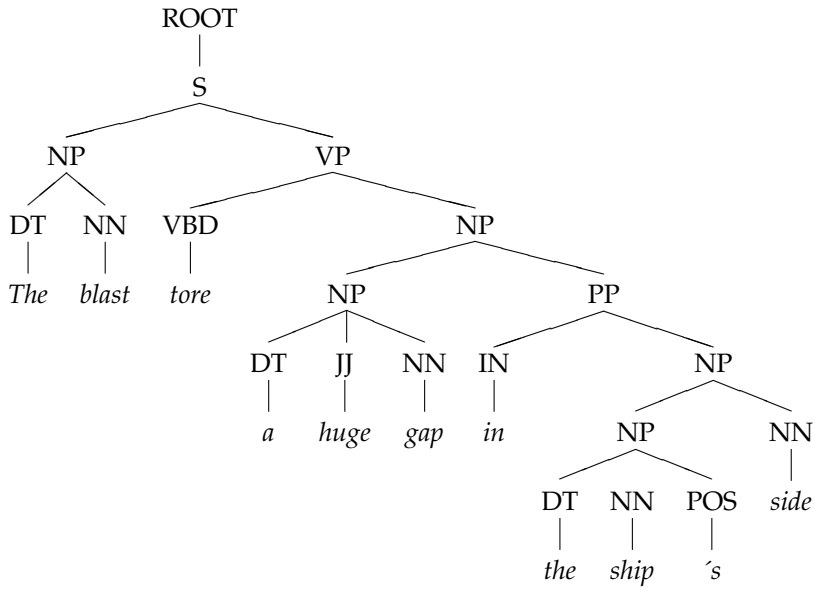

Figure 5.4: A variation of the syntactic parse with a lower PP attachment for the sentence *"The blast tore a huge gap in the ship's side."* in figure 5.2.

phrase structures
constituent

Instead of representing a sentence in terms of predicates, their arguments and the variation of relations between them, ***phrase structures*** provide us with a direct representation of all ***constituents*** (in phrase structure a constituent can be a single word, a group of words or even a whole clause that is represented as one unit in a given hierarchical structure) within the given sentence. A syntactic parse, as the one presented in figure 5.3, depicts all nominal phrases as constituents labeled with *NP* in their structure.

syntactic annotation

Even though ***syntactic annotation*** is one often used linguistic annotation layer provided for a wide range of languages, there can be various problems with using this layer for multilingual mention detection. One such problem is the overlap of mention boundaries to the used annotation scheme. For example, let us look at structural ambiguity. Constituent parsers may have a preference for higher or lower PP attachment, which results in a difference in the output structure and thus a difference in the resulting structure of the phrases. One such variation is displayed by figure 5.4 in which the PP attachment leads to an additional NP with respect to the phrases presented in figure 5.3.

Altogether, such variations are not problematic for parsers and thus annotations are consistent when they are achieved in an automatic way. Yet, if mentions are formed selecting one annotation scheme, and the phrase structure or dependencies are labeled according to another, the mismatch will lead to decrease in system performance.

#	token	POS	constituents	mentionID
1	The	DT	(ROOT(S(NP*	(1
2	blast	NN	*)	1)
3	tore	VBD	(VP*	_
4	a	DT	(NP*	(2
5	huge	JJ	*	_
6	gap	NN	*)	_
7	in	IN	(PP*	_
8	the	DT	(NP(NP*	(3(4
9	ship	NN	*	_
10	's	POS	*)	4)
11	side	NN	*))))	3)2)

Table 5.7: An example of mismatch of syntactic annotations and mention boundaries caused by a difference in the PP attachment for the sentence "The blast tore a huge gap in the ship's side." (see mention 2 and the noun phrase "*a huge gap*"). The example is extracted from the SemEval-2 English test dataset.

For example, let us assume that the annotations contain a structure, such as the one in figure 5.3, visualized in column *constituents* in table 5.7. This structure has a higher PP attachment leading to the shorter noun phrase *a huge gap*. However, the mention boundaries provided by the SemEval-2 shared task for that sentence are listed in column *mentionID*. They do not opt for higher PP attachments since mention 2 includes the prepositional phrase – leading to a mismatch between the noun phrases in the constituency structure and the derived mentions.

The CoNLL 2012 shared task provided constituent parses (see figures 5.3 and 5.4) on the basis of which a different mention detection method can be employed. Mention Detection based on Constituent Parse (mdCP) is a rule-based approach that uses a given constituent parse of a sentence to extract a mention for each separate NP of its structure. In table 5.8, we show an example sentence from the CoNLL 2012 English dataset. Column *Parse bit* of the table represents the constituent parse of the sentence within a bracketed structure. In column *mdCP* the output of the mdCP module is provided. The method is then applied to all three languages in the CoNLL 2012 shared task with the only difference that an additional and relatively straightforward, language specific modification was added for English that detects a mention for every possessive pronoun. This modification was not needed for Arabic and Chinese, because within those languages possessive pronouns were already separately marked as NPs. The last column of table 5.8, *mentionID*, lists the *gold* mentions labeled in the data. From a comparison between columns *Parse*

#	token	POS	Parse bit	mdCP	mentionID
0	The	DT	(TOP(S(NP(NP*	(1(2	(1
1	world	NN	*	_	_
2	's	POS	*)	2)	_
3	fifth	JJ	*	_	_
4	Disney	NNP	*	_	(2)
5	park	NN	*)	1)	1)
6	will	MD	(VP*	_	_
7	soon	RB	(ADVP*)	_	_
8	open	VB	(VP*	_	_
9	to	IN	(PP*	_	_
10	the	DT	(NP*	(3	_
11	public	NN	*))	3)	_
12	here	RB	(ADVP*)))	_	_
13	.	.	*))	_	_

Table 5.8: An example sentence from the CoNLL 2012 English dataset. Column *Parse bit* includes the constituency parse for the sentence, column *mdCP* – the output from the mdCP method and column *mentionID* shows the boundaries for the *gold* mentions in the data.

bit and *mentionID*, we can note that the key mentions correspond directly to either an existing noun phrase or a named entity (as the named entity *Disney*) from the annotation. This phenomenon was observed consistently within all 3 languages of the CoNLL 2012 shared task. Another observation that we can make explicit is that not all noun phrases correspond to a mention – only mentions that are not singletons are included in the key set. We previously introduced this problem in section 3.2.2.3. For this reason, extracting all NPs from the constituent structure should be approached in order to cover both singletons and coreferent mentions. Then, after the resolution process, the system can remove all mentions that it classifies as singletons.

5.1.5 *Mention Detection based on IOB Annotation*

Attempting to address the disadvantages of both mdPOSP and mdDS we continue our search for an efficient and reliable method for mention detection within a multilingual context of the task of coreference resolution. For this reason we turn to the alternative of rule-based approaches and propose a method based on machine learning techniques. Machine learning will provide us with more flexibility and adaptability of the mention detection module for new languages

and datasets. Thus, we will employ these techniques for the development of a mention detection method that does not need a predefined set of rules and is able to abstract away from the used training data. We start with the concept of ***IOB tagging*** that was first introduced by Ramshaw and Marcus [1995]. The authors implemented that approach for identifying noun chunks, while Veenstra and Buchholz [1998] extended it further. The idea behind IOB tagging is that each token within a sentence is labeled with one of the following classes: B (meaning that the current token is the beginning of an existing chunk), I (indicating that the token is inside of a chunk), and O (used when the token is outside of the boundaries of any chunk).

IOB tagging

As simple as it is, however, IOB tagging is not directly applicable to the mention detection problem for CR. ***Chunking*** or more often known as ***NP chunking*** is the process of identifying flat structures in text (most often NPs, but chunks are not restricted to a specific category). The identified phrases are flat in the sense that they do not contain other embedded phrases in them, they are non-recursive [Abney, 1991]. Mentions, however, are not flat structures. They may be embedded as mentions 2, 3 and 4 are in example (35) extracted from the SemEval-2 shared task English dataset. For this reason, IOB tagging needs to be redefined in order to be used for the multilingual MD subtask of coreference resolution.

chunking

NP chunking

(35) [She$_1$] is [an accomplished teacher with [22 years in [the profession$_2$]$_3$]$_4$].

In the current section, we define a variation of the IOB approach that allows the detection and extraction of recursive structures within text and is thus better applicable to MD. To our knowledge, only one variation of the IOB tagging is targeted at such embedded phrases – the one presented by Tjong Kim Sang [2001]. The author did not use one classifier but rather a full cascade of classifiers targeting a predefined structure type. He trained a classifier for each level of embedding starting from base NPs. Once the latter are identified, a classifier is trained for the recognition of the phrases that contain only one embedded structure. The process is repeated until a target level of embedding is reached.

Within the corpora distributed by the SemEval-2 and CoNLL 2012 shared tasks, most languages differ in the level of embedded phrases ranging approximately between 3 and 13[2]. Such variation in the embedding of the phrases makes the approach presented by Tjong Kim Sang [2001] not easily applicable to a multilingual setting and in the cases of high embedding (e.g. Catalan, English and Spanish) the computational complexity and lack of training data for each of the levels is highly inefficient (as Tjong Kim Sang [2001] discusses, the targeted level of embedding is directly correlated with the possibility to

[2] Approximate distributions in SemEval-2 (Catalan – 10, Dutch – 3, English – 13, German – 5, Italian – 10, Spanish – 9) and CoNLL 2012 (Arabic – 5, Chinese – 4, English – 5)

extract training examples – the higher the level of embedding, the less examples one can extract).

The variation of IOB tagging that we propose, further referred to as Mention Detection based on IOB Annotation (mdIOBA), incorporates all levels of embedding of phrases of different type within one single classifier which makes it comparatively more efficient and usable for multilingual mention detection than the method proposed by Tjong Kim Sang [2001]. Similar to the concept of ***regular expressions***, or more general ***finite state patterns***, that are used to define the state of objects or text, we redefine the set of accepted IOB labels. We extend this set (I, O, B) to a larger set – instead of label *I*, we allow for I^+ (e.g. *I, II, III, IIII,* etc.). *II*, for example, denotes a token that is inside two mentions simultaneously. Furthermore, instead of label *B*, we allow for B^+ (e.g. *B, BB, BBB, BBBB,* etc.). Similar to *II*, the label *BB* can be assigned to a token that is the beginning of two mentions at the same time, *BBB* indicates the beginning of three mentions, etc. Additionally, the combination of the *I* and *B* tags within the original IOB tagging approach, namely *IB*, does not exist, because there can be only one or the other in use. We add this label to the set of allowed labels and extend it as I^+B^+ (e.g. *IB, IIB, IBB, IIIB,* etc.). Analogically to the original IOB definition, the labels preserve their meaning, but in mdIOBA each symbol in a label also represents a level of embedding. Consequently, the label *IB* means that the current token is inside one mention and the beginning of another, *IIB* denotes a token that is in two distinct mentions and the beginning of a third mention, etc. There is no restriction to the number of labels that can be used by the method, yet that number depends solely on the level of embedding existing in the training instances for the given language. Each level of embedding n (apart from a flat structure, which has 3 labels) results in $n + 1$ additional labels. The full set of labels together with the corresponding frequency with which every label occurs as observed in the training sets for all languages within the SemEval-2 shared task, is listed in table 5.9.

regular expressions
finite state patterns

In order to visualize the functionality of the mdIOBA approach, let us look at the example in table 5.10. In column *mentionID* we list the *gold* mentions provided by the task, column *mdIOBA train* contains the labels that the mdIOBA method induces from the *gold* mentions for training a classifier. In the sentence *"Tragedies that test our strength and our resolve occur and have occurred in the past."*, it can be seen that the complexity of used mdIOBA labels derived from the *gold* mentions (column *mentionID* in the table) is already relatively high – *IBBB*, *IIBB*. This accounts for the complexity of the labels in the mdIOBA output as well. However, in ML, each distinct label in the training set represents a different class. Therefore, the higher the number of labels, the bigger the set of classes that the classifier needs to choose from. Moreover, for each of the labels, the classifier needs to have enough training examples in order to learn the proper representation of that class. For this reason, an increased set of labels can only be used when an abundant amount of training data is present or a

#	tag	CA	DU	EN	GE	IT	SP
1	B	34316	4916	12716	72339	12000	38384
2	BB	2224	39	1484	3300	563	2887
3	BBB	156	1	43	9	6	138
4	I	74179	10423	25672	93464	23872	82429
5	IB	21508	766	5232	21002	8028	23348
6	IBB	1341	4	752	740	467	1715
7	IBBB	87	-	23	4	3	62
8	IBBBB	2	-	1	-	-	1
9	II	34213	993	10529	25527	14153	40148
10	IIB	9852	51	1931	4614	4154	11064
11	IIBB	638	3	262	174	241	755
12	IIBBB	30	-	7	-	-	24
13	III	13581	49	3823	4701	6764	16219
14	IIIB	3754	2	611	748	1829	4186
15	IIIBB	220	-	86	18	84	236
16	IIIBBB	10	-	2	-	1	5
17	IIII	4868	-	1210	780	2807	5664
18	IIIIB	1235	-	168	112	670	1380
19	IIIIBB	80	-	28	3	34	82
20	IIIIBBB	3	-	1	-	-	1
21	IIIII	1412	-	302	85	942	1688
22	IIIIIB	361	-	48	8	251	382
23	IIIIIBB	17	-	13	-	12	26
24	IIIIII	389	-	110	4	380	451
25	IIIIIIB	106	-	16	1	85	100
26	IIIIIIBB	3	-	7	-	1	6
27	IIIIIII	93	-	34	-	140	85
28	IIIIIIIB	16	-	6	-	22	18
29	IIIIIIIBB	1	-	4	-	2	2
30	IIIIIIII	16	-	19	-	62	22
31	IIIIIIIIB	5	-	4	-	12	10
32	IIIIIIIIBB	1	-	1	-	-	-
33	IIIIIIIII	4	-	7	-	16	9
34	IIIIIIIIIB	3	-	1	-	4	-
35	IIIIIIIIIBB	-	-	1	-	-	-
36	IIIIIIIIII	-	-	4	-	5	-
37	IIIIIIIIIIB	-	-	1	-	-	-
38	IIIIIIIIIIBB	-	-	1	-	-	-
39	IIIIIIIIIII	-	-	1	-	-	-
40	IIIIIIIIIIIBB	-	-	1	-	-	-
41	IIIIIIIIIII	-	-	2	-	-	-
42	IIIIIIIIIIIB	-	-	1	-	-	-
43	O	90853	38811	30670	177126	20308	97105

Table 5.9: A full list of all IOB tags and the frequencies with which they occur across the training sets of all six languages (CA(talan), DU(tch), EN(glish), GE(rman), IT(alian), SP(anish)) within the SemEval-2 shared task.

#	token	POS	mentionID	mdIOBA train
1	Tragedies	NNS	(69	B
2	that	WDT	-	I
3	test	VBP	-	I
4	our	PRP$	(56 | (50 | (42)	IBBB
5	strength	NN	50)	III
6	and	CC	-	II
7	our	PRP$	(55 | (42)	IIBB
8	resolve	NN	55) | 56)	III
9	occur	VBP	-	I
10	and	CC	-	I
11	have	VBP	-	I
12	occurred	VBN	-	I
13	in	IN	-	I
14	the	DT	(66	IB
15	past	NN	66)	II
16	.	.	69)	I

Table 5.10: An example sentence from the SemEval-2 shared task English training dataset with the output from the mdIOBA method given in column *mdIOBA train*.

higher chance of data sparseness and thus error-prone system performance can occur. As we showed in table 5.9 each of the languages in the SemEval-2 shared task lead to a different number of used classes – in Catalan there are 35 unique labels, Dutch has only 12, English – 43, German – 21, Italian – 32 and Spanish – 33. Those numbers show that mdIOBA will face a different level of difficulty across the various languages – Dutch being the easiest to resolve with the lowest variation across the labels and English being the hardest as it has 43 different classes.

The feature set that the mdIOBA classifier uses includes both POS and dependency information for a context of 5 words before and after the target word. A full list of the features that the mdIOBA method uses is provided in table 5.11. A big advantage of ML and thus respectively of mdIOBA is that it is flexible in regards to the use and need for annotations and can be employed successfully only on POS information as well as only on dependency annotations.

The classification procedure is structured as follows. A feature vector containing all features is built for each of the tokens in both the training and the test set. For the learning process we make use of MBL (see section 4.1). The resolver aims at labeling the test instances according to the ones already seen

#	Feature Description
1	the target word
2	part-of-speech tag of the target word
3	dependency label of the syntactic head of the target word
4	part-of-speech tag of word$_{-5}$
5	part-of-speech tag of word$_{-4}$
6	part-of-speech tag of word$_{-3}$
7	part-of-speech tag of word$_{-2}$
8	part-of-speech tag of word$_{-1}$
9	part-of-speech tag of word$_{+1}$
10	part-of-speech tag of word$_{+2}$
11	part-of-speech tag of word$_{+3}$
12	part-of-speech tag of word$_{+4}$
13	part-of-speech tag of word$_{+5}$
14	dependency label of word$_{-5}$
15	dependency label of word$_{-4}$
16	dependency label of word$_{-3}$
17	dependency label of word$_{-2}$
18	dependency label of word$_{-1}$
19	dependency label of word$_{+1}$
20	dependency label of word$_{+2}$
21	dependency label of word$_{+3}$
22	dependency label of word$_{+4}$
23	dependency label of word$_{+5}$

Table 5.11: The full list of features used by the mdIOBA classifier consisting in general of POS and dependency information for a context window of 5 words before and after the target token.

in the training data. The output of the module is the test set labeled with IOB tags, which can be later converted to the bracketed form of mention boundaries. Yet, the mdIOBA classifier assigns a class to each word separately and independently form its previous decisions or the upcoming instances. Thus, the resulting mdIOBA labels can lead to inaccurate bracketing structures for the final mentions. For that reason, a postprocessing step is needed to ensure and validate mention boundary integrity. In that step we discard all mentions that do not have an opening bracket and the mentions that have been opened, but never closed, are terminated at the sentence boundary of the sentence in which they are found.

5.1.6 *Mention Detection via a Voting Technique*

Presenting various distinct methods that can be applied for mention detection within multilingual coreference resolution, we showed both their advantages as well as their disadvantages when applied to multiple languages. With the latter we indicated as well the differences between the approaches that lead to variations in their output. However, in the search for an optimal approach we implement as well a hybrid method that is a combination of all previously presented ones and thus combines a selection of their characteristics. Mention Detection via a Voting Technique (mdVOTE) assesses the mentions detected by mdPOSP (employing filter 5), mdDS and mdIOBA. We exclude mdNES from our investigation, because named entity annotations were not provided consistently across the datasets[3]. Yet, we include mdPOSP, mdDS and mdIOBA, because every additional method enriches the information considered by mdVOTE and thus improves on its output. This is so, because mdVOTE combines the mentions from the various methods as voting candidates for its own output. Moreover, a voting technique would require at least two candidates to be able to chose from, but an even number of candidates can easily lead to the occurrence of ties. In these cases, weighted techniques can also be applied. Since we use an uneven number of methods, weights are not necessarily needed and therefore not included. The approach incorporates the outputs into a voting scheme in which each of the distinct methods is entitled to an equal vote. mdVOTE evaluates the separate choices on per-word-basis and outputs the majority vote as its label.

Similar to mdIOBA, mdVOTE determines a class on per-word-bases, independently of previously made decisions and upcoming tokens. Analogically to the postprocessing step for the output produced by mdIOBA, mdVOTE labels can be corrected as well. Yet, the fact that mdVOTE combines three different approaches can lead to the emergence of mentions that do not correspond to any grammatical phrase, because the type of mentions produced by each of the separate methods is considerably different (i.e. mdVOTE will either inherit the type of errors or merging two mentions with wrong boundaries may lead to a new and still incorrect resulting mention). For this reason, the mdVOTE approach can lead to highly erroneous mentions and thus to a decrease in system performance overall. In table 5.12, we provide an example of the output produced by mdVOTE given the output of mdPOSP, mdDS and mdIOBA for a sentence extracted from the SemEval-2 shared task English dataset.

[3] Dutch contained only automatically acquired NE labels. Catalan, English, Italian and Spanish included only manual NE annotations and the German dataset did not provide any.

#	token	mdPOSP	mdDS	mdIOBA	mdVOTE
1	This	_	_	(0)	_
2	is	_	_	_	_
3	The	(2	(3	(1	(1
4	World	2)	_	1)	1)
5	,	_	_	_	_
6	a	(5	_	_	_
7	co-production	5)	_	_	_
8	of	_	_	_	_
9	the	(7|(3	(0	(2	(2
10	BBC	_	_	_	_
11	World	3)	_	_	_
12	Service	7)	_	2)	2)
13	,	_	_	_	_
14	PRI	(6	_	(3|(4	(3
15	and	_	_	4)	_
16	WGBH	6)	(1	(5	(4
17	in	_	_	_	_
18	Boston	_	0)|1)|(2)|3)	(6)|5)|3)	(5)|4)|3)
19	.	_	_	_	_

Table 5.12: An example sentence from the SemEval-2 shared task English dataset. Column *mdPOSP* lists the output of the mdPOSP method when filter 5 is used, column *mdDS* shows the output from the mdDS method, column *mdIOBA* includes the output of the machine learning method mdIOBA and the last column *mdVOTE* depicts the result from combining all methods via the hybrid approach mdVOTE.

5.1.7 *Applicability of the Mention Detection Methods within Both Multilingual Shared Tasks:* SemEval-2 *and* CoNLL 2012

As we noted in the introductory part of section 5.1, the diverse mention detection methods, presented so far, constitute algorithms developed for a specific type and format of annotations (e.g. POS, dependency information, NEs, etc.). All these layers were provided in the SemEval-2 or CoNLL 2012 shared tasks. Most of the approaches that we presented depend to a great extent on the layer of annotation and will be rendered unusable within a different setting. However, in real-world situations, it is most certain that a CR system will not be provided with the exact same data format and annotation layers as in both shared tasks. This was already proven by the SemEval-2 multilingual

method	SemEval-2	CoNLL 2012
mdNES	√	√
mdPOSP	√	√
mdDS	√	–
mdCP	–	√
mdIOBA	√	√
mdVOTE	√	√[4]

Table 5.13: Overview of the aplicability of the diverse mention detection methods on the datasets of the two multilingual shared tasks SemEval-2 and CoNLL 2012. √ indicates that the method is applicable (e.g. the necessary annotations are provided) and – that is not.

follow-up – CoNLL 2012 which attempted to be highly compatible with the SemEval-2 shared task by adopting the format of the data (e.g. each token is provided on a separate line and for each layer of annotation a separate column is used). Yet, the format of the data is not the only factor that needs to be considered.

CoNLL 2012 did not provide dependency information within its layers of annotation. For this reason mdDS is not applicable to that set of data. However, CoNLL 2012 included a different syntactic annotation, namely phrase structure. This allowed us to employ mdCP within that setting. An overview of the applicability of all methods presented in this chapter is given in table 5.13.

The availability of annotations and the applicability of the various methods in the different settings is highly important to our work. Comparing methods for multilingual MD, which is our aim further down in the chapter, we note that rule-based methods can lead to a better system performance. This is only so if annotations are provided that can unambiguously lead to direct identification of phrase structures or more specifically of nominal phrases. Unavailability of annotation layers in such cases, however, instantaneously affects multilinguality, which is crucial to our goals.

5.2 EVALUATION OF MULTILINGUAL MENTION DETECTION

In sections (section 5.2.2 and section 5.2.3) we report on two evaluation settings for the methods presented in section 5.1. We note, that all approaches are only comparable if they are evaluated on the same data. Thus, we divide the evaluation in two parts according to the dataset that is employed (being

[4] We note that the mdVOTE method will only be applicable if it uses mdCP instead of mdDS as one of its votees.

either the SemEval-2 or CoNLL-2012 data). Further, we evaluate how the proposed methods perform on their own, without considering the actual coreference performance of the system. Then, we integrate all approaches in the multilingual coreference pipeline and assess the overall system scores. However, we note that the achieved scores are not comparable across the two tasks not only because we employ different datasets but as well because various feature optimizations with regard to the underlying data sets (e.g. for SemEval-2 we consider a span of only three sentences as search space for coreference, while for CoNLL-2012 we increase this to 7 for Arabic and 10 for Chinese and English) are made use of. Such system improvements render the scores not comparable across the tasks. Furthermore, in both evaluation settings we will aim not only at quantitative but as well at qualitative analysis of the output and results. The latter will provide us with more knowledge about the type of annotations and annotation schemes needed for the development of efficient and robust multilingual approaches.

5.2.1 *Mention Detection Scoring*

We report system performance for all eight languages of both shared tasks (Arabic, Catalan, Chinese, Dutch, English, German, Italian, Spanish) via precision, recall and F-measure over all provided evaluation metrics (MUC, B^3, CEAF and BLANC (see section 2.3.4)). The scores in both intrinsic and extrinsic evaluations were acquired by version 1.04[5] of the scorer provided by and used during the SemEval-2 shared task as well as the newer version (4.0) of the scorer[6] (released for the CoNLL 2012 task) that is employed for the respective dataset. The two variants differ in their behaviour towards singleton mentions in the system output as discussed in section 2.3.4.3.

Before we present any system scores on mention detection, we would like to exemplify the effect of reducing singletons in the ***key set*** (the set containing the answers) on the overall scores and more precisely on precision, as reported by the evaluation software. For this purpose, let us consider a toy sentence extracted from the SemEval-2 English dataset, listed in table 5.14. Along with this sentence, we include four different sets of mentions available for it, marked as *FourM*, *ThreeM*, *TwoM* and *OneM* in table 5.14. Then, we present four evaluation settings with set *FourM* always being the test set and sets *FourM*, *ThreeM*, *TwoM* and *OneM*, being the key set, i.e. the gold standard. The scores that we can achieve by employing both the SemEval-2 and the CoNLL 2012 versions of the scorer (in this case both scorers achieve the same results) are listed in table 5.15. We list the F-scores across all four evaluation metrics. In column *ThreeM* of table 5.15 we can see the scores for using set *ThreeM* as the key set and set *FourM* as the system output. The only difference between the

key set

[5] http://www.lsi.upc.edu/~esapena/downloads/index.php?id=2
[6] http://conll.cemantix.org/2012/download/scorer.v4.tar.gz

#	token	FourM	ThreeM	TwoM	OneM
1	But	_	_	_	_
2	Eagle	(1)	_	_	_
3	said	_	_	_	_
4	the	(2	(2	_	_
5	financing	2)	2)	_	_
6	was	_	_	_	_
7	insufficient	_	_	_	_
8	and	_	_	_	_
9	sales	(3)	(3)	(3)	_
10	during	_	_	_	_
11	the	(4	(4	(4	(4
12	past	_	_	_	_
13	fiscal	_	_	_	_
14	year	4)	4)	4)	4)
15	sagged	_	_	_	_
16	.	_	_	_	_

Table 5.14: Four toy system outputs for scoring evaluation with both SemEval-2 and CoNLL-2012 scorers. Here set *FourM* is the test set in all cases and sets *FourM, ThreeM, TwoM, OneM* the key sets for which the number of mentions is gradually decreased.

two sets is the lack of mention 1 in set *ThreeM*. Similar to that is the decrease of mentions in the rest of the key sets.

The motivation for this example is the fact that key sets do not necessarily contain singletons, as the mentions in the *mentionID* column in table 5.7 on page 99. Yet, mention detection extracts all mentions – potentially coreferent and never coreferent mentions. The latter has a considerable effect on the reported precision, which can be falsely interpreted. Falsely, because there is no change in the actual precision of the mentions provided by the test set.

FourM			ThreeM			**TwoM**			OneM		
R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
100	100	100	100	75.00	85.71	100	50.00	66.66	100	25.00	40.00

Table 5.15: Five mention detection evaluations with *FourM, ThreeM, TwoM or OneM* used as key mentions and *FourM* as a test set.

Only the number of mentions is changed, which results in a false reduction in precision reported by the scoring software. If we again recall the definition of precision given in 2.1, that measure is defined by the ratio of the correct answers given by the system to the number of all given answers. Yet, in our toy example, all answers provided by the system in all evaluation settings are correct and therefore precision should not be reduced.

What can be seen from table 5.15 is that using set *FourM* as test set against sets *ThreeM, TwoM* and set *OneM* reduces precision. Yet, the later three sets do not differ in precision, or so to say, the mention boundaries for all included mentions are identical. We would like to clarify that precision in those cases is lower, not because the underlying mention boundaries are wrong, but because including more mentions in the test set than the mentions present in the key set conflicts with the definition of precision and leads to its substantial reduction. It is of utmost importance for us to discuss this issue here, because evaluating MD, outside of the coreference pipeline, means that we cannot rely on precision figures. Thus, evaluation in this stage completely ignores precision and respectively the F-measure calculated with it. Following, only recall is favored when mention detection is approached outside of the CR pipeline. In other words, the higher the recall values are, the better the MD method when evaluated outside of the coreference pipeline. We also need to favor recall and thus identify all potentially coreferent phrases, because mentions that are not identified by the MD module cannot be used in the CR pipeline further on. We will refer to this example later on in our discussion in sections 5.2.2 through 5.2.5.

5.2.2 SemEval-2 *Intrinsic Evaluation*

The following section provides both quantitative and qualitative analysis of the methods for mention detection that can be employed on the SemEval-2 datasets across all six languages (Catalan, Dutch, English, German, Italian and Spanish). We discuss extensively their applicability and effectiveness in a multilingual setting in the attempt to answer the research questions that we posed in the introduction of the chapter. Our main aim is to evaluate the approaches on the task of multilingual mention detection and to determine which approach provides the best performance in a multilingual setting.

5.2.2.1 *Quantitative Analysis*

Our quantitative analysis solely relies on the information that the SemEval-2 scoring software can provide. We present and analyze the results for two settings: *auto* (considering the automatically acquired linguistic annotations) and *gold* (employing only manually labeled data). Within the quantitative analysis evaluation, the most optimal outcome will be to achieve results that will clearly categorize one method as outperforming all the rest across all

targeted languages. Moreover, a method that consistently reaches surpassing scores within one language family will provide a higher guarantee that a random new targeted language from the same language family will lead to the highest performance across the MD methods.

In table 5.16, we present the scores within all evaluated settings (both *gold* and *auto* linguistic annotations). However, as we noted in section 5.1.3, mdDS only extracts one span per mention head, while mdIOBA does not follow such a restriction. For this reason, in order to achieve a better and more objective comparison between those approaches, in table 5.16 we include a modification of mdDS that allows for more than one possible span per head (i.e. excluding existing PP attachments and modifications of the target phrase instead of considering only the longest span). Additionally, we evaluate the initial version of mdDS against a modified version of mdIOBA, by restricting mdIOBA's output to select only the longest span per identified mention head. The respective scores are listed in table 5.17. We do not modify mdNES, since its output is merely mirroring the named entity annotations provided in the data. Furthermore, mdPOSP and all its variants are also excluded from this comparison, since we consider that method as a baseline heuristic and thus we do not increase its complexity, apart from the initially presented filters: considering no filtering at all (-), and $n = 5$, 10 and 20 for the frequency counts of filtered patterns respectively (i.e. $n = 5$ means that only patterns that are seen at least five times in the training set will be marked in the test set).

As we already noted in section 5.1.6, German does not provide any (*gold* or *auto*) named entity annotations and thus we cannot report system results for German within the mdNES method evaluation. This once more underlines the importance of provided and, in general, available layers of linguistic annotations for tasks, such as multilingual coreference resolution. The rest of the languages include evaluation of mdNES, depending on the annotations their datasets provide. Another peculiarity of the scores we provide is the lack of results within the *gold* setting for Italian, apart from the mdNES method. The reason for this is the lack of *gold* annotations in the Italian dataset. Similarly to Italian, Dutch does not provide any manually labeled information and thus that language is entirely excluded from the *gold* evaluation setting.

mdNES

The figures in table 5.16 give us the possibility to assess mdNES, which proves to be a hard task, since the output achieved by mdNES is exceedingly unbalanced (not all evaluation settings can be considered for lack of annotations in the data). This renders the method hardly comparable to all other presented approaches. Within the *auto* setting, only the results for Dutch can be compared across all MD methods and even though mdNES outperforms all mdPOSP variants in terms of F-scores, it does not achieve this in terms of recall, which is more important

			mdNES	mdPOSP				mdDS	mdIOBA	mdVOTE
				-	5	10	20			
auto	CA	R	-	54.74	40.38	35.54	32.70	**75.12**	50.65	51.69
		P	-	8.79	49.90	56.16	68.08	45.96	53.01	53.50
		F_1	-	15.15	44.67	43.53	44.18	57.03	51.80	52.58
	DU	R	31.00	51.42	36.33	32.88	23.60	**73.42**	42.12	54.25
		P	49.70	9.30	30.85	29.82	30.37	27.43	51.74	39.39
		F_1	38.18	15.76	33.37	31.27	26.56	39.94	46.44	45.64
	EN	R	-	52.18	37.89	33.65	19.65	**82.34**	59.59	57.19
		P	-	21.05	60.36	62.04	60.61	44.50	66.50	58.05
		F_1	-	30.00	46.55	43.64	29.68	57.77	62.85	57.61
	GE	R	-	49.93	42.44	41.00	35.49	**78.96**	67.32	66.39
		P	-	21.07	62.21	65.56	64.45	59.16	70.23	65.88
		F_1	-	29.64	50.45	50.45	45.77	67.64	68.74	66.14
	IT	R	-	52.38	34.32	27.23	19.35	**62.64**	41.40	40.55
		P	-	23.25	58.02	64.49	63.10	42.35	46.42	45.89
		F_1	-	32.20	43.13	38.20	29.62	50.54	43.77	43.06
	SP	R	-	55.08	41.54	38.41	31.08	**76.82**	50.51	44.80
		P	-	9.10	50.39	59.01	71.01	45.59	53.61	48.87
		F_1	-	15.62	45.54	46.53	43.24	57.22	52.01	46.75
gold	CA	R	25.23	38.90	40.40	38.93	32.20	**76.28**	53.58	53.03
		P	88.31	17.15	50.13	51.69	70.66	47.03	55.84	54.32
		F_1	39.24	23.81	44.74	44.41	44.24	58.18	54.69	53.67
	EN	R	20.64	52.65	37.89	33.65	19.65	**82.76**	63.95	58.92
		P	55.88	21.61	60.36	62.04	60.61	44.48	70.55	59.96
		F_1	30.15	30.68	46.55	43.64	29.68	57.86	67.09	59.43
	GE	R	-	45.22	42.24	39.33	35.49	**79.48**	71.18	68.50
		P	-	36.64	62.44	63.75	64.45	59.82	74.59	68.40
		F_1	-	40.48	50.39	48.65	45.77	68.27	72.84	68.46
	IT	R	**85.95**	-	-	-	-	-	-	-
		P	99.64	-	-	-	-	-	-	-
		F_1	92.29	-	-	-	-	-	-	-
	SP	R	24.34	39.20	41.15	35.63	31.08	**78.25**	54.36	47.94
		P	88.03	15.25	53.05	60.18	71.01	46.23	57.23	51.51
		F_1	38.14	21.96	46.35	44.76	43.24	58.12	55.76	49.66

Table 5.16: Mention detection across all six languages of the SemEval-2 shared task with all spans for each mention in both *auto* and *gold* settings. The highest recall is marked in bold. - marks the lack of annotations for the setting.

		mdNES	mdPOSP				mdDS	mdIOBA	mdVOTE
			-	5	10	20			
CA	R	-	54.74	40.38	35.54	32.70	**73.01**	46.57	51.77
	P	-	8.79	49.90	56.16	68.08	68.62	54.69	57.88
	F_1	-	15.15	44.67	43.53	44.18	70.75	50.30	54.66
DU	R	31.00	51.42	36.33	32.88	23.60	**71.15**	41.91	53.21
	P	49.70	9.30	30.85	29.82	30.37	35.93	51.86	38.70
	F_1	38.18	15.76	33.37	31.27	26.56	47.75	46.36	44.80
EN	R	-	52.18	37.89	33.65	19.65	**79.67**	56.59	56.80
	P	-	21.05	60.36	62.04	60.61	54.75	67.83	61.80
	F_1	-	30.00	46.55	43.64	29.68	64.90	61.70	59.19
GE	R	-	49.93	42.44	41.00	35.49	**78.77**	63.84	65.33
	P	-	21.07	62.21	65.56	64.45	67.69	71.13	70.77
	F_1	-	29.64	50.45	50.45	45.77	72.81	67.28	67.94
IT	R	-	52.38	34.32	27.23	19.35	**61.88**	29.86	35.03
	P	-	23.25	58.02	64.49	63.10	55.84	46.71	46.95
	F_1	-	32.20	43.13	38.20	29.62	58.70	36.43	40.12
SP	R	-	55.08	41.54	38.41	31.08	**74.63**	47.09	41.83
	P	-	9.10	50.39	59.01	71.01	71.14	55.43	49.13
	F_1	-	15.62	45.54	46.53	43.24	72.84	50.92	45.19

Table 5.17: Mention detection across all six languages of the SemEval-2 shared task considering the longest span per mention in both *auto* and *gold* settings. The highest recall figures are marked in bold. mdNES and mdPOSP scores are kept for comparison. - marks the lack of annotations for the setting.

to us in the context of this evaluation setting. Furthermore, mdNES does not outperform mdDS, mdIOBA and mdVOTE according to both observed criteria. With respect to the scores in the *gold* setting, mdNES is also outperformed by the rest of the methods in both recall and F-scores, with the exceptional case of Italian. The only mention detection that could be achieved for Italian, within the *gold* setting (we remind the reader that no *gold* annotations apart from named entities were included in the data for that language), reaches exceptionally high performance leading to F-score of 92.29% and a recall of 85.95%. This is due to the distinct annotation scheme for NEs (see section 3.2.1) that marks as an entity all named entity instances, all pronouns as well as abstract noun phrases. This significantly extends the semantic types of phrases that are included within that annotation layer in comparison to other languages. Thus, the overlap between the resulting NEs for Italian and the actual *gold* mentions is exceedingly high. However, the high performance of the mdNES method is not transferable to all languages and even less to both (*auto* and *gold*) evaluation settings. This

is so for two main reasons: First, not all languages have such an exceedingly high correspondence between the named entities provided in the data and the included *gold* mentions; second, named entities are not always available for both *auto* and *gold* annotation types. This is proven by the exceptionally high variation in the results achieved by the mdNES method across all six languages – being lowest for English with recall of 20.64% and highest for Italian with recall of 85.95%. Additionally, no parallel between *gold* and *auto* is possible, because none of the languages provided both annotation types – there is either a lack of *auto* (Catalan, English, Italian, Spanish) or *gold* (Dutch) or even both (German) annotations.

Altogether, we can conclude that the mdNES method leads to rather low performance with the exception of Italian. For this reason, we do not include mdNES as a stand-alone method within our further investigations. However, our results and observations showed that named entities can be valuable indicators of mentions (as in the case of Italian), which means that NEs should be reevaluated not in the form of a stand-alone mention detection method, but rather as supplementary information to other employed methods.

mdPOSP

The next approach for multilingual MD that we want to review, as listed in table 5.16, is the mdPOSP method. Even though mdPOSP outperforms mdNES in all comparable cases, it also performs worse, across all its filter variants (no filter, 5, 10 and 20), than mdDS and mdIOBA and mdVOTE when recall is assessed.

Across all variants of mdPOSP, we can see that altering the values for n directly affects the variation relation between precision and recall. Selecting a filter with a higher value for n, as 10 or 20, means that the patterns used as mentions often occur in the training set, thus those filters result in less, but more precise mention boundaries. However, if the mdPOSP method is employed for applications other than coreference resolution, being able to select a filter that favors either precision or recall can be a great advantage during system development.

The highest recall for mdPOSP is achieved when no filter is applied to the method with a considerable increase over filter 5 (with minimum difference for German – 7.49 percent points and maximum for Italian – 18.06 percent points). However, as we will discuss further in our qualitative analysis, we cannot favor recall exclusively for the mdPOSP method without using a frequency filter. This is so, since every mention pattern that is seen in the training set (including erroneous phrases) is consistently replicated in the test set leading to ***overgeneration*** (producing well formed mentions as well as numerous incomplete phrases that are also marked as mentions) to a great extent and thus the resulting phrases either do not correspond to correct mention boundaries or are so many that the coreference pipeline is overloaded with potential mentions

overgeneration

to choose from. Yet, too many mentions can be as bad as having too few of them. In order to avoid such erroneously produced phrases, only for mention detection as well as in the coreference pipeline, we compare mdPOSP with filters 5, 10 and 20 as potential candidates. For this analysis we also compare the F-scores of the filters, as they give a good representation of both precision and recall together. According to the figures in table 5.16, mdPOSP 5 outperforms filters 10 and 20 with respect to recall and F-measure and in both *gold* and *auto* settings across all languages. The only exceptions are the F-scores for Spanish and German. The former indicates a higher result for filter 10 with respect to filter 5 with 0.99 percent points. The F-scores of both filters for German are equal.

Within the *auto* setting the difference in recall between the four filters ranges across the languages as shown in table 5.16. For that setting, filter 5 outperforms filter 10 with 1.44 percent points (lowest difference) for German and 7.09 percent points (highest difference) for Italian, while the results for the recall achieved by filter 20 are even lower. This tendency is kept for the *gold* setting as well (filter 5 reaches higher scores for recall with respect to filter 10 and 20) with the lowest difference for Catalan, being 1.47 percent points and highest difference for Spanish, being 5.52 percent points (see table 5.16). Moreover, filter 5 consistently achieves highest recall for all targeted languages and across all evaluation settings and avoids the errors when no filter is applied. This designates filter 5 as best performing in comparison to filter 10, 20 and no filter at all and thus we employ only this filter in all our further analysis.

With respect to F-scores, as we noted above, filter 10 outperforms filter 5 for Spanish and has equal score for German. The highest variation in that setting between those two filters can be observed for Italian – 4.93 percent points. The tendencies for the *gold* setting are more consistent, because filter 5 outperforms filter 10 for each of the languages for which results can be compared. The lowest difference can be seen for Catalan with an improvement of filter 5 of 0.33 percent points and highest improvement is reached for English with 2.91 percent points.

mdDS

The results in table 5.16 rank mdDS as the best performing method across all the different approaches when recall is observed. The scores in the *auto* setting indicate that there is a big variation between the achieved recall from all four evaluated methods ranging with differences of more than 60 percent points (being 19.35% for Italian (the performance of mdPOSP 20) and 82.34% for English (the performance of mdDS)) when recall is favored for evaluation. We can also note that there is a clear cut between the best performing method, mdDS, and the performance of all other approaches. When recall is assessed, using dependency information shows best results for both *gold* and *auto* settings

(see table 5.16), with the exception of Italian and Dutch in the *gold* setting for which *gold* dependency annotations were not provided. Moreover, mdDS reaches highest recall across all languages also in its basic implementation (when only the longest spans per mention head are selected (see table 5.17)). Altogether, mdDS does not profit as much from the extraction of various spans per mention head (the increase in recall across all languages is 1.69 percent points) as mdIOBA loses when only the longest span is selected (4.29 percent points across all six languages).

With regard to multilingual performance, we can note that mdDS performs best, but not optimally balanced over all languages. When all spans are used (table 5.16) mdDS reaches highest recall for English (82.34%) and lowest for Italian (62.64%). The latter results in a variation of 19.70%. This means that in case a new random language is included in the multilingual coreference resolution pipeline, one cannot predict mdDS's performance with a high confidence. However, this difference is not only across six languages, but also across two language families. Within the Romance language family the variation is higher (14.18 percent points between Italian (62.64%) and Spanish (76.82%)), while within the Germanic language family the difference is considerably lower (8.92 percent points between Dutch (73.42%) and English (82.34%)). The latter results show that mdDS reaches higher scores for Germanic languages with less variation across the scores. This may be either caused by the similarities across the languages of one family or the fact that in our case the annotations for those languages were more consistent. If the former is the case, the introduction of a new language from a language family integrated in the system will have higher chances of performing close to the languages within that family than a language within a different one. In order to be sure in our assumption that the reason lies within the similarity of languages, additional families should be observed.

Furthermore, what the differences tell us is that even if dependency information is provided via either automatically achieved or manually labeled annotation layers, we cannot expect consistently robust performance over all languages if there is not much overlap between the *gold* mention layer and the phrases underlined by the dependency structure – good examples for that again are the best and lowest performing languages: English (for which gold mentions overlap considerably with the underlying noun phrases) and Italian (for which less correspondence between noun phrases and gold mentions is present). The latter peculiarities with respect to the correspondence of mentions to the NP structure were described in more detail in section 3.2.2.

mdDS is also one of the methods for which the contrast between *gold* and *auto* labels does not make a big difference in system performance overall. The figures show that there is a variation between both settings of only 0.88 percent points when the four languages for which there are *gold* labels are targeted (Catalan, English, German, Spanish), while the variation for mdIOBA,

for example, is relatively higher – 3.75 percent points across the same four languages. This is counterintuitive since mdDS is expected to be more sensitive to the quality of the provided annotations. One reason, for example, is the fact that as a rule-based method mdDS completely relies on the annotations provided in the data without any capabilities of abstracting over the observed information. This rule-based nature of the approach prevents it from accounting for exceptions and errors in the automatically labeled data.

The latter fact again raises one of our main concerns within our work – how suitable is the method at hand within a multilingual context? Considering the quantitative data, mdDS is reliable and well performing for mention detection in a multilingual coreference resolution system. It reaches best scores when recall is favored, thus this method could be used whenever dependency annotations are provided for all targeted languages. However, mdDS is fully dependent on the presence of this type of annotation layer and cannot be used if dependencies are not included in the data.

mdIOBA

mdIOBA is the next mention detection method that we proposed and want to discuss with respect to the results presented in table 5.16 and table 5.17. In both settings *gold* and *auto* and when all spans are included in the set of mentions (see table 5.16), mdIOBA performs worse than mdDS when recall is assessed. On average across all languages, mdIOBA also performs worse than mdVOTE – within the *auto* setting and when all spans are used (table 5.16), mdIOBA reaches 51.93% as an average recall score across all six languages, while mdVOTE performs slightly better – 52.48%; within the *gold* setting (when only Catalan, English, German and Spanish are compared), mdIOBA leads with 60.77% ahead of mdVOTE with 57.10%. This makes the comparison between the two methods not that straightforward. Yet, within the *auto* setting, mdIOBA reaches lower recall scores only for Catalan and Dutch with respect to mdVOTE, where for Dutch the difference is much higher (being 12.13 percent points). This gap accounts for the lower recall overall achieved by mdIOBA with respect to mdDS.

mdIOBA also reaches higher recall when all potential mention spans are included (table 5.16, *auto* setting) in comparison to the longest span per mention head (table 5.17) with an average of 4.29 percent points over all languages.

Since mdIOBA is a machine learning approach it is not as dependent on the provided annotations as the best performing rule-based approach – mdDS. Having a machine learning nature, mdIOBA is easily adaptable to the provided multilingual data annotations. It can also be easily optimized further and tuned to a language specific approach, if needed, by selecting language specific features and parameters.

However, as noted in the analysis for mdDS, mdIOBA is more sensitive to the quality of the provided annotations (*gold* and *auto*) with respect to mdDS, leading to a difference of 3.75 percent points between the two settings and across all four languages that provide that type of annotations as opposed to 0.88 percent points for mdDS.

The results that we report in table 5.16 are based on mdIOBA's performance when both part-of-speech and dependency annotations are made use of. However, it is important to know to what extent mdIOBA is applicable to different types of annotations or in other words to what extent is it sensitive to the presence or absence of the highly informative dependency labels. In this way, we will gain a better awareness of the flexibility of this method and thus of its adaptability to multilingual approaches. Additionally, we carry out an investigation of detection of base NPs, because this is more intuitive and close to the initial IOB tagging approach, which drastically reduces the number of labels/classes the classifier has to choose from. Thus, having only 3 equivalence classes (being I(inside), O(utside) and the B(eginning) of a potential mention) decreases the difficulty for the machine learning resolver. Therefore, we can expect that a better overall performance will be reached. For these reasons, we designed an experiment that compares three different variations of mdIOBA to the best performing mdPOSP approach – using filter 5. We select this comparison, because mdPOSP is a baseline approach that can be easily applied in a multilingual setting and for which only part-of-speech information is needed. The mdIOBA variations are as follow:

- **mdIOBA** - the original mdIOBA method using dependency annotations and POS information.
- **mdIOBA-POS** - uses POS information only.
- **mdIOBA-BASE** - aims at the identification of base noun phrases – flat, not embedded structures.

Table 5.18 lists the system performance across all languages for all three mdIOBA variations together with the baseline mdPOSP approach. The figures show that using dependency information boosts the performance of the mdIOBA method. The difference for recall between mdIOBA (the variant including dependency information) and MDIOBA-POS (the variant based solely on POS labels) does not increase by 3% for any of the considered languages. This shows that dependency information can help improve the performance of machine learning methods, but it also indicates that the presence of dependency information is far less important to mdIOBA than to mdDS which completely relies on it. This, we note, is a highly positive feature of mdIOBA.

Another comparison that we can achieve by the results in table 5.18 is the contrast between the baseline pattern matching of part-of-speech tags and the memory-based learning approach that MDIOBA-POS offers. Both methods use

	mdPOSP 5			mdIOBA-POS			mdIOBA			mdIOBA-BASE		
	R	P	F	R	P	F	R	P	F	R	P	F
CA	40.38	49.90	44.67	49.15	51.33	50.22	50.65	53.01	51.80	31.60	70.60	43.66
DU	36.33	30.85	33.37	39.53	48.96	43.74	42.12	51.74	46.44	43.53	58.32	49.85
EN	37.89	60.36	46.55	56.56	67.87	61.70	59.59	66.50	62.58	43.90	79.63	56.60
GE	42.44	62.21	50.45	62.73	72.24	67.14	67.32	70.23	68.74	54.51	77.27	63.23
IT	34.32	58.02	43.13	37.53	45.96	41.81	41.40	46.42	43.77	44.20	60.43	52.32
SP	41.54	50.39	45.54	48.96	52.26	50.56	50.51	53.61	52.01	30.35	71.46	42.61

Table 5.18: The performance of the three variants of mdIOBA across all six languages of the SemEval-2 shared task listed as: *mdIOBA-POS*, *mdIOBA* and *mdIOBA-BASE*. The baseline *mdPOSP 5* is included for comparison.

solely the POS annotation layer and thus their results can be directly compared. The figures show that mdIOBA-POS consistently outperforms mdPOSP for all languages, reaching a difference of more than 20% (for German) when recall is assessed. One of the reasons for this big variation is the fact that mdIOBA-POS is able to abstract away from the data and better account for exceptions and unseen examples, while mdPOSP can only detect a mention if its pattern was already present in the training set. The performance achieved by the mdIOBA-POS approach is a highly positive outcome for our work. It indicates that a machine learning method for mention detection can be designed in a competitive way on a language independent level, since this method relies on the most widely distributed annotation layer – POS.

The last variation that we investigated within the mdIOBA analysis is the mdIOBA-BASE alternative. The figures in table 5.18 indicate that mdIOBA-BASE is outperformed by both mdIOBA and mdIOBA-POS in terms of recall for all languages, apart from Dutch and Italian. Altogether, the drop in system performance is not consistent across the languages which can be the consequence of the variation in mention embedding within the set of languages and the type of annotation scheme used for mention labeling. What is more important to us, however, is the fact that identifying base NPs is not sufficient within the task of coreference resolution. This is indicated by the significant drop of recall within the mdIOBA-BASE results – even though precision figures are increased substantially for mdIOBA-BASE in comparison to both mdIOBA and mdIOBA-POS, we cannot consider mdIOBA-BASE as a competitive method.

mdVOTE

The last approach to multilingual mention detection that we review is the combination of mdPOSP, mdDS and mdIOBA, namely mdVOTE. As the figures in the last column of table 5.16 reveal, a combination of such different methods

		mdDS	mdIOBA	mdDS ⋃ mdIOBA
CA	R	75.12	50.65	**81.84**
	P	45.96	53.01	40.26
	F_1	57.03	51.80	53.97
DU	R	73.42	42.12	**79.93**
	P	27.43	51.74	26.67
	F_1	39.94	46.44	40.00
EN	R	82.34	59.59	**85.92**
	P	44.50	66.50	40.97
	F_1	57.77	62.85	55.49
GE	R	78.96	67.32	**86.94**
	P	59.16	70.23	53.27
	F_1	67.64	68.74	66.06
IT	R	62.64	41.40	**71.88**
	P	42.35	46.42	36.58
	F_1	50.54	43.77	48.48
SP	R	76.82	50.51	**82.97**
	P	45.59	53.61	40.07
	F_1	57.22	52.01	54.05

Table 5.19: Mention detection with combined rule-based *mdDS* and machine learning performance *mdIOBA* via the unification of both approaches as *mdDS* ⋃ *mdIOBA* across the six languages of the SemEval-2 shared task. Highest recall figures are highlighted in bold.

improves the recall achieved by the system only for Catalan and Dutch. Yet, when only longest spans are considered (table 5.17), mdIOBA's performance reaches such a drop that all languages apart from Spanish show higher figures for recall for mdVOTE. Moreover, in comparison to mdDS, mdVOTE reaches scores for recall that are up to about 30 percent points lower (Spanish) as well as a variation across the languages of the same extent (German and Italian). The problem, as we will better exemplify in the qualitative analysis of mdVOTE (see section 5.2.2.2), is that often enough the two weaker approaches (mdIOBA and mdPOSP) overrule the best performing method – mdDS. Another disadvantage of mdVOTE is the fact that to achieve a decision, that approach needs output of at least two reliable mention detection methods which is not easily accomplished in most coreference resolution tasks.

As we mentioned in section 5.1.6, mdVOTE can be modified as a weighted method, which can be used to solve ties when even number of votees is used or when a preference can be given to one of the voting approaches. Another possibility to combine the strengths of mdDS and mdIOBA is simply to unify their

respective sets of identified mentions. This means that we combine all mentions from mdDS and mdIOBA and exclude all redundant instances (this means that mentions identified by both methods are included only once). The results achieved by the union are shown in column *mdDS* ⋃ *mdIOBA* in table 5.19. The recall figures across all languages are consistently higher than both methods, mdDS and mdIOBA, separately. The overall performance for recall is with lowest scores of 71.88% for Italian and highest scores of 86.94% for German with an average score over all languages of 81.58%. These results indicate that when a method based on the syntactic structure does not reach optimal performance, it can be easily enhanced by a machine learning approach such as mdIOBA. As the scores in column *mdDS* ⋃ *mdIOBA* in table 5.19 show, the unification reaches an improvement of performance from 3.58 percent points for English to 9.24 percent points for Italian. That is an improvement with an average, across all languages, of 6.70 percent points leading to best recall scores across all tested approaches and for all targeted languages. For this reason, we consider the unification of the outcomes of a syntactic and machine learning approach as a good and quantitatively well performing strategy within a multilingual setting.

5.2.2.2 *Qualitative Analysis*

After we have assessed the numerical results within our quantitative analysis of the intrinsic multilingual mention detection in section 5.2.2.1, it is time to look at the actual output of the various methods and gain a more detailed understanding of the type of the detected mentions and the errors made by the approaches. For this, we looked at the first 100 sentences and assessed the errors that occur within them. The new observations will give us a deeper insight on the qualitative compatibility of the developed algorithms according to the employed datasets and annotation schemes.

mdNES

Within the *auto* setting, the mdNES method could only be employed for the Dutch language, because there is no annotation layer for automatically achieved named entities for the rest of the languages. In order to exemplify some of the problems of the mdNES method, let us look at the sentence in table 5.20. In column *NE*, we include the labels for the named entities provided in the dataset, column *mentionID* lists the *gold* labels for the coreferent mentions and column *mdNES* reveals the mention boundaries that the mdNES method identifies. What can be seen from the example sentence is that named entities do not correspond to full noun phrases, which carries multiple consequences for the mention identification process. First, the mentions that the mdNES identifies that do not have identical boundaries with the *gold* mention boundaries, such as mentions 1 and 2 in column *mdNES*, are completely ignored by the scoring software and thus drastically reduce the performance of this method. Second, with respect to

#	token	POS	NE	mentionID	mdNES
1	In	VZ	_	_	_
2	de	LID	_	(1	_
3	driemaandelijkse	ADJ	_	_	_
4	peiling	N	_	_	_
5	van	VZ	_	_	_
6	Marketing	N	PER	_	(1)
7	Unit	SPEC	_	1)	_
8	behaalt	WW	_	_	_
9	het	LID	_	(3	_
10	Vlaams	SPEC	ORG	_	(2
11	Blok	SPEC	ORG	3)	2)
12	zijn	VNW	_	(4	_
13	hoogste	ADJ	_	_	_
14	score	N	_	_	_
15	ooit	BW	_	4)	_
16	(	LET	_	_	_
17	16,8	TW	_	(4	_
18	procent	N	_	4)	_
19	,	LET	_	_	_
20	+	LET	_	_	_
21	0,3	N	_	_	_
22	)	LET	_	_	_
23	.	LET	_	_	_

Table 5.20: An example sentence with the annotations provided by the mdNES module (column *mdNES*) from the SemEval-2 Dutch dataset. Column *NE* lists the named entity annotation layer and column *mentionID* includes the set of gold mentions.

mention 1 (*de driemaandelijkse peiling van Marketing Unit* (English: *the quarterly poll by Marketing Unit*)) in column *mentionID*, we can see that the named entities in the Dutch data do not always correspond to the same entities from the *gold* mentions. Therefore, the syntactic heads of the mentions identified by the system in the process directly following mention detection, namely mention head detection (see chapter 6) are not always the correct mention heads. In this case, the head of the *gold* mention 1 is *Unit*. It is also the token that should have been selected by the system as well. Yet, the mdNES method elicits only *Marketing* as a mention and, therefore, the system can select only this token

#	token	POS	NE	mentionID	mdNES
1	Fa	VI	(gsp	(3	(1
2	parte	SS	gsp)	3)	1)
3	del	ES	(organizzazione	(133	(2
4	Comprensorio	SPN	_	_	_
5	Alto	SPN	(gsp	(4	(3
6	Garda	SPN	gsp)	4)	3)
7	e	C_coo	_	_	_
8	Ledro	SPN	(gsp)|organizzazione)	(215)|133)	(4)|2)
9	.	XPS	_	_	_

Table 5.21: An example sentence with the annotations provided by the mdNES module (column *mdNES*) from the SemEval-2 Italian dataset. Column *NE* lists the named entity annotation layer and column *mentionID* includes the set of gold mentions.

as the head of the mention. This is crucial for the mention pair approach that we employ, because phrases can only be rendered coreferent on the basis of their syntactic heads plus additional information about the phrase. A failure to identify the head of a phrase can lead to information about the mention that is highly misleading for the classifier. Additionally, the named entity annotation for Dutch covers only few different entity types[7], and as can be seen from the example sentence in table 5.20, it does not cover simple entities as ordinal and cardinal numbers, that are generally easy to extract and can be helpful to mention detection.

Within the *gold* setting only Catalan, English, Italian and Spanish provided NE annotations. Similar to the problems that we discussed for Dutch, the NE annotation does not provide optimal overlap between entities and *gold* mentions with the exception of Italian. In table 5.21, we present a sentence from the SemEval-2 Italian dataset from which we can see that the mentions identified from the mdNES completely overlap with the boundaries of the *gold* mentions in column *mentionID*. Moreover, the large number of entity types, covered by the Italian NE annotation scheme (12 entity types), as well as the fact that pronouns are also marked as entities, accounts for almost all mentions in the dataset.

The examples from Dutch and Italian show that there is an extreme variation between the usability of named entities as indicators for mentions within a multilingual mention detection task for coreference resolution. Our analysis shows, that NEs can only be used as support for more robust methods for MD, unless exceptional cases (e.g. Italian) are targeted. Knowledge about the

[7] In fact Dutch covers only four different entity types: PER, MISC, LOC and ORG.

semantic type of the phrases can also be helpful during the resolution process, since depending on the language and annotation scheme in use some entity types are seldom coreferent (for example, cardinals or ordinals).

mdDS

The method based on dependency structures achieves not only highest quantitative but as well highest qualitative performance over all approached languages. This is due to the fact that mdDS extracts mentions that correspond to well-formed phrases overlapping reliably with the boundaries of the *gold* mentions in the data. Again, both Dutch and Italian are outliers in performance with respect to the other four languages. In order to better visualize the problems that can occur within that method, let us consider the sentence presented in table 5.22. The table lists the *Head* (the head of the phrase) and *DepRel* (dependency relation) columns from the annotation, column *mentionID* includes all *gold* mentions (including singletons) and in the last *mdDS* column we include the output from the mdDS method.

One difference that we can note between the mentions in column *mentionID* and *mdDS* is the mismatch of the use of postmodification. *Gold* mentions do include all syntactic dependents of the head up to the first postmodifier. This, however, is a language specific difference. Since the main aim of our work is to investigate the needed steps for the development of a multilingual coreference resolution pipeline, we do not account for such language specific variations. Thus, the mdDS variant for Dutch extracts the full, according to the dependency structure, noun phrases. This is fiercely penalized by the scoring software as complete unrecognition of the given phrase (recall that both SemEval-2 and CoNLL-2012 scorers require that mentions are identified with the exact boundaries of the mentions provided in the key set, otherwise the mentions detected by the system are discarded). Such language specific details can be accounted for, if improvements on given languages are needed. Yet, our main aim is to achieve a multilingual and even language independent behaviour and thus tuning the approach at this point defeats the purpose of our work.

Various other types of inconsistencies in the annotations are also a considerable part of the errors of the mention detection module for Dutch. They are either with respect to the dependency labels or the key boundaries, as for example mention #4 in column *mentionID* in table 5.23. In that case, the *gold* mention #4 includes token #20 in its boundaries. Yet, according to the information provided in the *Head* column, that token is dependent on #15 and not on any of the tokens included in mention #4 (#17, #18 and #19). Thus, mdDS does not include this token in the boundaries of mention #2 in column *mdDS*.

Such inconsistencies are not easy to predict and analyze for a single language and in a context of a multilingual coreference resolution system, such as UBIU, they are even harder to cope with. Our analysis also confirms that the

#	token	POS	Head	DepRel	mentionID	mdDS
1	Lezer	N	6	su	(1	(1
2	Luc	SPEC	1	app	_	_
3	Vanacker	SPEC	2	mwp	1)	_
4	uit	VZ	1	mod	_	_
5	Koksijde	N	4	obj1	(2)	(2)|1)
6	kon	WW	0	ROOT	_	_
7	bij	VZ	23	mod	_	_
8	wapenhandel	N	7	obj1	(3	(3
9	Doucet	N	8	app	3)	_
10	in	VZ	8	mod	_	_
11	Koksijde	N	10	obj1	(4)	(4)
12	,	LET	11	punct	_	_
13	die	VNW	8	mod	_	_
14	nog	BW	15	mod	_	_
15	altijd	BW	16	mod	_	_
16	bestaat	WW	13	body	_	3)
17	,	LET	16	punct	_	_
18	geen	VNW	19	det	_	(5
19	bevestiging	N	23	obj1	_	_
20	van	VZ	19	mod	_	_
21	dit	VNW	22	det	(5	(6
22	verhaal	N	20	obj1	5)	6)|5)
23	krijgen	WW	6	vc	_	_
24	.	LET	23	punct	_	_

Table 5.22: An example sentence with the annotations provided by the mdDS module (column *mdDS*) from the SemEval-2 Dutch dataset. Column *Head* lists the ID of the syntactic head for the token, column *DepRel* shows the dependency relation label of the word and column *mentionID* includes the set of gold mentions.

quality of linguistic annotations provided in the various datasets may differ immensely. Accordingly, rule-based approaches, such as mdDS, that do not abstract over the data and fully rely on its validity are correspondingly affected. This drastically reduces the flexibility of the method to new languages and their associated datasets, which is in conflict with the the main purview of the current investigation.

#	token	POS	Head	DepRel	mentionID	mdDS
12	de	LID	13	det	(3	(1
13	Universiteit	SPEC	11	obj1	_	_
14	Antwerpen	SPEC	13	mwp	3)	1)
15	analyseerden	WW	0	ROOT	_	_
16	in	VZ	15	mod	_	_
17	het	LID	18	det	(4	(2
18	blad	N	16	obj1	_	_
19	Huisarts	SPEC	18	app	_	2)
20	Nu	SPEC	15	mod	4)	_
21	hoe	BW	15	vc	_	_
22	artsen	N	23	su	_	_
23	beslissen	WW	21	body	_	_

Table 5.23: An example excerpt with the annotations provided by the mdDS module (column *mdDS*) from the SemEval-2 Dutch dataset. Column *Head* lists the ID of the syntactic head for the token, column *DepRel* shows the dependency relation label of the word and column *mentionID* includes the set of gold mentions.

mdPOSP

In table 5.24 we present an example sentence from the SemEval-2 English dataset with the help of which we want to exemplify some of the problems of the mdPOSP method. In column *mdPOSP* we list the output of the method when no filter is applied to it. One thing that can be seen from the resulting mentions is that using each and every sequence of patterns present in the training set does not always prove to be a good approach. Some mentions, as for example 1, 2 or 3, in column *mdPOSP* have wrong boundaries, because they represent patterns of exceptional cases in the training set. Yet, once such a pattern is recorded it is always applied in the test set. This can be exceptionally misleading, because one error in the training set can lead to a replication of that type of error over the full test set. Altogether, the mere overgeneration of mentions that harms the overall performance leads to the use of filter 5 in mdPOSP. As can be seen from column *mdPOSP 5*, the number of mentions is drastically reduced and only mentions that occur in the training set with higher frequency can be applied. This, however, as shown in the examples, can lead to overall shorter mentions, because longer patterns have less chance of appearing frequently than shorter ones.

In fact, as the mdPOSP output shows, the mentions or phrases, identified by the module, do not always correspond to grammatically correct noun phrases

#	token	POS	mentionID	mdPOSP 5	mdPOSP	mdIOBA-POS	mdIOBA
1	A	DT	(55	(1	(1	(1	(1
2	committee	NN	_	1)	(2	_	_
3	representing	VBG	_	_	2)|1)	_	_
4	the	DT	(56	_	(3|(4	(2	_
5	unsecured	JJ	_	(2	_	_	_
6	creditors	NNS	56)|55)	2)	4)	2)|1)	1)
7	agreed	VBD	_	_	3)	_	_
8	to	TO	_	_	_	_	_
9	accept	VB	_	_	_	_	_
10	24	CD	(57	(3	(5|(6	(3	(2
11	cents	NNS	_	3)	_	_	2)
12	on	IN	_	_	6)	3)	_
13	the	DT	(58	(4	_	(4	(3
14	dollar	NN	58)|57)	4)	5)	4)	3)
15	,	,	_	_	_	_	_
16	Eagle	NNP	(59)	_	(8)	(5)	(4)
17	said	VBD	_	_	_	_	_
18	.	.	_	_	_	_	_

Table 5.24: An example sentence with the annotations provided by the mdPOSP module (columns *mdPOSP 5* and *mdPOSP*) and the mdIOBA module (columns *mdIOBA-POS* and column *mdIOBA*) from the SemEval-2 English dataset. Column *mentionID* includes the set of gold mentions.

which is less often seen within the rest of the approaches. There are numerous modifications and improvements that can be initiated in order to overcome errors of various types. However, we proposed mdPOSP as a baseline approach. A rule-based attempt to overcome such modifications does not offer a good trade off for this baseline. Moreover, approaching such a modification in a multilingual task will need a considerable effort for development and will significantly reduce the flexibility of the baseline. The latter can be regarded as an additional reason for the introduction of filters based on the frequency of patterns, such as the ones we employ.

When we compare both *mdPOSP* columns with the output of the machine learning approach based on the same annotation layer listed in column *mdIOBA-POS*, we can see that MBL can already easily account for exceptional cases and lead to a more reasonable representation without the need to assemble sets of rules for each separate language.

mdIOBA

As we just peeked into *mdIOBA-POS*'s qualitative performance, let us see how mdIOBA's output can change when dependency information is added to the knowledge the method can use for mention detection. From the quantitative analysis and the results in table 5.18 on page 120 we saw that *mdIOBA-POS* and mdIOBA proper have a very close overall performance, which is reflected in the difference in the mentions presented in columns *mdIOBA-POS* and *mdIOBA* in table 5.24. One of the positive effects of the additional information is the corrected identification of the boundaries of the mention *"24 cents"*, which were wrongly identified by *mdIOBA-POS*.

However, using more information during classification also means being more restrictive, which can lead to other errors as excluding mention *"the unsecured creditors"* for example. What the data shows us is that using dependency information, additional to part-of-speech tags, prompts for selecting flatter phrase structures, which in general decreases recall. Yet, correcting erroneous mention boundaries accounts for that loss as the data in table 5.24 shows.

The closer look into the output of both mdIOBA and *mdIOBA-POS* indicates that merely looking at the evaluation numbers does not necessarily give us the best performing method. Completely ignoring mentions with wrongly identified boundaries is an often seen penalty by the scoring procedure for this method, thus if the task was to find approximations of mentions rather than detecting the exact boundaries of every mention, this method will be a lot more successful. If a way to provide a better account for such entities is found, *mdIOBA-POS* can prove more helpful in the attempt to detect mentions and can certainly be easy to apply within a multilingual approach in which no dependency information is provided. Nevertheless, as our quantitative evaluation showed, mdIOBA-POS reaches a highly competitive performance. Its increased flexibility to new datasets, resulting from the low dependability on annotation layers other than POS, elicits mdIOBA-POS as an exceedingly good candidate for language independent approaches.

mdVOTE

The similarities in mention boundaries between mdIOBA and mdPOSP (listed in table 5.24) indicate that mdIOBA is strongly influenced by the part-of-speech information that it uses. Thus, combining such a POS-biased approach with mdPOSP (being a heuristic method solely driven by POS information) and mdDS in a single annotation attempt leads to a highly biased voting scheme. Voting on per-token basis proves difficult in such a setting, because the two weaker approaches, with higher similarity of their output, often overrule the stronger mdDS approach. Additionally, as the example in table 5.25 shows, selecting a filter for mdPOSP (needed for more accurate performance) as well as using

#	token	POS	mentionID	mdDS	mdIOBA	mdPOSP5	mdVOTE
18	the	DT	(19	(1	(1	(1	(1
19	Yemeni	NNP	_	(2)	_	1)	1)
20	port	NN	_	_	1)	_	_
21	of	IN	_	_	_	_	_
22	Aden	NNP	(75) | 19)	(3)|1)	(2)	(2)	(2)

Table 5.25: An example part with the annotations provided by the mdVOTE method (column *mdVOTE*) from the SemEval-2 English dataset. Columns *mdDS*, *mdIOBA* and *mdPOSP 5* show the output for its votees. Column *mentionID* includes the set of gold mentions.

dependency information for mdIOBA leads to more flattened phrase structures, resulting in the detection of less mentions altogether. The output from the voting scheme inherits this property which is also confirmed by the low recall figures in table 5.16. This is not a property that we aim for in MD for multilingual coreference resolution. For this reason, either using a different combination of methods as votees, or completely discarding this approach is an acceptable option.

The best performing variant of combining different approaches that we showed in the mdVOTE section in our quantitative analysis in section 5.2.2.1, namely the unification of the outputs from mdDS and mdIOBA, directly inherits the behaviour of both methods. As table 5.26 shows, only the mentions with identical spans are merged and thus no further qualitative differences and problems can be seen apart from those discussed in the sections for each respective method. Similar to the quantitative analysis of the unification approach, here we can see that considering both outputs leads to an increase in the recall

#	token	POS	mentionID	mdDS	mdIOBA	mdDS⋃mdIOBA
18	the	DT	(19	(1	(1	(1|(2
19	Yemeni	NNP	_	(2)	_	(3)
20	port	NN	_	_	1)	2)
21	of	IN	_	_	_	_
22	Aden	NNP	(75) | 19)	(3)|1)	(2)	(4)|1)

Table 5.26: An example part with the annotations provided by the mdDS⋃mdIOBA method (column *mdDS⋃mdIOBA*) from the SemEval-2 English dataset. Columns *mdDS* and *mdIOBA* show the output from the respective methods. Column *mentionID* includes the set of gold mentions.

of mentions with no decrease in their quality, because there is no change in the already existing mention boundaries identified by the separate mdDS and mdIOBA methods. The latter fact shows that the unification of a machine learning and a rule-based approach, such as the one we investigate (mdDS$\bigcup$mdIOBA), proves to be highly desirable in both quantitative and qualitative performance. We assume that rule-based methods, such as mdDS, that are based on syntactic information, can always profit from a ML approach whenever they do not achieve optimal recall figures.

5.2.2.3 *Discussion*

We presented both quantitative and qualitative analysis of the methods proposed in our work (introduced in section 5.1) that can be employed on the SemEval-2 dataset. Our results indicate that the selection of a method for detecting mentions within multilingual CR is dependent on multiple crucial factors.

First, and most importantly, we showed that the presence of syntactic annotations, such as dependency parses, is highly beneficial to mention detection. We affirmed that the method based on this type of annotations performs reliably and efficiently through all targeted languages. The results mdDS achieved showed consistent recall performance throughout all settings in which it could be evaluated.

We also showed that named entities are, indeed, phrases that have a very high probability of being a mention (depending on the selected annotation scheme), but the performance of mdNES also indicated that this layer of annotations cannot be reasonably used for a stand-alone mention detection procedure across a diverse set of languages and thus a diverse set of annotation schemes. However, named entities, if provided, can be used as a complimentary support to the MD method in use. For this reason, we exclude mdNES from our further investigations.

One further remark, based on our observations, is the fact that a memory-based learning approach, in the form of mdIOBA, can be easily and competitively used when syntactic annotation layers are not provided. We showed that mdIOBA can be employed by only using part-of-speech information and that additional information, such as dependency structure, can have a beneficial effect on the learner's performance. Achieving a comparatively good behavior *mdIOBA-POS* clearly outperforms the baseline method mdPOSP indicating that memory-based learning can successfully and reliably be used for mention detection across all languages for which part-of-speech information is provided. This is a substantial advantage in a multilingual setting that can always be used to enhance existing methods based on the syntactic structure.

As a matter of fact, the unification of mdDS and mdIOBA, namely mdDS$\bigcup$mdIOBA, led to highest system performance within the *auto* setting across all languages. Thus, we expect that mdIOBA, as a representative of a machine

learning approach, will be a good method for combination and enhancement of any rule-based approach, as mdDS in our case or a competitive substitute for it in the rare case that no dependency information is provided.

Considering the language specific tuning employed in our methods, as for example the language dependent sets of dependency labels (mdDS), motivates the search for similarities between languages that can be generalized according to various criteria. One such generalization can be approached around the concept of language families, because various linguistic phenomena are similarly represented across the annotation schemes of languages within one language family (e.g. Catalan and Spanish). Moreover, providing language specific information in a form that can serve the languages of a whole language family provides system flexibility and robustness for new languages from an already known family.

5.2.3 SemEval-2 *Extrinsic Evaluation*

In this section we provide a different evaluation for the mention detection methods employed on the SemEval-2 datasets, namely, by including them in the coreference resolution pipeline. This attempt will show if the best performing method, mdDS, as well as the combination of mdDS and mdIOBA, keep their good and robust performance when integrated in the full system with respect to the performance for mdPOSP-5 and mdIOBA. The optimal outcome of this approach would be that the system performance overall is highest when mdDS$\bigcup$mdIOBA is used across all languages. In section 5.2.3.1 we present all results and in section 5.2.3.2 a following discussion of the evaluation is offered.

5.2.3.1 *Results*

Table 5.27 and table 5.28 list the performance of the CR system across all six languages within the full coreference pipeline. As in table 5.16 on page 113, the results are listed in recall (R), precision (P) and F-measure. Using so many and divergent in nature evaluation metrics renders the attempt to calculate an average system score not very informative. For this reason we show the results of the metrics separately and report calculated total scores for easier comparison. Additionally, it is important to note that all results in table 5.27 and table 5.28 are achieved by the use of *auto* linguistic annotations. The last setting, *gold mentions*, uses *auto* linguistic annotations, but instead of using the mentions detected by the MD module includes the set of *gold* mentions. That setting will allow a better comparison of the performance of the MD module within the CR pipeline, because this is the optimal performance it can achieve, or in other words, it is the ***upper bound***. There are two reasons for the use of only *auto* linguistic annotations: First, we showed that the tendencies among both, *gold* and *auto*, settings are kept for all targeted languages with the only difference that results on *gold* annotations are higher than results on *auto* data,

upper bound

which is an expected outcome within most NLP tasks. Second, as we showed in section 5.2.2, no *gold* annotations were provided for Dutch and Italian (apart from NEs for Italian).

We would also like to note that there is a highly important difference between the mention detection scores in table 5.27 and 5.28 (listed in column *MD* in the tables) and the scores that we presented in the intrinsic evaluation in section 5.2.2. Once a CR system uses the mentions identified by its MD module, it keeps only the ones that it finds coreference relations for. This means, that singletons are removed from the final system output, which affects the MD scores reported by the SemEval-2 scoring software.

The overall evaluation presented by the figures in table 5.27 and table 5.28 reveals several facts interesting with respect to our work and goals. First, the performance of the baseline mdPOSP within the CR pipeline is considerably lower than both mdDS and mdIOBA as well as their union. It reaches a cross-language difference of almost 20 percent points (the cross-language scores calculated as an average of the total scores per language are listed in table 5.29 on page 136). Within the analysis of mdPOSP, we noted that this method over-generates in terms of identification of a very high number of potential mentions (which was the reason for us to integrate different filters) with erroneous mention boundaries. This fact is well demonstrated by the low performance of the method within the full coreference pipeline. For the latter, not only the quantity but as well the quality of the mentions is important, since phrases that were rendered coreferent but have erroneous boundaries will be discarded by the scoring software. Second, unlike the more ample variation between the performance of the mdDS and mdIOBA methods from their employment outside of the coreference pipeline, the figures in table 5.29 indicate that there is only a small difference of exactly 1 percent point across the languages with respect to the extrinsic evaluation of these two methods. The biggest gap can be observed for German (mdDS 64.29% and mdIOBA 61.90%). The more detailed figures in table 5.28 for that language show that this difference is mainly due to the improved performance reported by the MUC metric for mdDS. Moreover, the union of the outcome of both methods, mdDS$\bigcup$mdIOBA, enhances the performance of the MD module further leading to highest total scores for almost all targeted languages (see table 5.29). The only outlier is English, for which the performance of mdDS alone is better that mdDS$\bigcup$mdIOBA. Again, those changes are mainly for the variations reported by the MUC metric, which shows higher scores for all languages apart from English for mdDS$\bigcup$mdIOBA. As exemplified in table 2.3 on page 38 in section 2.3.4.3, MUC is highly sensitive to overmerged entities. With respect to that, the scores in table 5.27 and table 5.28 indicate that the mentions identified by mdDS reach higher entity chaining in the overall coreference output than the mentions provided by the mdIOBA method. We assume that this is not only the result of the higher recall figures typical for mdDS. mdDS provides mentions that are syntactically more precise than the ones

		CA			DU			EN		
		R	P	F_1	R	P	F_1	R	P	F_1
mdPOSP-5	MD	40.38	49.99	44.67	36.33	30.85	33.37	37.89	60.36	46.55
	MUC	0.01	1.75	0.03	1.10	6.22	1.87	1.24	14.63	2.29
	B^3	35.18	61.82	44.85	16.93	32.33	22.23	37.94	74.45	50.27
	$CEAF_M$	38.52	46.15	41.99	22.98	19.51	21.11	38.71	58.41	46.56
	$CEAF_E$	53.45	39.39	45.36	40.25	13.09	19.76	48.07	53.61	50.69
	BLANC	49.99	48.65	45.32	50.01	51.52	35.18	50.11	57.36	48.35
	TOTAL F_1			35.51			20.03			39.63
mdDS	MD	75.12	45.96	57.03	73.39	27.45	39.95	82.35	44.51	57.78
	MUC	6.31	27.34	10.26	4.83	11.85	6.86	11.87	21.66	15.33
	B^3	63.41	93.13	75.45	47.65	80.23	59.79	74.72	87.90	80.77
	$CEAF_M$	61.85	61.85	61.85	35.96	35.96	35.96	69.26	69.26	69.26
	$CEAF_E$	83.89	57.44	68.19	48.77	29.18	36.51	82.25	70.69	76.03
	BLANC	50.94	59.29	51.24	50.10	50.64	49.51	51.83	56.75	52.73
	TOTAL F_1			53.40			37.73			**58.82**
mdIOBA	MD	50.65	53.01	51.80	42.12	51.74	46.44	59.59	66.50	62.85
	MUC	6.49	22.38	10.06	4.20	38.60	7.59	4.56	23.71	7.65
	B^3	64.53	91.25	75.60	37.15	95.37	53.48	72.37	95.63	82.39
	$CEAF_M$	61.54	61.54	61.54	36.94	36.94	36.94	71.25	71.25	71.25
	$CEAF_E$	81.41	57.90	67.67	67.58	26.57	38.14	89.74	68.18	77.49
	BLANC	50.82	56.78	51.05	50.36	63.77	49.21	50.59	58.29	50.79
	TOTAL F_1			53.18			37.07			57.91
mdDS$\bigcup$mdIOBA	MD	74.44	44.82	55.95	77.92	28.58	41.83	80.37	43.08	56.10
	MUC	13.52	29.38	18.52	16.35	23.85	19.40	11.87	20.21	14.95
	B^3	66.64	87.57	75.68	54.83	73.87	62.94	75.02	87.03	80.58
	$CEAF_M$	62.86	62.86	62.86	41.12	41.12	41.12	68.91	68.91	68.91
	$CEAF_E$	79.27	61.13	69.03	46.74	36.21	40.80	81.38	70.97	75.82
	BLANC	51.87	59.82	52.81	51.21	55.88	51.52	51.84	55.76	52.67
	TOTAL F_1			**55.78**			**43.16**			58.59
gold mentions	MD	100	100	100	100	100	100	100	100	100
	MUC	13.00	39.79	19.60	6.92	47.57	12.08	13.21	38.69	19.70
	B^3	63.37	91.47	74.87	36.14	94.34	52.26	72.80	93.35	81.81
	$CEAF_M$	63.41	63.41	63.41	37.55	37.55	37.55	72.29	72.29	72.29
	$CEAF_E$	85.33	59.82	70.33	70.37	26.86	38.88	89.42	70.60	78.90
	BLANC	51.96	63.71	52.95	50.59	65.05	49.59	52.16	65.56	53.52
	TOTAL F_1			56.23			38.07			61.24

Table 5.27: Results for the different mention extraction modules in UBIU within the SemEval-2 shared task; MD evaluates the extraction of mentions; the best F-scores per metric and language are marked in bold.

		GE			IT			SP		
		R	P	F_1	R	P	F_1	R	P	F_1
mdPOSP-5	MD	42.44	62.21	50.45	34.32	58.02	43.13	41.54	50.39	45.54
	MUC	0.94	26.05	1.80	0.67	13.68	1.27	1.94	23.05	3.58
	B^3	42.63	73.21	53.88	39.92	74.60	52.01	38.51	59.49	46.76
	$CEAF_M$	44.89	64.05	52.79	37.78	58.71	45.98	40.78	46.85	43.60
	$CEAF_E$	54.48	59.42	56.84	48.67	54.97	51.63	55.49	41.15	47.25
	BLANC	50.02	56.18	48.03	50.03	56.45	48.04	50.02	54.62	45.44
	TOTAL F_1			42.67			39.79			37.32
mdDS	MD	78.93	59.16	67.63	62.63	42.35	50.53	74.63	71.14	72.84
	MUC	18.81	34.31	24.29	3.12	17.74	5.31	11.07	29.76	16.14
	B^3	78.73	90.74	84.31	72.46	95.85	82.53	65.38	89.67	75.62
	$CEAF_M$	76.11	76.11	76.11	70.05	70.06	70.06	63.52	63.52	63.52
	$CEAF_E$	87.66	76.96	81.96	89.60	68.00	77.32	82.98	61.28	70.50
	BLANC	52.97	64.65	54.79	50.15	53.11	49.98	51.54	59.52	52.30
	TOTAL F_1			64.29			57.04			55.62
mdIOBA	MD	67.32	70.23	68.74	41.41	46.43	43.78	50.51	53.61	52.01
	MUC	10.61	30.48	15.75	1.60	17.46	2.90	7.03	24.14	10.89
	B^3	77.56	93.89	84.95	71.86	97.81	82.85	65.15	91.54	76.12
	$CEAF_M$	75.36	75.36	75.36	70.69	70.69	70.69	62.70	62.70	62.70
	$CEAF_E$	88.95	73.98	80.78	91.75	67.56	77.82	82.69	59.36	69.11
	BLANC	51.58	63.38	52.64	50.08	55.12	49.80	50.93	57.39	51.28
	TOTAL F_1			61.90			56.81			54.02
mdDS$\bigcup$mdIOBA	MD	84.34	56.01	67.32	67.19	39.96	50.12	75.16	45.72	56.85
	MUC	22.33	38.51	28.26	3.87	22.83	6.62	11.95	29.58	17.03
	B^3	79.46	90.83	84.76	72.57	96.22	82.73	66.41	89.08	76.09
	$CEAF_M$	76.83	76.83	76.83	70.59	70.59	70.59	63.65	63.65	63.65
	$CEAF_E$	87.96	77.99	82.68	90.17	68.31	77.73	81.72	61.74	70.34
	BLANC	53.70	66.35	55.89	50.22	57.09	50.11	51.66	58.77	52.48
	TOTAL F_1			**65.68**			**57.56**			**55.92**
gold mentions	MD	100	100	100	100	100	100	100	100	100
	MUC	25.33	47.12	32.95	6.24	45.40	10.98	13.92	40.96	20.78
	B^3	79.03	92.04	85.04	71.70	97.70	82.71	64.24	91.45	75.47
	$CEAF_M$	77.53	77.53	77.53	71.33	71.34	71.34	64.23	64.23	64.23
	$CEAF_E$	89.84	78.11	83.56	92.46	68.44	78.66	85.77	61.20	71.44
	BLANC	53.87	69.11	56.27	50.38	65.16	50.37	52.02	63.93	53.10
	TOTAL F_1			67.07			58.81			57.00

Table 5.28: Results for the different mention extraction modules in UBIU within the SemEval-2 shared task; MD evaluates the extraction of mentions; the best F-scores per metric and language are marked in bold.

	CA	DU	EN	GE	IT	SP	average
mdPOSP-5	35.51	20.03	39.63	42.67	39.79	37.32	35.83
mdDS	53.40	37.73	**58.82**	64.29	57.04	55.62	54.48
mdIOBA	53.18	37.07	57.91	61.90	56.81	54.02	53.48
mdDS⋃mdIOBA	**55.78**	**43.16**	58.59	**65.68**	**57.56**	**55.92**	**56.12**
gold mentions	56.23	38.07	61.24	67.07	58.81	57.00	56.40

Table 5.29: Repeated total scores from table 5.27 and table 5.28 and the calculated for them corr-language average.

provided by mdIOBA, which has a direct influence on the identification of the head words for those phrases (see chapter 6). Since during the coreference resolution process our approach strongly relies on the mention information provided mainly by the head words, mdDS has a considerable advantage over mdIOBA. The combination of the output of both methods (mdDS⋃mdIOBA) provides both higher recall as well as a mixture of the types of identified phrases. This also leads to an enhancement of scores in comparison to the performance by mdDS mainly for MUC, because all languages apart from English (mdDS – 15.33% in comparison to mdDS⋃mdIOBA – 14.95%, see table 5.27).

Another fact worth noting is the comparison between the system performance of the CR pipeline when using mdDS⋃mdIOBA mentions and when *gold* mentions are employed. First, across almost all languages, *gold* mentions lead to higher overall scores. Yet, the improvement for any of the languages does not increase by more than 3 percent points, while for Dutch the scores even decrease by 5.09 percent points (mdDS⋃mdIOBA – 43.16% and *gold* mentions – 38.07%, see table 5.29). The scores indicate that employing a combination of both approaches, mdDS⋃mdIOBA, for mention identification reaches almost optimal performance for our system.

5.2.3.2 *Discussion*

The results presented in section 5.2.3.1 show that selecting a MD method for a multilingual coreference resolution system is not a straightforward task and that this selection should not be solely based on the intrinsic evaluation of the mention detection approach, because the type, quality and quantity of resulting mentions does affect the overall system performance. For this reason an evaluation within the whole CR pipeline is a necessary prerequisite.

However, modifications and improvements on the knowledge that the final coreference resolver uses (for example in the form of feature selection, as presented in chapter 7) will also have a direct effect on the interaction of the MD module with the resolution pipeline. This fact signifies that adjusting and tuning a multilingual CR pipeline can prove to be a highly complex task,

because all combinations of options within the set of selected languages must be carefully examined and constantly reevaluated before an optimal setting can be chosen.

The most valuable outcome of our evaluation until now is the fact that the combination of a rule-base and a machine learning method has a positive effect on the performance of most languages, which motivates its employment further in other mention detection modules.

5.2.4 CoNLL-2012 *Intrinsic Evaluation*

Unfortunately, an objective quantitative comparison between the methods presented to this point and mdCP is not possible because of the differences in the linguistic annotations of the datasets in SemEval-2 and CoNLL 2012. Moreover, there are various adaptations of the data model that we use (e.g. considered sentence window, extracted features, etc.), as a consequence of the differences in annotation. For this reason, in order to create a possibility for comparison, in the following evaluation we integrate the baseline method mdPOSP as well as the mdIOBA method that makes use solely of POS information. Note that for simplicity, we refer to *mdIOBA-POSP* as mdIOBA in the current section. The baseline and mdIOBA will allow us to gain a more precise and concrete idea of the efficiency and robustness of the mdCP method in this evaluation setting. We employ only POS information, since in section 5.2.2.1, we showed that POS information is sufficient for the implementation of mdIOBA and that syntactic information can only minimally improve the method's performance. Similar to the evaluation of the methods employed on the SemEval-2 datasets, we provide both quantitative (section 5.2.4.1) and qualitative (section 5.2.4.2) analysis of the output of the mdCP method and discuss the findings overall in section 5.2.4.3.

5.2.4.1 *Quantitative Analysis*

Unlike the SemEval-2 datasets, the CoNLL 2012 datasets did not include singletons in their annotations. This renders mention detection evaluation, when it is not included in the CR pipeline, more complicated and hard to analyze. This is the result of the scoring problem, due to a higher number of singletons in the response in comparison to the key, that we described in section 5.2.1. Thus, the results that we present in the following section put an even stronger emphasis on the significance of recall over precision and F-score.

Using the scoring software provided by the CoNLL 2012 shared task, we evaluated all three approaches (mdPOSP, mdIOBA and mdCP) on both *gold* and *auto* linguistic annotations provided in the training and development sets[8] of all three languages.

[8]Note that the final test set is not freely available at the time this chapter is written.

			mdPOSP	mdIOBA	mdCP	mdCP⋃mdIOBA
auto	AR	R	34.55	46.01	87.93	**89.08**
		P	14.34	56.37	17.95	17.58
		F_1	20.27	50.67	29.81	29.37
	EN	R	41.27	56.10	89.96	**91.44**
		P	23.77	63.93	30.77	29.48
		F_1	30.17	59.76	45.85	44.59
	ZH	R	30.57	44.45	88.54	**90.43**
		P	11.88	57.02	31.81	30.49
		F_1	17.12	49.96	46.80	45.61
gold	AR	R	35.35	47.49	95.65	**96.17**
		P	15.94	56.65	19.69	19.10
		F_1	21.97	51.67	32.67	31.88
	EN	R	41.55	56.43	94.23	**95.45**
		P	24.14	64.15	32.41	30.97
		F_1	30.54	60.04	48.23	46.77
	ZH	R	31.32	45.84	98.23	**99.31**
		P	12.48	57.49	34.19	32.43
		F_1	17.85	51.01	50.72	48.89

Table 5.30: Results from the *mdPOSP*, *mdIOBA*, *mdCP* and *mdCP⋃mdIOBA* mention detection methods across all three languages of the CoNLL 2012 shared task in both *auto* and *gold* settings. The highest recall figures are marked in bold.

We note that for Arabic there is an important difference between the POS information provided in the *gold* and *auto* settings. While within the *auto* setting the datasets contain morphologically-poor POS labels (e.g. NN, JJ, PRP, etc.), the *gold* part-of-speech annotations contained morphologically-rich tags (e.g. DET+NOUN+CASE_DEF_NOM, DET+ADJ+CASE_DEF_NOM, PV+PVSUFF_SUBJ:3MS, etc.). As we discussed in the introduction section of the mdPOSP method, section 5.1.2, mdPOSP would only reach its expected performance when morphologically-poor tagsets are used. The employment of complex tags increases the complexity of the memorized patterns significantly. Thus, the frequency counts for the various pattern types are drastically decreased. For this reason, within the *gold* setting we make use of a n:1 mapping of morphologically-rich to morphologically-poor tags. The n:1 mapping combines all morphologically-rich tags of a given category (n number of tags) in one morphologically-poor tag of the same category.

For example, the following POS tags are a small subset of all morphologically-rich variants of the proper noun category from the Arabic *gold* POS annotation layer:

- DET+NOUN_PROP
- DET+NOUN_PROP+NSUFF_FEM_SG
- DET+NOUN_PROP+NSUFF_FEM_SG+CASE_DEF_ACC
- NOUN_PROP+NSUFF_FEM_SG

Instead of using these complex tags, the n:1 mapping will substitute them with their morphologically-poor variant – NNP.

The system results are presented in table 5.30. We note again that the reason to favor recall is that, as explained in section 5.2.1, every mention that is not detected at this stage cannot be considered at all by the coreference pipeline further on. The results in table 5.30 confirm the fact that the lack of singletons in the key set increases the difficulty in evaluating mention detection on its own.

Although we cannot directly compare the datasets of the two shared tasks, when recall is considered, the baseline mdPOSP reaches a performance close to the range reached for the six languages in the SemEval-2. This indicates that with respect to mention detection both tasks were of similar difficulty. The latter is also confirmed by the performance of the mdIOBA method, which also keeps the trends established within the evaluation on the SemEval-2 datatsets. mdIOBA reaches considerably higher recall than the baseline across all three targeted languages using identical annotation layers.

From the figures in table 5.30, we can also see that mdCP reaches very high recall, which indicates that the approach is able to identify almost all mentions, both potentially coreferent and singletons. However, the English dataset included event coreference, which we did not target. This means that all non-nominal mentions that mdCP does not label have a direct detrimental effect on recall for English.

Altogether, considering that we favor recall, the mdCP method performs highly competitively and can be employed easily in a multilingual setting, given that the needed annotation layer (phrase structure) is provided. Moreover, similar to mdDS, this method is completely dependent on the existence and quality of this type of annotations.

During the evaluation of mdDS and mdIOBA we showed that a combination of rule-based and machine learning methods leads to enhanced mention detection performance. Thus, we approached the evaluation of the unification of the sets of mentions detected by both mdCP and mdIOBA (referred to as mdCP$\bigcup$mdIOBA further on) for the CoNLL 2012 datasets as well. The results of this evaluation setting are given in table 5.30. The figures confirm our findings from the evaluation of the methods applied on the SemEval-2 data. For all

#	token	ParseBit	mdCP	mentionID
0	Then	(TOP(S(ADVP*)	-	-
1	welcome	(VP*	-	-
2	to	(PP*	-	-
3	the	(NP(NP*	(1|(2	-
4	official	*	-	-
5	writing	*	-	-
6	ceremony	*)	2)	-
7	of	(PP*	-	-
8	Hong	(NP(NML*	(3|(4	(1|(2
9	Kong	*)	4)	2)
10	Disneyland	*))))	3)|1)	1)
11	.	*))	-	-

Table 5.31: An example sentence from the CoNLL-2012 English dataset with mdCP annotations listed in column *mdCP*. Column *ParseBit* shows the syntactic parse for the sentence and column *mentionID* gives the set of *gold* mentions.

three languages and within both *gold* and *auto* evaluation settings the recall for mention detection is highest for mdCP$\bigcup$mdIOBA.

5.2.4.2 *Qualitative Analysis*

mdCP is one of the methods that provided best overlap between the mention annotation scheme of the CoNLL 2012 shared task and the identified by the module mentions. As the example sentence in table 5.31 shows, all extracted mentions (listed in column *mdCP*) have identical boundaries as the NPs found in the syntactic parse, plus the additional NE for the phrase *Hong Kong*. Because those boundaries are derived from constituent phrases within the syntactic parses they represent well formed noun phrases which increases the quality of the resulting markables. As we mentioned in the quantitative analysis of this method, errors may be due to a lack of overlap between some NEs and the key mentions as well as the fact that we do not cover event coreference for which verbs should have been additionally extracted.

In contrast to the mentions identified by mdDS, which we needed to adapt additionally for the task, mdCP provides the capability of extracting not only longest span NPs, but as well smaller, embedded phrases. This increases the method's flexibility and adaptability to various annotation schemes, since mdCP will be able to provide a highly adequate mention detection procedure for both *longest* and *all spans* annotation scheme approaches without any additional and language dependent adaptation.

5.2.4.3 *Discussion*

Overall, mdCP bears the means of a highly valuable mention detection method that can be successfully and easily applied within a multilingual coreference resolution system. However, similar to all rule-based methods presented within the SemEval-2 evaluation, it also suffers from high dependency on provided annotation layers – in that particular case the presence of constituency parses. Thus, mdCP shows once more that rule-based approaches to multilingual mention detection may prove highly efficient, easily applicable for more than one language and exceptionally well performing, but rather inapplicable when the annotations they strongly rely on are not provided in the data. In general, shared tasks, such as SemEval-2 and CoNLL-2012, aim at including various linguistic annotations, but this is an aim that to some extent restricts the number of languages that can be targeted. Languages that have less resources are not as easily supported, such as English and German for example. For this reason, once again it becomes important to concentrate on algorithms and methods, such as the solely POS driven mdIOBA variant, that do not require such complex data analysis and can still perform competitively within the multilingual coreference resolution task.

5.2.5 CoNLL-2012 *Extrinsic Evaluation*

This section aims at evaluating the mdCP method within the full coreference pipeline analogous to the evaluation presented for the methods employed on the SemEval-2 datasets. With this, we again aim to revise not only the performance of the algorithm on its own but as well to see what its effect on the overall system performance is. Thus, in section 5.2.5.1 we present the achieved results and in section 5.2.5.2 we offer a follow up discussion.

5.2.5.1 *Results*

The scores that the coreference resolution system achieves when employing the various mention detection methods within the CoNLL-2012 shared task datasets are listed in table 5.32. We report results for all three CoNLL-2012 languages: Arabic (AR), English (EN) and Chinese (ZH). We also list results from all evaluation metrics in addition to the averaged scores, because the separate figures from the various metrics give a more objective and detailed presentation of the actual system performance. We also note once more that the performance of mention detection listed in rows *MD* in table 5.32 indicates mention detection after coreference has been performed and respectively after the mentions identified as singletons have been removed.

The figures in table 5.32 show highly interesting aspects of the evaluation. While mdCP outperformed mdIOBA with a wide margin within the intrinsic evaluation in section 5.2.4, the scores from the extrinsic evaluation in the current

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
POSP	MD	14.28	33.33	20.00	21.92	65.18	32.80	17.30	69.23	27.69
	MUC	6.66	20.83	10.10	18.68	59.03	28.38	12.50	40.00	19.04
	B^3	36.26	91.50	51.94	38.58	89.28	53.88	33.85	93.93	49.76
	$CEAF_M$	29.29	29.29	29.29	36.26	36.26	36.26	30.32	30.32	30.32
	$CEAF_E$	44.12	17.46	25.03	46.89	18.97	27.01	49.18	17.57	25.90
	BLANC	49.82	47.89	48.38	57.08	77.12	59.94	50.22	55.98	49.25
	TOTAL			32.95			41.09			34.85
mdIOBA	MD	22.64	77.58	35.06	45.97	76.86	57.53	31.69	77.92	45.06
	MUC	12.22	43.23	19.05	38.22	57.78	46.00	26.69	56.91	36.34
	B^3	36.22	86.73	51.10	48.98	68.28	57.04	44.81	77.36	56.75
	$CEAF_M$	35.25	35.25	35.25	40.20	40.20	40.20	38.12	38.12	38.12
	$CEAF_E$	51.61	20.71	29.56	40.81	24.10	30.30	45.81	22.14	29.85
	BLANC	52.66	60.86	53.29	63.79	60.97	62.11	60.54	58.17	59.14
	TOTAL			37.65			47.13			44.04
mdCP	MD	20.71	74.83	32.45	62.04	66.53	64.21	39.23	69.55	50.16
	MUC	15.03	57.92	23.87	49.76	50.38	50.07	33.08	53.29	40.82
	B^3	38.06	91.55	53.77	60.71	57.72	59.18	51.55	73.71	60.67
	$CEAF_M$	37.37	37.37	37.37	43.49	43.49	43.49	41.71	41.71	41.71
	$CEAF_E$	52.88	21.06	30.12	34.44	33.86	34.15	42.21	25.98	32.16
	BLANC	53.59	73.61	54.89	65.21	59.20	61.03	63.04	60.71	61.73
	TOTAL			**40.00**			**49.58**			**47.42**
mdCP⋃mdIOBA	MD	46.01	43.38	44.66	62.38	64.55	63.44	42.03	59.22	49.17
	MUC	24.19	18.72	21.11	50.23	49.00	49.60	34.56	41.73	37.81
	B^3	60.27	44.36	51.10	61.89	56.46	59.05	56.75	60.08	58.37
	$CEAF_M$	27.11	27.11	27.11	43.21	43.21	43.21	35.09	35.09	35.09
	$CEAF_E$	21.33	28.11	24.25	33.15	34.31	33.72	31.61	25.77	28.39
	BLANC	50.47	50.11	48.12	65.95	59.53	61.50	58.78	52.36	51.60
	TOTAL			34.34			49.42			42.25
gold mentions	MD	100	100	100	100	100	100	100	100	100
	MUC	41.66	81.37	55.11	65.59	76.70	70.71	45.88	76.20	57.28
	B^3	47.08	90.04	61.83	61.54	61.10	61.32	51.49	75.83	61.34
	$CEAF_M$	51.58	51.59	51.59	51.00	51.01	51.00	47.26	47.26	47.26
	$CEAF_E$	70.24	31.35	43.35	53.63	36.59	43.50	58.19	28.27	38.05
	BLANC	58.31	81.40	61.91	69.57	62.10	64.05	66.68	62.49	64.17
	TOTAL			54.76			58.12			53.62

Table 5.32: Results for the MD modules in UBIU within the CoNLL-2012 shared task; the best total scores per language are marked in bold.

section are not this far apart. Spread considerably balanced across all metrics, mdIOBA's average performance within the coreference pipeline is exceptionally close to the performance reached by mdCP with differences smaller than 3.5 percent points (Arabic – 2.35 percent points, English – 2.45 percent points and Chinese – 3.38 percent points) across all three languages. This indicates that even though mdIOBA does not reach comparably high recall with respect to mdCP, it provides mentions with which the coreference resolution system can reach almost the same performance as the one achieved by mdCP. Moreover, comparing the mention detection performance of the latter methods for Arabic (*MD* rows for mdIOBA and mdCP in table 5.32) we can see that after coreference resolution, mention detection for mdIOBA reaches higher scores than MD for mdCP for all reported measures (recall, precision and F-measure). This is due to the fact that mdCP identifies all noun phrases in the Arabic data. Yet, Arabic is a highly NP-rich language. As a result, too many mentions are included in the resolution process. This drastically increases the complexity level for the MBL classifier and leads to a decrease in MD performance when the full CR pipeline is employed.

Another interesting outcome of the evaluation in the current section is the fact that the combination of mdCP and mdIOBA, or alternatively the combination of the mentions they identify, does not lead to enhancement of performance as was observed for the methods within the SemEval-2 shared task. Yet, this can again be explained with the increased resolution difficulty posed by the high number of mentions resulting by the unification of the two sets of mentions. mdDS⋃mdIOBA reached better scores than both methods apart, because neither of the methods achieves an exceptionally high recall alone, unlike mdCP. Since mdCP reaches optimal recall on its own, mdCP⋃mdIOBA provides collections of mentions that are a mere overgeneration of phrases. Thus, the resolution process is overloaded, which is directly reflected in the scores achieved by mdCP⋃mdIOBA in comparison to those for mdCP only. mdCP⋃mdIOBA leads to a decrease in performance of 5.66 percent points for Arabic, 0.16 percent points for English and 5.17 percent points for Chinese.

Comparing the overall scores between *gold mentions* and *mdCP* in table 5.32, we can see that there is a large difference between the overall performance across all three languages (14.76 percent points for Arabic, 8.54 percent points for English and 6.2 percent points for Chinese). While mdCP achieves scores for English and Chinese closer to the ones gained by the use of *gold* mentions, the figures for Arabic indicate a bigger gap. Again, we assume that this is the result of the fact that Arabic is a language with a highly NP-rich structure and exceptionally long sentences which increases the number of mention pairs proportionally and decreases the performance of the resolution process itself.

5.2.5.2 *Discussion*

The extrinsic evaluation of the mdCP method once more reveals that the selection of a mention detection algorithm for a multilingual approach within the coreference resolution task is a difficult endeavor. mdIOBA's performance shows that even though achieving high recall is important for mention detection, lower recall, such as the one achieved by mdIOBA, can prevent excess numbers of mentions and thus overgeneration or overloading of the resolver, which is observed in mdCP's performance. Thus, mdIOBA reaches almost as good a performance as mdCP within the CR pipeline. The latter also shows that mdIOBA is highly reliable as a machine learning method, being able to lead to competitive performance for all three languages. However, when syntactic parses are present in the annotations, rule-based methods could be used as more reliable and better performing multilingual mention detection methods than the machine learning one – mdIOBA. Yet, the language independent nature of mdIOBA, as a result of its exceptional flexibility and competitiveness is further confirmed as one highly positive outcome of our work.

5.3 SUMMARY AND CONCLUSION

One of the main goals of this exploratory chapter was to investigate in more detail to what extent the mention detection subtask of coreference resolution can be performed in a close to language independent manner. What we learned from our diverse observations is that such an attempt is easily achievable even when basic layers of standard linguistic annotations, such as part-of-speech tags, are provided. We showed that a machine learning method, such as mdIOBA can achieve competitive performance to rule-based approaches to the problem and can be designed in a fully language-independent manner. The machine learning nature of the method also provides various possibilities of improvement based on the selection of different learner algorithms, parameter optimization and feature selection.

Even though that we can point to POS information as rather sufficient for the development of a competitive, robust and reliably performing approach to multilingual MD, we do not consider it commensurate with the design of a high performance system. We showed that using further information as dependency relations, for example, can boost the system performance. Our investigation also demonstrated that annotation layers such as NEs do not provide efficient and uniform representation for the development of stand-alone methods across languages. However, we can conclude that named entities can always be employed as an additional enhancement of the backbone method in use, because they have a high chance of being regarded as a mention and do not always correspond to well formed noun phrases.

The availability of additional annotation layers, apart from POS, has certainly an enhancing effect on multilingual mention detection. Yet, another question

that we aimed to provide an answer to is which layers can supply the most indicative information for that process. Our approaches to both SemEval-2 and CoNLL-2012 attest that mention detection can be achieved with high accuracy across multiple languages if annotations that provide the possibility for the direct or indirect derivation of phrase structure, such as all types of syntactic parses, are included in the targeted datasets. Syntactic parses include a highly accurate and abundant delineation of the noun-phrase structures, as our investigations on the mdDS and mdCP methods showed, which ranks them as most helpful and providing most indicative information for tackling multilingual mention detection.

A highly important conclusion that we can accomplish, based on our observations, is that phrase structures can be indicative only if the annotation scheme of the underlying mentions overlaps to a high degree with the noun phrases existing in the data. Yet, the results achieved by mdCP strongly confirm our assertion. The presence of high correspondence between noun phrases and mentions also leads to the acquisition of close to language independent MD methods that can perform equally across the distinct languages.

Altogether, we reviewed multiple issues concerning the application of each method to more than one language and approached an exhaustive quantitative and qualitative analysis of the problems. Moreover, we also exemplified that a reliable evaluation of the mention detection module can only be achieved if singleton mentions are included in the set of key mentions.

We believe, that our findings should be considered further during the annotation, accumulation and dissemination of datasets for that highly important and complex task that serves as the basis for every coreference resolution approach. Additionally, our observations can be exceedingly beneficial for future attempts to the development of multilingual MD modules in diverse system architectures.

CHAPTER

6

MENTION HEAD DETECTION

Our work in chapter 5 showed that mention detection is a highly important and a relatively complex subtask of CR, because mentions that are not identified in the text cannot be rendered coreferent. We discussed how this task can be approached in a multilingual context and what the various advantages and disadvantages of the employed methods are when more than one language is targeted. However, within the CR pipeline that we presented and exploited, the MBL coreference classifier does not make use of the full mentions, or nominal phrases, for the resolution process. In fact, a representation of all extracted mentions is assembled by collecting only their syntactic heads and additional features describing a variation of their context characteristics. The latter implies that mention detection is tightly followed by the identification of mention heads when pairwise classification are attempted.

Since most approaches to coreference resolution only target one language, the identification of the heads of the extracted mentions was mostly seen as a technicality by the computational linguistic community. It was either a subtask of mention identification or performed by simple heuristics further down in the CR pipeline. However, in the current chapter we address the increased complexity of the problem when the task is intended to cover more than one specific language. The problem arises by the difference in the typology of the languages according to the position of the heads of their nominal phrases.

Within this chapter, we also propose that for pairwise approaches, such as the one that we employ, Mention Head Detection (MHD) is considered as a separate subtask of coreference resolution equally important to mention detection rather than being a minor subtask within mention detection. We consider this as an important aspect, because multilinguality introduces new problems that are hardly solvable by simple heuristic rules. Moreover, the qualitative identification of mention heads is significantly important for the coreference resolution process, as syntactic heads determine the syntactic category of the phrase and carry highly relevant semantic information about it. Thus a failure to identify the actual mention head can be compared to a failure to identify the actual mention.

In order to provide an exhaustive delineation of the problem, we first outline the differences between head-initial and head-final languages (section 6.1) and the problems phrase directionality may pose for the head detection problem. In section 6.2, we present the way multilingual mention head detection was approached within the two multilingual shared tasks that we concentrate on: SemEval-2 and CoNLL 2012. Following, in section 6.3, is the presentation of three different solutions to the multilingual MHD task, which are thoroughly evaluated and compared in section 6.4. We propose a new, machine learning method for mention head detection that allows for more flexibility and language independency and assess it against the a heuristic and rule-based approach to the problem. Finally, section 6.5 sums up our findings and concludes the current chapter.

6.1 HEAD-INITIAL VS. HEAD-FINAL LANGUAGES

head

In phrase structure, the ***head*** of a phrase is the lexical item within that phrase that determines its syntactic nature. Thus, the head, or also known as the

syntactic head

syntactic head of a noun phrase can either be a common or a proper noun or a pronoun.

Across different languages, the head of a phrase can be placed in various positions within that phrase. In most languages and especially in those in which the verb precedes both the subject and the object in a sentence, the noun phrases

head-initial
head-first

are ***head-initial*** or also known as ***head-first***. These languages are known as

VSO languages
SVO languages
SOV languages

V(erb)S(ubject)O(bject) languages (e.g. Hebrew, Irish, Zapotec). Other, more commonly seen variations are the ***SVO languages*** (e.g. English, Mandarin, Russian) or ***SOV languages*** (e.g. Hindi, Japanese, Latin). Head-initial phrases are phrases in which the syntactic head is placed in the beginning of the full structure. This often correlates the directionality of branching within the phrase

left-branching
right-branching

structures and thus these are also called ***left-branching*** languages. However, in reversed directionality, or in ***right-branching*** languages the syntactic head of the phrases is placed in final position and analogically to the already presented

head-final

head-initial and head-first languages, these languages are known as ***head-final***

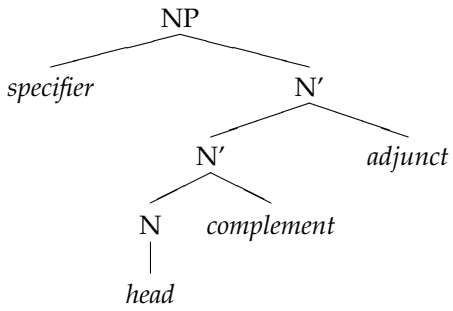

Figure 6.1: A potential N′ phrase structure for a head-initial language with the syntactic head placed after the specifier, but before the complement and the adjunct.

or ***head-last*** languages. ***head-last***

Within the standard ***X-bar theory*** notation [Chomsky, 1970, Jackendoff, 1977], which accounts for the structural integrity of phrases across languages and tries to identify syntactic similarities between them, the structure of a phrase regarding its head, complements, specifiers and adjuncts can be represented as in the rules in (36). ***X-bar theory***

(36) $X' \rightarrow X$, (complement)
$X' \rightarrow X$
$X' \rightarrow$ (complement), X
$XP \rightarrow X'$
$XP \rightarrow$ (specifier), X'
$XP \rightarrow X'$, (specifier)
$X' \rightarrow X$
$X' \rightarrow$ (adjunct), X'
$X' \rightarrow X'$, (adjunct)

Without focusing on a specific language, these rules can be combined in a variation of structures allowing for all possible permutations of the positioning of the daughter nodes of the phrases. One possible variant of the noun phrase structure according to the above presented rules, representing head-initial phrases, is as shown in figure 6.1. If the language is head-final, the structure will have the form as shown in figure 6.2. We note again that the head in those structures must not necessarily be a common noun, but can also be represented by a proper noun or a pronoun and that the specifiers, complements and adjuncts are only optional components of the phrase.

In general, languages tend to be consistent in their nature and are thus either uniformly head-initial or head-final with respect to one type of phrase. Japanese is often given as an example of a language that places the head consistently in last position across all types of phrases. For example, let us

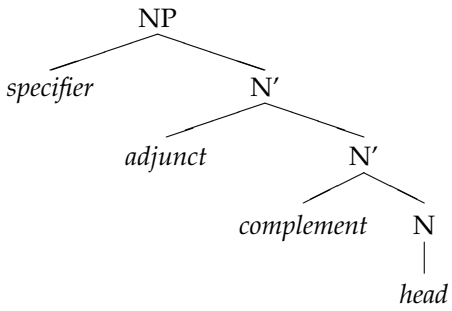

Figure 6.2: A potential N′ phrase structure for a head-final language with the syntactic head placed after the specifier, the adjunct and the complement.

consider the noun phrase *yellow flower*. In Japanese the head noun *flower* takes the last position in the phrase as in example (37). However, in consistently head-initial languages as Spanish, for instance, that order will be reversed and the head noun *flower* will be in first position, as shown in example (38).

(37) 黄色の花
yellow flower

(38) *flor* *amarill-a*
flower yellow-*fem*

In natural language, it is not always the case that noun phrases are as simple as in examples (37) and (38). If the slots for specifier, complement and adjunct are filled complex structures, such as the one in figure 6.3, are built. Such phrases pose a greater difficulty for the head detection procedure.

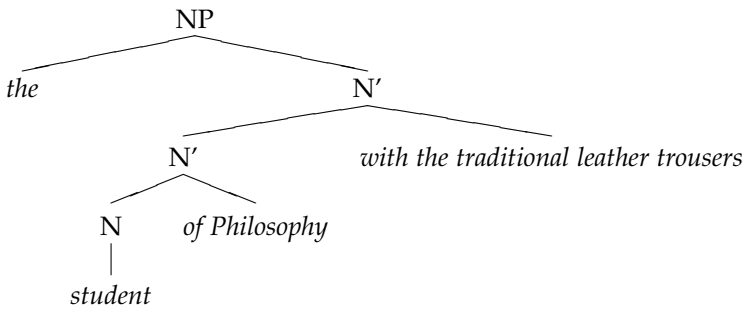

Figure 6.3: The outline of the N′ phrase structure for the complex noun phrase *"the student of Philosophy with the traditional leather trousers"*.

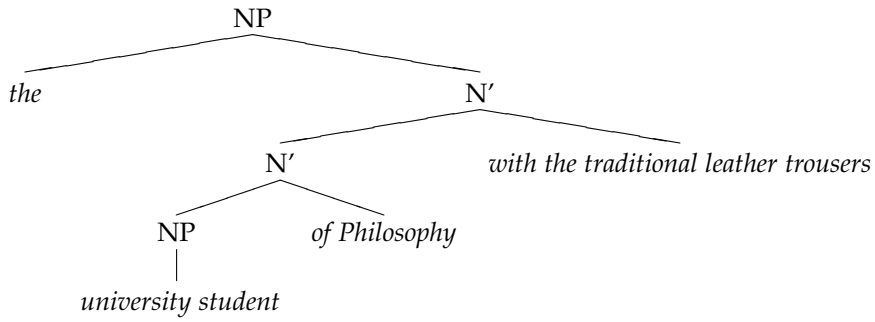

Figure 6.4: The outline of the N′ phrase structure for the complex noun phrase *"the university student of Philosophy with the traditional leather trousers"*.

Furthermore, if the head noun is also premodified, as shown in figure 6.4, a deeper knowledge of the phrase structure of the given language is needed in order to be able to identify the correct token as the syntactic head of the phrase. For one targeted language, this is often an easily-solvable task. However, if diverse languages with increased phrase complexity are considered, the problem of head identification becomes more ambitious. The following section addresses this issue in more detail.

6.2 ISSUES IN MULTILINGUAL MENTION HEAD DETECTION

Targeting a single language for coreference resolution implies that only one directionality of branching has to be taken into account and thus the head of a given mention can be identified easily via simple heuristics (see section 6.3.1). However, including multiple languages in the resolution task can lead to diversity within the observed nominal structures, meaning that more language-specific knowledge needs to be considered and represented in a more complex description than simple heuristics. Additionally, heuristics are not always capable of capturing difficult cases, such as the one presented in figure 6.4. In the current section, we discuss the way contemporary multilingual CR systems handle directionality (see section 6.2.1) and then show further problems that they might face when dealing with different languages (see section 6.2.2).

6.2.1 *Overcoming Directionality*

In the CoNLL-2012 shared task the three languages, Arabic, English and Chinese, constituted a good mixture of noun phrase directionality: while in Chinese the head of a noun phrase is always in phrase-final position, English is generally known to be a head-initial language, for it places "heavier" con-

construct state
status constructus

stituents, as adjuncts, more towards the end of a phrase. In Arabic, phrase directionality is even more elaborate considering the ***construct state*** or also known as ***status constructus***, which leads to a mixed directionality of that language (see section 6.2.2). In short, the construct state is composed for semantically definite nouns that are modified by another noun that is in a genitive construction.

For this reason, most of the systems participating in the shared task that needed to perform MHD relied on more complex rules that captured the peculiarities of the language-specific noun phrase structures. Chen and Ng [2012], for example, use rules, such as the ones presented by Collins [1999]. Furthermore, Björkelund and Farkas [2012] employ Choi and Palmer [2010]'s perculation rules for Arabic and English and the rules of Zhang and Clark [2011] for Chinese. Li et al. [2012] also use different sets of rules for English[1] and for Chinese[2]. The system presented by Martschat et al. [2012] relies on the SemanticHeadFinder (an implementation[3] of the rules presented by Collins [1999]) for English, while the head detection for Chinese is provided by the SunJurafskyChineseHeadFinder (an implementation[4] of the rules presented by Sun and Jurafsky [2004]), both components of the Stanford Parser[5], yet the authors do not target coreference resolution for Arabic. Uryupina et al. [2012] also employ the rules by Collins [1999] for English and different heuristic rules for the rest of the languages: for Arabic, the first noun/pronoun in a sequence is selected to be the head, in Chinese, the last noun/pronoun is appointed as the head. The authors also note the importance of such predefined and well documented collections of rules and address the fact that the absence of more detailed linguistic knowledge can become an issue when such rules are to be developed manually for each separate language.

The various approaches to mention head detection show that noun phrase directionality is an important part to consider during that process. However, developing rule sets for each separate language targeted in a multilingual MHD approach requires in-depth knowledge of the nominal phrase structure of each of the languages. For example, let us consider the role of common titles in the three languages of the CoNLL-2012 shared task. While in English and Arabic common titles precede the proper name as in figure 6.5 a), in Chinese they are placed in phrase-final position figure 6.5 b). This poses a problem during head identification, because selecting the last noun in such phrases for Chinese (because Chinese is a head-final language) would wrongly appoint the title to be the head of the phrase. In English, this phenomenon does not cause further

[1] http://w3.msi.vxu.se/~nivre/research/headrules.txt
[2] http://w3.msi.vxu.se/~nivre/research/chn_headrules.txt
[3] http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/trees/SemanticHeadFinder.html
[4] http://tides.umiacs.umd.edu/webtrec/stanfordParser/javadoc/edu/stanford/nlp/trees/international/pennchinese/SunJurafskyChineseHeadFinder.html
[5] http://nlp.stanford.edu/software/lex-parser.shtml

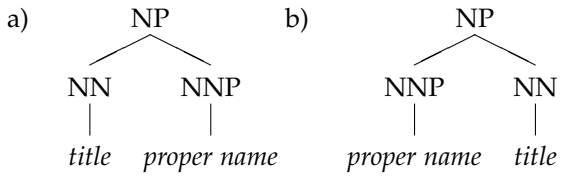

Figure 6.5: The structure of noun phrases that contain common titles. In a) the head is in phrase-initial position, while in b) it is in phrase-final position.

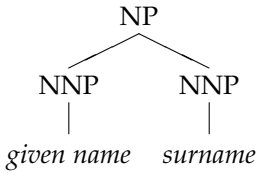

Figure 6.6: Phrase structure of personal proper names situating the given name in phrase-initial position and the surname in phrase-final position.

complications, because in general the last noun is selected as the head and in the structure given in figure 6.5 a) that is the proper name. Arabic, however also follows the type of structure as the one shown in figure 6.5 a). However, the head of Arabic noun phrases is generally in phrase-initial position. Thus, for this language, as for Chinese, the title will be erroneously selected as the head of the phrase. Heuristic approaches to mention head detection will not manage to capture such exceptions. More complex rules can make use of predefined lists of titles for each language and employ them during the identification process.

However, common titles are not the only examples in which directionality is not the most precise indicator. The case of complex proper names is very similar. Let us consider for instance the general pattern for personal proper names shown in figure 6.6. Typically, the surname is appointed to be the head of such types of phrases. Yet, heuristic approaches will not capture this difference for head-initial languages selecting the given name to be the head, as in Arabic within the CoNLL-2012 shared task. Again, in such cases, additional rules can be considered to exclude proper nouns from the head-initial heuristic.

The common title and proper names exceptions are rather simplistic deviations from the standard position of the head in these three languages. However, more complex situations might be posed by various language specific phenomena for which a deeper linguistic knowledge is necessary.

6.2.2 *Construct State in Arabic*

Apart from the language differences demonstrated by the head-initial and head-final nature of noun phrases across the various languages, further complications might be caused by exceptional cases such as the construct state in Arabic.

Semitic languages

As Shlonsky [2003] presents, the Construct State (CS) is a type of genitive noun phrase form that is often seen in ***Semitic languages***, such as Arabic and Hebrew for example. Another form that the noun phrases in Semitic languages can take is the Free State (FS). Both states differ significantly in their headedness. While the FS is close to the prepositional genitive construction in Romance languages, the CS strictly requires that the head noun precedes the determiner. Let us take for instance Shlonsky [2003]'s Moroccan Arabic comparisons displayed in example (39) representing the FS and example (40) representing the CS respectively.

(39) *d-dar* *dyal* *l-wazir*
the-apartment of the minister
'the minister's apartment'

(40) *d-dar* *l-wazir*
apartment the minister
'the minister's apartment'

Another peculiarity of the construct state is that there is a strict adjacency of the members of the phrase imposed by the construct. This means that no modifiers can appear between a noun and its complements. The difference is depicted in example (41), displaying this case for the FS, and example (42) and (43) showing the contrast for the CS.

(41) *d-dar* *l-was^{c}a* *dyal* *l-wazir*
the-apartment the-spacious of the minister
'the minister's spacious apartment'

(42) $*$ *d-dar* *l-was^{c}a* *l-wazir*
apartment the-spacious the minister
'the minister's spacious apartment'

(43) *d-dar* *l-wazir* *l-was^{c}a*
apartment the minister the-spacious
'the minister's spacious apartment'

(44) *l$^{\circ}$-amn-i* *majlis-i* *qaraaraat-i* *jamii^{c}-i* *taTbiiq-u*
security council resolutions all application
'the application of all of the resolutions of the Security Council'

In addition, as Ryding [2005] reports, the CS allows for more than one embedded genitive construction with up to five included nouns as in example (44). In this form the construct state contains a head for each embedded genitive construction and one for the construct itself. As can be seen in example (44), the actual head of the full construct, namely *application*, is placed in phrase-final position. This phenomenon can hardly be captured by a heuristic and thus more language specific approaches have to be employed.

6.3 METHODS FOR MULTILINGUAL MENTION HEAD DETECTION

Since the main aim of this thesis is to assess the possibility of performing multilingual coreference resolution and to evaluate potential language independent approaches, in the following section we present three different methods to MHD that can be employed in a coreference resolution system and discuss their applicability and performance when more than one language is targeted.

We include a baseline heuristic approach that we present in section 6.3.1, a rule-based approach that requires more language specific knowledge (see section 6.3.2) and finally we present a machine learning method (section 6.3.3) that attempts to capture the various language phenomena in a more language independent manner.

We concentrate our discussion on these methods and their performance within the setting of the CoNLL-2012 shared task. We evaluate them on an excerpt of the dataset for each of the separate languages. The limitation to the smaller part of this data and not to the full set is necessary, because the development of a memory-based learning approach, such as the one we propose (see section 6.3.3), requires manually annotated instances. The latter are not provided within the data sets of the SemEval-2 and within those of the CoNLL-2012 shared task. Section 6.3.3 also introduces the employed excerpt of the data.

6.3.1 *Heuristic Approach*

The Mention Head Detection Heuristic (mhdH) approach only includes phrase directionality of the targeted language for the identification of the mention head. Within head-initial languages it would appoint the first noun, proper noun or a pronoun in a noun phrase to be its head, while in a head-final approach the last one is picked. Since this is a baseline approach that should require a minimal implementation effort, we do not look at any further language specific conditions.

As simple as it is, we assume that this method is well applicable to languages with strong and consistent directionality, such as Arabic (being head-initial, apart from CS forms, such as in example (44) on page (154)) and Chinese (being head-final). Let us take as an example the three mentions presented in

#	token	POS	ParseBit	mention ID	MHDH
0	the	DT	(NP(NP	(1|(2	-
1	long	JJ	-	-	-
2	call	NN	)	2)	1|2
3	from	IN	(PP	-	-
4	the	DT	(NP	(3	-
5	president	NN	)))	3)|1)	3

Table 6.1: Example of the mhdH output for toy example mentions.

table 6.1. As we noted in the beginning of section 6.2.1 English places "heavier" constituents such as complements and adjuncts after the head, which makes it a head-initial language. For this reason the mhdH method will identify the noun *call* as the head for both mentions 1 and 2, while the noun *president* will be extracted as the head of mention 3.

Because of the fact that simple noun phrases in English, such as *the long call* or *the president* that do not contain complements or adjuncts are indeed head-final, English may be observed as a head-final language with respect to noun-phrase directionality. However, if this is the case, in the example shown in table 6.1, the mhdH method will select correctly the nouns *call* and president for mentions 2 and 3 respectively, but it will wrongly pick the noun *president* to be the head for mention 1. This confirms that English should be regarded as head-initial language with respect to noun-phrases.

The simplicity of the mhdH heuristic allows its employment in an exceptionally straightforward and relatively language independent way. However, we assume that mhdH will not be able to account for a wide range of language specific exceptions as well as mixed language directionality, which will not render it competitive against other methods for mention head detection. Yet, the only knowledge needed to employ this heuristic is the predominant directionality of the targeted language.

6.3.2 *Rule-Based Approach*

The Mention Head Detection Rule-Based (mhdR) approach is developed around the idea of the mhdH heuristic. However, mhdR extends the approach with additional language specific knowledge by defining a set of rules for each separate targeted language. Each set includes rules that cover various phenomena specific to the language in order to provide a more exact and precise detection of the phrase heads for the distinct languages. Such rules are generally easy and straightforward to create, yet they need the presence of deeper knowledge of the language that they are developed for.

#	token	POS	ParseBit	mention ID	mhdH	mhdR
0	a	DT	(NP(NP*	(1|(2	-	-
1	King	NNP	(NML*	(3	1|2|3	-
2	Kong	NNP	*)	3)	-	3
3	type	NN	*)	1)	-	1|2
4	of	IN	(PP*	-	-	-
5	animal	NN	(NP*))))	(4)|2)	4	4

Table 6.2: Example of the mhdR output for toy example mentions.

In our work, for both SemEval-2 and CoNLL-2012 shared tasks, the sets of rules created for all languages were achieved by assembling a rule for each of the separate language specific phenomena occurring in the first 100 sentences of the training data. For all languages apart from English, language experts were consulted to ensure the correctness of the resulting rules. Since Arabic and Chinese have a highly consistent directionality, the rules for these languages included mainly specifications for the position of the head with respect to common titles and proper names as well as rules for coordinated phrases that have more than one head. However, for English additional rules were added improving on the identification of heads that are preceded by modifying nouns/pronouns.

In order to gain a better idea of the functionality of this approach, let us consider an English example and the tokens listed in table 6.2. As shown in column *mhdH*, the mhdH method would extract token #1, *King*, for the first three mentions, since it is the first noun in all those phrases. Yet, what mhdR would return is the heads listed in column *mhdR* in table 6.2. In order to achieve that, mhdR uses multiple additional rules. One of them ensures that mhdR selects as a head not the first, but rather the last noun in a sequence of nouns or proper nouns. This sequence should not be preceded by other nouns/pronouns in the phrase. For this reason, instead of selecting *King* as the head of mention 3, mhdR takes the last noun in that sequence and thus outputs *Kong*. Another rule defines that the head should be positioned before any complements or adjunct, which indicates that the head for mention 1 in table 6.2 should be found before the preposition in token #4. Moreover, before token #4 there is a sequence of nouns and according to the first rule we select the last noun in that sequence – token #3, which is thus appointed as the head of mention 2. The first rule also helps for the identification of the head of mention 1 – token #3, while no specific decisions need to be met for extracting the head of mention 4 as it consists of a single token.

The various rules that the mhdR approach includes provide a more sophisticated, language specific and linguistically reasoned approach to the selection

#	token	POS	ParseBit	mentionID	head
0	a	DT	(NP(NP*	(1|(2	-
1	focus	NN	*)	2)	1|2
2	of	IN	(PP*	-	-
3	worldwide	JJ	(NP*	(3	-
4	attention	NN	*))))))	3)|1)	3

Table 6.3: Example of the mhdML annotations within the excerpt data of the English dataset from the CoNLL 2012 shared task. The *head* column lists the manually annotated heads of the *gold* mentions presented in column *mentionID*.

of the heads for all mentions found by the mention detection module. Yet, assembling such rules assumes a highly detailed knowledge of the grammar of the targeted language, which defeats the purpose of an easily applicable and multilingual concept.

6.3.3 *Machine Learning Approach*

Similar to the task of mention identification, detection of mention heads via a specifically designed and trained memory-based classifier can provide a language independent solution to the problem. Since developing, assessing and integrating multilingual approaches in the coreference resolution pipeline and thus improving it in a general, not language specific way, is one of our main aims in the current work, we employ machine learning also for the task of MHD. The alternative to the rule-based method for detecting mention heads, mhdR, which we presented in section 6.3.2 is the Mention Head Detection Machine Learning Based (mhdML) approach.

Unfortunately, neither the SemEval-2 nor the CoNLL 2012 shared tasks provided gold data annotations labeling the heads of the mentions included in the datasets. Yet, in order to make the implementation of such an approach possible and as well to have an output from mhdML that is reasonable and comparable to the other approaches (e.g. mhdH and mhdR), we manually annotated an excerpt of the dataset for each of the three languages in the CoNLL 2012 shared task. An example of the *gold* annotations for English is shown in the *head* column in table 6.3. Manual annotation of examples is a tedious and expensive endeavour. Thus, we are not able to target all languages within both shared tasks. Moreover, the CoNLL 2012 shared task provides us with a greater typological diversity over three language families and for this reason we use its datasets for the following investigation.

The dataset for Arabic consists of an excerpt of 42 documents for training and 8 for testing. For English, we include 20 documents as a test set and 100

as the training set, while for Chinese 84 documents serve the purpose of a training set and 16 are the test set. This is a feasible amount of data to train and test a machine learning model across the three languages. Even though Arabic has a lower number of included documents, it results in a high number of extracted mentions, since Arabic has a very NP-rich syntactic structure. We note that annotating the full datasets would not be manageable within the scope of the current work. However, our investigation will provide a deeper insight into the potential representation of the problem and its solution within a machine learning framework. This knowledge can be used in the future for the preparation of further datasets and as the basis for the development of such ML-based frameworks.

Using the training examples that could be created from the data of the manually annotated training set, we designed and trained a MBL classifier. It observes the instances on per-token basis and uses features representing that token within a given mention. The features also describe the context outside of the mention – for example, whether the mention is positioned in a PP, SBAR, VP, S. For the representation of that information, we include features that depict the POS annotations of the target token as well as of the tokens before and after it. Furthermore, we include features describing the position of that token within the mention and providing more detailed information about it in the context of that mention. A full list of the features is given in table 6.4. As an implementation of a memory-based classifier, we again used TiMBL (see section 4.2) without additional parameter optimization. The optimal output from the machine learning approach to mention head detection for our exploration is the one in which a highly accurate classifier can be trained via the use of a set of simple and not language dependent features, such as the ones we employ (see table 6.4). In the future, a more detailed investigation of the feature selection for a wider diversity of learning algorithms can be approached. At that point, we propose a base set of features that is language independent and that can provide a good estimate for the syntactic head across the languages we target.

The reason for us to consider mhdML is that we aim to evaluate the possibility for casting the subtask of coreference resolution, namely MHD, as a more sophisticated machine learning method rather than simple heuristics or a collection of language specific rules. This will allow for more flexibility and adaptability of that particular CR subtask to a wider number and type of languages. This is a key point to multilingual approaches that perform mention head detection as a separate task within their full coreference resolution pipeline. Additionally, we would like to raise the question of the need for providing that type of annotations (annotation for mention heads) within standard annotation distributions, in order for good, efficient and most importantly language independent approaches to be developed.

#	Feature Description
1	the target token
2	part-of-speech tag of the target token
3	part-of-speech tag of token$_{-1}$
4	part-of-speech tag of token$_{+1}$
5	Y if it is the only token in the mention; else N
6	Y if it is not in a PP, SBAR, VP, S; else N
7	Y if it is the first token in the mention; else N
8	Y if it is the last token in the mention; else N
9	Y if the target token is a noun
10	Y if the target token is a pronoun
11	Y if the target token is a noun or a pronoun
12	Y if the target token is followed by a noun
13	Y if the target token is followed by a pronoun
14	Y if the following token is possessive and the last token in the mention

Table 6.4: The proposed set of 14 initial features used by the mhdML classifier. The features provide information on per token basis within a given mention.

6.4 EVALUATION OF ALL MHD METHODS WITHIN THE EXCERPT DATA

As we already explained in section 6.3.3, the only fully comparable evaluation (both intrinsic and extrinsic) between all three methods, presented in sections 6.3.1 through 6.3.3, can be achieved by using only the manually annotated subsets for all three languages of the CoNLL 2012 shared task: Arabic, English and Chinese. For this reason, the following section (section 6.4.1) presents the results of an intrinsic evaluation of the MHD performed by the three approaches within that data excerpt. For the sake of completeness and objectiveness, we also offer an extrinsic evaluation for that data set in section 6.4.2.

Our main aim is to evaluate the possibilities that mhdML offers in terms of overall performance within the coreference pipeline as well as alone with respect to its rule-based opponent and the baseline heuristic. We also aim to assess the multilinguality and flexibility of all methods as well as the type of errors their outputs have, which is further mirrored in our qualitative analysis in section 6.4.1.

language	metric	mhdH	mhdR	mhdML
AR	R	0.79	0.83	**0.85**
excerpt	P	0.87	0.88	**0.91**
	F_1	0.83	0.85	**0.88**
EN	R	0.65	**0.92**	0.87
excerpt	P	0.70	0.97	**0.98**
	F_1	0.67	**0.95**	0.92
ZH	R	0.84	0.96	**0.97**
excerpt	P	0.98	0.98	**0.99**
	F_1	0.90	0.97	**0.98**

Table 6.5: Mention head detection for the excerpt datasets for all three languages of the CoNLL 2012 shared task (Arabic (AR), English (EN) and Chinese (ZH)), considering all spans for each mention; highest F-scores are marked in bold.

6.4.1 *Intrinsic Evaluation*

The intrinsic evaluation presented in the current section introduces both a quantitative (see section 6.4.1.1) and a qualitative (see section 6.4.1.2) analysis of the performance of all three approaches. We note, that this assessment examines the performance of all mention head detection approaches alone without being included in the full coreference pipeline. We use the excerpts for which mentions were extracted automatically via the mdCP method presented in section 5.1.4.

6.4.1.1 *Quantitative Analysis*

Providing a quantitative evaluation of MHD as a task on its own is not as straightforward an enterprise as the quantitative intrinsic evaluation of mention detection was. The reason for this is that the scoring software provided by both shared tasks, SemEval-2 as well as CoNLL 2012, does not include calculation of the scores for the detection of heads for all included mentions (the only figures the scorers provide are precision, recall and F-measure for identification of mentions and coreference performance). For this reason, we calculated these metrics (precision, recall and F-measure) for each of the approaches against the manually annotated data. We used the respective formula listed in equations 2.1, 2.2 and 2.3 on page 33. Table 6.5 lists the scores achieved by the mhdH, mhdR and mhdML approaches separately when all possible spans for each mention are considered. The best scores are marked with bold.

mhdH

The figures achieved by mhdH support the baseline nature of this method. Selecting only the first noun/pronoun in a given mention as its head leads mostly to low recall scores, which is directly mirrored in the final F-measure. The weakness of the mhdH is best exemplified by its performance for English, which is a head-initial language but also a language in which nouns can be placed in the specifier position and thus be situated before the head. As can be seen in table 6.5, for English, the mhdH method reaches overall performance of 25 percent points lower than the performance reported for the other two approaches. However, an F-score of 67% for a heuristic method as mhdH is a good overall performance. Languages with strong phrase directionality for which the head is placed more consistently in first or last position within the phrase, such as Arabic and Chinese, reach higher overall performance. As we previously discussed, phenomena such as the CS in Arabic may also pose a problem to a potential heuristic approach. This is directly mirrored in the overall lower F-score for this language (83%) with respect to Chinese (90%). Easy and efficient applicability to languages, such as Chinese, is a noteworthy advantage of the heuristic method and proves that it can be employed for multilingual approaches targeting languages with strong and consistent phrase directionality. However, languages, such as English, show that the heuristic is not always sufficient and in a multilingual approach this finding provides the motivation for further exploration of possible solutions to the problem.

mhdR

The next scores included in table 6.5 are those of the rule-based head detection approach – mhdR. What the figures in table 6.5 show, is that mhdR achieves higher scores with respect to mhdH for all languages. For Arabic the improvement is the smallest consisting of 0.04 percent points for recall, 0.01 percent points for precision and 0.02 percent points for the final F-measure. This shows that the exceptional cases, such as common titles, proper names and difficulties posed by the CS, which are handled by the rules for this approach do not occur very often in the excerpt data. For English the improvement is a lot higher: recall – 27 percent points, precision – 27 percent points and respectively F-measure – 28 percent points. The figures for English confirm the difficulty that nominal specification poses for the head detection problem and the fact that manually created rules are more efficient than the mhdH heuristic. With an F-score of 95%, mhdR accomplishes the best performance across all three approached methods for English. For Chinese, similar to Arabic, the improvement in scores is due to increased recall – 0.12 percent points higher. This improvement also results to a 0.07 percent points increase of the final F-score for this language.

mhdR's good performance shows once more that rule-based approaches to the task can reach exceptionally good overall scores that are reliable in both

recall and precision. As we described in section 6.3.2 the set of rules assembled for all languages, is based on observations on the phenomena occurring in the first 100 sentences in the training data. The latter fact implies that not all exceptional cases are covered by the employed rules and thus the performance of the mention head detection module can be improved if additional rules are developed and added to the existing collection. However, language expertise is necessary for every new targeted language and, thus, highly inconvenient for a multilingual coreference resolution approach, such as the one we present in our work. Bearing that in mind, we can state that close to optimal performance can be achieved by the mhdR, but only if deeper knowledge of the targeted language is present.

The last observation once more underlines the need of approaches that can provide the efficiency of a rule-based method, but at the same time can be more language independent in terms of less language specific knowledge and the capability to account for exceptional cases in a better and more flexible manner. Thus, the last approach that we consider in our experiments is the one presented in section 6.3.3, namely mhdML.

mhdML

The results presented in table 6.5 indicate a highly interesting and as well very successful outcome. The machine learning method mhdML achieves better scores with respect to its rule-based opponent for Arabic (with 0.03 percent points for the F-measure) and Chinese (0.01 percent points for the F-measure), while for English the final score is decreased with 0.03 percent points. These figures show that the machine learning approach can easily cope with languages that have comparatively consistent phrase headedness, such as Arabic and Chinese, for which mhdML achieves results better than those reached by mhdR. In general, it is not often the case that machine learning approaches outperform rule-based ones, but in this case this behaviour can be explained by the relative simplicity of the problem and the capability of the learner to account for new and unseen examples. However, the behaviour observed for English is more typical, which is due to the increased difficulty for this language.

The overall outcome of this evaluation is highly positive and the benefits of our results are twofold. We introduced a method for MHD that is language independent, relying on the features presented in table 6.4. The mhdML approach performs highly competitively to a language dependent rule-based approach across three very typologically different languages. Additionally, the high scores achieved by the method indicate that the mention head identification problem is easily applicable to machine learning and particularly to memory-based learning, as in our case. Moreover, the fact that such high scores were achieved with a relatively small amount of training data (100 documents) shows that corpora can be prepared for a wide number of languages with-

#	token	POS	mention ID	mhdH	head
0	leather	NN	(1	1	-
1	jacket	NN	1)	-	1

Table 6.6: Example of a noun placed in specifier position.

out much annotation effort. Including a bigger set of labeled instances will additionally improve the robustness and accuracy of the method.

Altogether, the quantitative intrinsic analysis of the performance of the mhdML method shows that the task of mention head detection can be easily and successfully approached on a multilingual level with a high efficiency, which is in favor of one of our main aims – transforming the coreference resolution pipeline to a more flexible, language independent and efficient combination of approaches.

6.4.1.2 *Qualitative Analysis*

Similar to the analysis presented in section 5.2.2.2 and 5.2.4.2, the qualitative analysis of the mention head detection approaches assessed by us will undergo a deeper examination that will provide a more detailed understanding of the type of errors seen in the outcomes of all methods. We believe that further improvements on all developed approaches can be achieved only if an exhaustive evaluation for them is conducted. The following paragraphs provide an overview of the errors seen within the first 100 sentences of the outcome of each separate method.

mhdH

The baseline heuristic mhdH selects by definition the first/last seen noun/pronoun of each phrase as its syntactic head, depending on the phrase directionality the targeted language has. With respect to English and Arabic, which are head initial languages, the mhdH method selects the first noun/pronoun of each phrase. For Chinese, the last one is selected.

The outcome of the mhdH, however, reveals several problems with that approach. First, as we already noted in the quantitative analysis, English has a phrase structure in which the specifier precedes the head within the phrase. Yet, the position of the phrase specifiers in English can be filled in various ways. For example, let us consider the Finite State Automata (FSA) representation of possible specifier fillers represented in figure 6.7. We note that for simplicity reasons the FSA represents an overgenerating automata allowing for multiple occurrences of each of the transitions apart from a transition via a *head noun*. Furthermore, the order in which the transitions may occur is also not specified

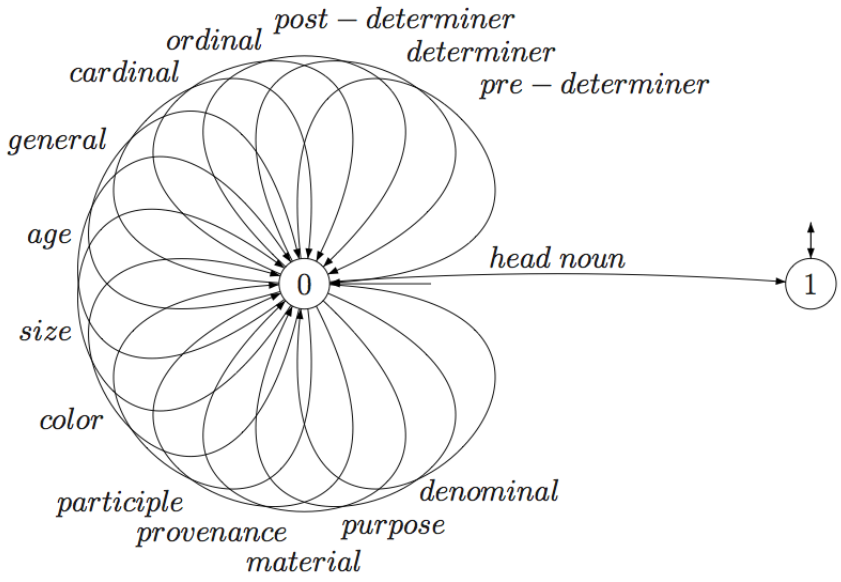

Figure 6.7: A finite-state representation of the possible specifiers in an English noun phrase.

by that representation. An important observation that can be made from the automaton in figure 6.7 is that various nouns can precede the head noun, in the slot for provenance or material for example. One such example is the mention presented in table 6.6 in which the noun *leader* appears before the head noun *jacket* filling in the material specifier position. In those cases, mhdH would wrongly select the first noun as a head (visualized in column *mhdH*) with respect to the correct selection, the token marked as the head in the gold data (visualized in column *head*).

Similar to the latter example is the manipulation of compound nouns consisting of several nouns and written separately (e.g. *phone call*, *fish tank*, etc.), proper names consisting of more than one token (e.g. *Hong Kong*, *Lantau Island*, *Daisy Duck*, etc.), common titles preceding proper names (e.g. *Professor Liu Jiangyong*, *President Barack Obama*). All these cases are wrongly handled by mhdH for English, because always the first noun is selected whereas the last one is the head.

Furthermore, not only nouns preceding the head noun cause erroneous selections by the mhdH method. Such can be also observed when a pronoun precedes the head noun as in the mention listed in table 6.7.

Another type of error that leads to a decrease in recall is the fact that mhdH does not extract a head for the mentions in which no noun or pronoun is

#	token	POS	mention ID	mhdH	head
0	their	PRP$	(1	1	-
1	unique	JJ	-	-	-
2	charm	NN	1)	-	1

Table 6.7: Example of a pronoun preceding the head noun.

#	token	POS	NE	mention ID	mhdH	head
0	first	JJ	(ORDINAL)	(1)	-	1

Table 6.8: Example of a mention that is not a noun phrase and respectively does not include a noun or a pronoun.

included for any of the targeted languages. In case named entities, which do not necessarily correspond to noun phrases, are extracted by the MD module, phrases that do not include nouns or pronouns can be present in the set of mentions that the MHD module further uses. For example, consider the mention presented in table 6.8. Since the token is marked as a named entity (see column *NE*) that phrase is considered for mention head detection. Yet, the MHD heuristic will not allow adjectives to be heads of mentions and thus will not extract this token.

The last type of error that we will look at in our qualitative analysis is connected with the inability of mhdH to cope with any kind of coordinated phrases. Since each coordinated noun phrase embedded in a NP has its own head, the mhdH method cannot properly deal with complex NPs that contain coordination. For example, let us consider the three mentions listed in table 6.9. As can be seen in column *head*, the heads in the gold annotations are altogether four – one for each embedded noun phrase (mention 1 and mention 3) separately and two for each of the heads of the complex NP (mention 2). mhdH identifies only two of those four heads correctly by picking the first noun for mention 3 as the head, although the second one is the syntactic head of the phrase, and ignoring the coordination in mention 2 and selecting only the first coordinate as the head of the complex noun phrase.

The only solution to avoiding the errors that this simple heuristic leads to can be found in enhancing it with further language specific rules, which account for the above mentioned phenomena. Yet, the usage of sets of rules is the scope of the rule-based approach to mention head detection. So, let us have a look at the problems that more advanced and language specific rules cannot cope with either.

#	token	POS	ParseBit	mention ID	mhdH	head
0	dolphins	NNS	(NP(NP*)	(1)|(2	1|2	1|2
1	and	CC	*	-	-	-
2	sea	NN	(NP*	(3	3	-
3	lions	NNS	*))))	3)|2)	-	2|3

Table 6.9: Example of the output of mhdH for coordinated phrases.

#	token	POS	ParseBit	mention ID	mhdR	head
0	two	CD	(ADVP(NP*	(1|(2	-	-
1	years	NNS	*)	2)	1|2	2
2	ago	RB	*)	1)	-	1

Table 6.10: Example of the output of mhdR for non-nominal mentions.

mhdR

As a rule-based approach, mhdR aims at addressing most of the shortcomings of the previously discussed heuristic – mhdH. Respectively, mhdR correctly extracts the last noun within a series of nouns/pronouns in mention initial position and thus appropriately handles common titles, proper names and other noun modifiers containing nominals and preceding the head noun. It also allows for the extraction of heads for mentions that do not contain a noun/pronoun. As we showed in our quantitative analysis, the rule-based approach achieves significantly higher scores for all languages with respect to the heuristic, however, there are multiple issues that it cannot cope with properly at that stage.

For example, let us consider mention 1 in table 6.10. As is shown in the *ParseBit* column, that mention (*two years ago*) is an adverbial phrase and thus its correct syntactic head is token #2 – the adverb *ago*. However, under the assumption that all mentions are noun phrases, mhdR has a strong preference for nouns or pronouns when selecting a head. Thus, in that case it prefers token #1 – the plural noun *years*. Including additional rules that would allow the selection of other word classes when a noun or a pronoun is present, depending on the given constituency labels in the *ParseBit* column is not a hard task and can easily be accomplished if needed. Yet, the majority of those phrases represent dates for which the adverbs do not carry sufficient semantic information. Thus, the trade off between lower MHD performance but more informative semantic information in that particular case can be accepted.

#	token	POS	ParseBit	mention ID	mhdR	head
0	a	DT	(NP(NP*	(1|(2	-	-
1	fishing	NN	*	-	-	-
2	harbor	NN	*)	2)	2	1|2
3	one	CD	(ADVP(NP(QP*	(3|(4	-	-
4	hundred	CD	*)	-	-	-
5	years	NNS	*)	4)	1|3|4	4
6	ago	RB	*)))	3)|1)	-	3

Table 6.11: Example of the output of mhdR.

Another problem, specific to English, that mhdR cannot cope with is the issue depicted in table 6.11. The correct head for mention 1 is token #2 – the noun *harbor*, which is not accurately identified by mhdR. This is so for cardinal numbers as well as nouns are allowed in specifier position, preceding the syntactic head. Thus, token #5, *years*, is selected as the head of the phrase. Yet, something that figure 6.7 did not unambiguously show is that in English the multiple word classes used as noun modifiers can only appear in a relevant order [Lam, 2004], shown in example (45) below. That order shows that cardinals cannot appear after nouns within the specifier slot and thus a rule specific to the English order of word classes allowed in specifier position can be included to the set of rules mhdR relies upon.

(45)
1. pre-determiner
2. determiner
3. post-determiner
4. ordinal
5. cardinal
6. general
7. age
8. size
9. colour
10. participle
11. provenance
12. material
13. purpose
14. denominal

#	token	POS	ParseBit	mention ID	mhdR	head
0	Bank	NNP	(NP(NP*)	(1)|(2	1|2	1
1	of	IN	(PP*	-	-	-
2	China	NNP	(NP*	(3	-	-
3	Tower	NNP	*)))	3)|2)	3	2|3

Table 6.12: Example of the output of mhdR for non-nominal mentions.

Another possibility for the resolution of the problem is to consider the constituency information provided in the *ParseBit* column. If such information is available, the head of recursive noun phrases can be defined as the head of the first embedded NP in a complex noun phrase, such as mention 1, in table 6.11. Looking deeper into the syntactic structure can also help with more complex problems posed by coordination as for example the identification of the heads for the phrase "*the 2005 Ningxia Investment and Trade Fair and Ningxia-Taiwan Economic and Trade Cooperation Seminar*", as *Fair* and *Seminar*.

Yet, neither the information provided in the *ParseBit* column nor the rules can cope with exceptional cases such as the one in table 6.12. As can be seen from the syntactic parse (see column *ParseBit*) the given representation for that mention is as visualized in figure 6.8 a). However, that structure implies that *Bank* is the head of the mention, yet, in the given discourse, it is not the bank that is referred to but the tower of the Bank of China. Thus, *Tower* should be correctly selected as the head for which a different phrase structure should have been included in the annotations, as visualized in figure 6.8 b). Such exceptional examples, however, are not easy to handle, especially if the structural ambiguity has affected the correctness of the constituency annotations.

Altogether, the qualitative analysis of the mhdR approach shows that, as any rule-based method, mhdR can be further improved and additional language specific rules can be added in order to cope with exceptional or more complex cases as the few presented here. However, such improvements can be time-consuming and even impossible if a deeper knowledge of the targeted language is not present. This is a great difficulty in a multilingual setting, such as the one we investigate.

mhdML

mhdML meets the needs for our multilingual investigation exceptionally well. As the quantitative analysis for this approach showed, mhdML reaches very good and competitive overall performance for English and it outperforms its rule-based opponent for Arabic and Chinese. The errors that this approach has in its output, though, are similar in nature to the ones we reported for

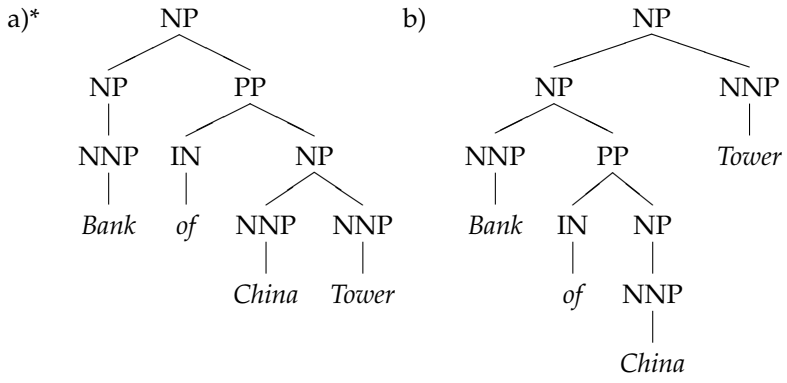

Figure 6.8: Possible phrase structures for the mention *Bank of China Tower*.

mhdR. For example, let us look at the mention in table 6.13. Since the features used by the mhdML classifier (listed in table 6.4 on page 160) do not include any information about the constituency label of the mention in which the tokens are, it is difficult for the resolver to pick the correct head in that particular case. Such errors can be addressed by an extended feature set that contains more detailed information about the phrase itself in addition to the information provided about the token.

Another error type that can be found in mhdML's output is the selection of nouns/pronouns as heads that occur after punctuation marks within the mention, which are clear indicators for post-modification of the head. Such an example is, for instance, the head selected by mhdML for mention 1 that is listed in table 6.14. As can be seen in column *head*, displaying the gold annotations for mention heads, the head of mention 1 is token #1, the proper noun *Kong*. Instead, mhdML selects token #5, the proper noun *Paradise* as the head of mention 1. Such errors are not often seen in the output of mhdML, which is confirmed by its good overall performance. Yet, we assume that extending the feature set with a feature to designate the existence or lack of punctuation marks, as well as other types of indicators for post-modification (e.g. prepositions, WH-words, etc.) will exclude that type of error from the output of the memory-based method.

#	token	POS	ParseBit	mention ID	mhdML	head
0	high	RB	(NP(NP(ADJP*	(1	-	1
1	above	JJ	*))	1)	1	-

Table 6.13: Example of the output of mhdML for non-nominal mentions.

#	token	POS	ParseBit	mention ID	mhdML	head
0	Hong	NNP	(NP*	(1|(2	-	-
1	Kong	NNP	*))))	2)	2	1|2
2	:	:	*	-	-	-
3	a	DT	(NP(NP*	(3	-	-
4	Shopping	NNP	-	-	-	-
5	Paradise	NNP	*)	3)|1)	3|1	3

Table 6.14: Example of the output of mhdML for post-modification.

#	token	POS	ParseBit	mention ID	mhdML	head
0	Dr.	NNP	(TOP(S(NP(NP*	(1|(2	-	-
1	Robert	NNP	*	(3	-	-
2	Mann	NNP	*)	3)|2)	2|3	1|2|3
3	a	DT	(NP*	(4	-	-
4	forensic	JJ	*	-	-	-
5	anthropologist	NN	*))	4)|1)	1|4	4

Table 6.15: Example of the output of mhdML for post-modification without a good indicator.

Similar to all previously presented methods, all errors and inconsistencies in the annotations also have direct detrimental effect on the performance of mhdML as well as ambiguities similar to the one presented by the mention *Bank of China Tower*, shown in table 6.12.

Additionally, features representing the constituency structure within the mention can be considered in order to better capture coordinated phrases or head post-modification that is not easily identifiable by the POS information, as for example in the instance presented in table 6.15.

Our qualitative analysis for mhdML shows that the types of errors that this method makes are very similar to the errors reported for mhdR, although both approaches are of completely different nature. As we noted above, we assume that further improvement can be achieved by a more comprehensive and exhaustive set of features that will account for exceptional phenomena, similar to adding new rules to the existing set used by mhdR. However, not only the quantitative but also the qualitative analysis shows, that mhdML provides a highly competitive performance to mhdR and thus we consider it an exceptionally good solution towards robust and multilingual mention head detection.

6.4.1.3 *Discussion*

Section 6.4.1 provided an exhaustive quantitative and qualitative intrinsic evaluation of all three approaches to the MHD subtask of coreference resolution for all three targeted languages – Arabic, English and Chinese. We dealt with MHD on its own and not within the coreference pipeline. Our analysis showed a highly interesting outcome indicating that mention head detection is a task that can easily be solved by simple heuristics when it is approached for one specific language, especially if that language has a consistent either head-initial or head-final directionality. This fact explains the well established practice within state-of-the-art approaches to use such heuristics to tackle the task. Yet, as we showed in the current section, heuristic approaches are not sufficient when a high overall accomplishment is targeted for more than one language. In such cases rule based approaches provide better performance as they allow for the development of sets of rules that precisely describe the phenomenon in each language. However, those rules are language dependent and need an exceptionally thorough and deep knowledge of the targeted language in order for accurate rules to be acquired. Following our main aim, namely the exploration of multilingual and as well highly efficient approaches, we presented and evaluated a memory-based method for MHD.

From both, our quantitative and our qualitative evaluation, it is evident that MBL is easily applicable to the problem and provides a highly competitive solution. We showed that mhdML outperforms mhdR for languages consistent with respect to their directionality, such as Arabic and Chinese and reaches very competitive performance for more complex cases, such as English. mhdML produces errors of similar types to the rule-based approach. Within our analysis, we also provided suggestions for enhancing the performance of both approaches, mhdR and mhdML, according to the errors we observed in the data. Even though our investigation indicated that mhdML is reliable and efficient on its own, we need to confirm its consistently good performance and competitiveness within the full coreference pipeline.

6.4.2 *Extrinsic Evaluation*

An extrinsic evaluation of the approaches within the CR system is needed, in order to ensure the correctness of our findings and to examine the effect of the proposed approaches to MHD in the full coreference pipeline. Thus, the current section, section 6.4.2, aims to provide an extrinsic evaluation of all three approaches for mention head detection presented by now: mhdH, mhdR and mhdML. Similar to the intrinsic evaluation (section 6.4.1), the only comparable evaluation of all three methods is only possible if the manually annotated excerpt of the datasets for all languages is taken into account. For this reason, the results we report in section 6.4.2 are based on the employment of these excerpts only.

Following, in section 6.4.2.1 we will provide the results the UBIU multilingual coreference resolution system (see section 4.3) achieves and in section 6.4.2.2, we provide a consecutive discussion after which we conclude the chapter.

6.4.2.1 *Results*

Table 6.16 lists the overall system performance for all three languages (Arabic, English and Chinese) and for all three methods (parts in the table: *mhdH*, *mhdR* and *mhdML*) as well as the final scores of the coreference resolver when *gold* mention heads are made use of (part *gold heads*). We report the evaluation scores for mention detection (MD) and additionally we provide all detailed figures that are reported by each evaluation metric by the CoNLL-2012 shared task scoring software (for more information on the scoring procedure, see section 5.2.1). For an easier comparison and overview, averaged numbers from all metrics' F-scores in the form of total scores are provided (marked as TOTAL in table 6.16).

The results achieved by the system, listed in table 6.16, indicate that all the tendencies between the evaluated methods from the intrinsic evaluation (presented in more detail in section 6.4.1) are kept within the extrinsic evaluation as well.

The heuristic method mhdH reaches overall performance lower than both mhdR and mhdML across all three languages. The lower performance of mhdH is characterized mainly by low scores from the *MUC* and *BLANC* metrics, which increase for English and Chinese significantly for the mhdR and mhdML approaches, as well as for the *gold heads* setting, while for Arabic the change is not very large. The latter may be caused by the overall low improvement of scores for the Arabic language across the results achieved by the different approaches. We assume that the overall low scores of the mhdH methods are the result of the fact that by selecting the first noun or the first pronoun in a mention, mhdH cannot properly identify the links between the mentions and thus many chains are left disjoint and represented by a collection of "subchains". This is very well supported by the exceedingly low scores for the MUC metric, which is leaning towards the upper bounds when the entities are overmerged and very low, as in this case, when there are less coreference links in the output.

The results obtained by mhdR are considerably higher than those achieved by mhdH. These changes adhere to the tendencies set by our previous results. We obtain a very slight improvement for Arabic with 1.20 percent points of TOTAL score enhancement, while the changes for English (8.30 percent points) and Chinese (11.32 percent points) are more substantial. These figures confirm our findings that rule-based approaches provide a better solution to the problem. Similar to the intrinsic evaluation, the best system performance for English (48.40) was achieved by the rule-based approach, indicating that languages,

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
mhdH	MD	4.85	43.05	8.72	42.18	51.39	46.33	58.73	40.43	47.89
	MUC	1.70	18.18	3.11	24.31	28.87	26.39	42.29	30.33	35.32
	B^3	31.82	94.60	47.63	52.29	62.17	56.81	61.54	45.82	52.53
	$CEAF_M$	29.11	29.11	29.11	34.96	34.96	34.96	28.78	28.78	28.78
	$CEAF_E$	49.33	16.36	24.58	32.06	27.16	29.41	15.92	24.02	19.15
	BLANC	50.05	51.54	48.25	52.60	53.66	52.94	52.09	51.79	51.89
	TOTAL			30.54			40.10			37.53
mhdR	MD	7.35	48.95	12.78	57.09	57.57	57.33	71.80	65.37	68.44
	MUC	3.41	27.11	6.06	43.80	42.30	43.03	59.31	58.90	59.10
	B^3	33.10	93.43	48.88	61.49	59.08	60.26	51.39	62.79	56.52
	$CEAF_M$	29.79	29.79	29.79	42.55	42.55	42.55	40.59	40.59	40.59
	$CEAF_E$	49.04	17.07	25.32	32.90	34.24	33.56	24.83	25.16	25.00
	BLANC	50.22	54.74	48.66	63.94	61.59	62.61	58.93	68.44	59.83
	TOTAL			31.74			**48.40**			48.21
mhdML	MD	8.76	61.53	15.34	56.97	58.59	57.76	72.12	65.37	68.58
	MUC	4.05	35.18	7.26	42.51	42.10	42.30	59.54	58.90	59.22
	B^3	31.90	94.26	47.66	59.05	59.56	59.30	51.57	62.67	56.58
	$CEAF_M$	30.41	30.41	30.41	41.38	41.38	41.38	40.65	40.65	40.65
	$CEAF_E$	52.28	17.28	25.98	33.40	33.77	33.58	24.78	25.29	25.03
	BLANC	50.37	60.43	48.83	59.09	59.44	59.26	58.92	68.38	59.81
	TOTAL			**32.03**			47.16			**48.26**
gold heads	MD	13.30	51.51	21.14	57.09	58.49	57.78	71.66	65.94	68.68
	MUC	4.69	20.18	7.61	42.83	42.28	42.56	58.61	58.90	58.76
	B^3	36.24	87.14	51.19	59.40	59.54	59.47	49.33	62.52	55.15
	$CEAF_M$	30.87	30.87	30.87	41.65	41.65	41.65	39.37	39.37	39.37
	$CEAF_E$	46.06	18.87	26.77	33.48	33.97	33.72	24.77	24.54	24.60
	BLANC	50.26	52.50	49.09	59.28	59.50	59.38	58.08	67.40	58.51
	TOTAL			33.11			47.36			47.28

Table 6.16: Results for the three mention head detection methods: mhdH, mhdR and mhdML compared to the use of gold heads within the full coreference pipeline for both mdDS and gold mentions; the best total score is marked in bold (not regarding the *gold heads* setting).

such as English, that do not have a consistent head directionality pose a greater problem to the machine-learning approach as well. However, mhdML shows itself to be well capable of representing the problem for Arabic and Chinese, for which this method achieves best scores – 32.03 for Arabic and 48.26 for Chinese.

The last evaluation setting that we approached is shown in table 6.16 as *gold heads*, listing the scores that the multilingual coreference resolution pipeline reaches when the manually annotated heads are considered for the resolution process. Those results are of great importance to us, because they show that both mention head detection methods, mhdR and mhdML, achieve optimal or close to optimal performance with mhdR even reaching higher overall scores than the use of the actual *gold* data for English and Chinese as well as mhdML for Chinese. This is possible because the output that the methods provide is based on a consistent way of selecting a token as the head of a mention. Such consistent behaviour is more helpful for the coreference resolver, for it bases its decisions on previously seen examples, thus uniformity in the latter data enhances its performance.

The fact that mhdML achieves scores so close for English, and even better ones for Arabic and Chinese, to the ones acquired by the used manual annotations is even more important to us than the good performance of mhdR. As the numbers in table 6.16 show, the machine learning method mhdML is highly competitive and also reaches optimal scores according to all evaluation metrics. This outcome was the main aim of our work within the current chapter. It shows that mhdML is capable of providing a competitive solution to the mention head detection problem and can easily be used in a language independent manner without the need for large annotated corpora.

6.4.2.2 *Discussion*

In the current section, section 6.4.2, we aimed at providing a comparison between all three mention head detection methods that furthers the intrinsic assessment and reveals the capabilities of each approach within the full coreference pipeline. We provided results for each of the three presented methods as well as a comparison of their performance to the one achieved by manually extracted mention heads.

The extrinsic evaluation revealed observations highly important to us that confirmed the previously achieved results within the intrinsic evaluation: First, although very well-performing for a simple heuristic, the mhdH baseline does not provide a sufficient and accurate enough manner of extracting mention heads. Second, mhdR achieves optimal overall performance and could be used when a single language is targeted and when sufficient knowledge for that language is present. Yet, our most valuable observation is the fact that the machine learning mention head detection method mhdML achieves highly competitive performance to its rule-base opponent and thus also leads to optimal overall performance. This is a highly important fact for us, since mhdML is a machine learning approach that does not require the development effort of a rule-based approach, as well as an abundant amount of manual annotations (as we noted we used a small amount of annotated documents for training). Furthermore, the approach is developed in a way that can be

considered not only multilingual but rather language independent. We assume that the lower scores achieved by this method could be significantly improved with a larger manually annotated training set, so that the memory-based learner is provided with enough examples of all exceptional or less represented cases.

6.5 SUMMARY AND CONCLUSION

In the current chapter, chapter 6, we discussed the problem that mention head detection poses in a multilingual coreference resolution approach. We demonstrated that approaches targeting more than one language at a time, such as the ones participating at both shared tasks (SEMEVAL-2 and CONLL 2012) do not have a proper solution to the mention head detection problem when the most widely used baseline coreference model is made use of, namely the mention-pair model.

We showed that the heuristic baseline used from a large number of state-of-the-art approaches so far does not provide sufficient functionality for that purpose. Furthermore, we confirmed the state-of-the-art tendency that rule-based methods can be applied very successfully to the task. Yet, our analysis revealed that such methods require deeper and most importantly language specific knowledge that is not always present when more than one language is considered.

In our work, we proposed a machine learning method, mhdML, capable of tackling both the efficiency and the multilinguality problem and showed that it can be used reliably without much annotation effort. For this reason, we suggest that such annotation layers are further provided within standard layers of linguistic annotations in data distributions for the coreference resolution task and respectively for its multilingual coreference resolution extension. The latter will provide the possibility for multilingual or language independent approaches to make direct use of any provided dataset without additional language adaptations.

In addition to the analysis of the output that all discussed methods provide, we proposed suggestions for improvement according to the observed errors that each of the approaches makes. The machine learning method best supports our main goal. In order to evaluate and develop the ways in which it can be modified so that multiple languages can be targeted without much language specific tuning, we provided an initial set of features and suggestions for additional features that address its current weaknesses.

Altogether, the problems we discussed in the current chapter support our initiative to consider mention head detection not a subtask of mention detection but a proper subtask of coreference resolution and respectively of multilingual coreference resolution. We showed that this is necessary when a multilingual pipeline is assembled and that the suggested machine learning model does

provide the needed efficiency and capability to represent the problem. For this reason, the mhdML approach should be employed in the context of multilingual coreference resolution instead of the state-of-the-art tendency to use different variations of heuristic or rule-based methods.

CHAPTER

7

FEATURE SELECTION FOR MULTILINGUAL COREFERENCE RESOLUTION

Being able to properly identify all mentions in the context as well as correctly extracting the syntactic heads of all already detected mentions are the two important subtasks of coreference resolution that are directly affected by recasting the enterprise to a multilingual setting, which we thoroughly discussed in the previous two chapters, chapter 5 and chapter 6. However, as we already mentioned in chapter 4, there is a third procedure also highly important to the overall pipeline. This is the assembling of features that can most reliably and precisely describe the pairs of syntactic heads, their respective mentions and the discourse they are extracted from, which also needs to be reconsidered when more than one target language is included in the desideratum.

Our overall concerns can be covered by the following most important questions:

1. Which of the features used in monolingual approaches are applicable in multilingual or even language independent methods?

2. Is there a difference in the information gain and importance between the types of features considered by the pipeline and which type carries most descriptive and helpful information for the coreference resolver? Does that trend change across languages and which is the setting most helpful to all considered languages?

3. Are there other layers of annotations or external sources of knowledge that can enhance the resolution process and what kind of limitations do those sources have?

In order to investigate these questions, in the current chapter we intend to look deeper into the importance and the effect of the features used in the multilingual coreference pipeline. For this reason, we need to consider various languages similar to our explorations in chapter 5 and chapter 6. However, the differences in the annotations between the corpora provided in the two shared tasks, SemEval-2 and CoNLL 2012, makes a comparative study between all eight languages hardly possible. Thus, only one set of data can be selected for this investigation. SemEval-2 provides the possibility to explore six different languages at the same time, yet it covers only two different language families (Romance and Germanic) including very similar languages (e.g. Catalan and Spanish). On the other hand, CoNLL 2012 includes three typologically very different languages from three distinct language families (Semitic, Sino-Tibetan and Germanic). This gives us a better motivation to select the datasets provided from the CoNLL 2012 shared task in the following investigation. Yet, the outcome of this research and the knowledge gained from it will be further applicable to any corpora selection employed in multilingual coreference resolution.

Our exploration begins with section 7.1 and an overview of the annotation layers used for building state-of-the-art feature sets for coreference resolution (mostly monolingual coreference resolution). Then, in section 7.2, we extend the investigation to the examination of feature selection for more than one language within the UBIU multilingual coreference resolution system and the CoNLL 2011 shared task datasets. In section 7.3, we look deeper into the effect of basic linguistic annotations, such as part-of-speech tags, to the full MCR process and in section 7.4, we divide the full set of features into multiple groups depending on the type of information they provide and evaluate the informativeness of those groups for the coreference resolver. Section 7.6 concludes our findings and observations.

7.1 FEATURES FOR COREFERENCE RESOLUTION

Before we further explore the multilinguality aspect, we need to extend the knowledge on feature selection that was presented shortly in section 4.1.2 consider more precisely the types of features that are employed in state-of-the-

art approaches with respect to the information they provide about the mention pair. Thus, the following section will discuss the general selection of features for the state-of-the-art monolingual systems, presented in the context of the CoNLL 2011 shared task on English (see section 7.1.1). We will discuss the participation of UBIU in the CoNLL 2011 shared task and show an evaluation of the minimal feature set that we used for the task, as well as the effect of extending this set with ontological knowledge. On the basis of this information we will further assess the features in UBIU within the multilingual environment. Section 7.1.2, concludes the section with a short discussion.

7.1.1 *The* CoNLL 2011 *Shared Task on English*

The CoNLL 2011 shared task on "Modeling Unrestricted Coreference in OntoNotes" [Pradhan et al., 2011] is the predecessor of the multilingual CoNLL 2012 shared task presented in section 3.1.2. In general, the definition of the 2011 task is identical with the one used for the 2012 competition, with the difference that it employed English as the only language targeted by the systems. For this reason, we will not repeat the details from section 3.1.2 here as well. One of the conclusions made by the CoNLL 2011 shared task organizers that is of high interest to us directly relates to the topic of the chapter, namely feature selection. As Pradhan et al. [2011] discuss, the features that are used within state-of-the-art machine learning coreference resolution systems are highly complex. This is one of the reasons why rule-based approaches to CR are still highly competitive in comparison to machine learning ones. This was proved by the best performing system of the CoNLL 2011 shared task, which is of a rule-based nature [Lee et al., 2011].

Both Lee et al. [2011] as well as Klenner and Tuggener [2011] employed a rule-based approach to tackle the problem. However, the rest of the participating systems [Sapena et al., 2011, Chang et al., 2011, Björkelund and Nugues, 2011, Nogueira dos Santos and Lopes Carvalho, 2011, Cai et al., 2011, Uryupina et al., 2011, Zhou et al., 2011, Kobdani and Schütze, 2011, Xiong et al., 2011, Irwin et al., 2011, Lalitha Devi et al., 2011, Charton and Gagnon, 2011, Kummerfeld et al., 2011, Li et al., 2011, Zhekova and Kübler, 2011, Yang et al., 2011, Stoyanov et al., 2011, Chen et al., 2011, Song et al., 2011] were machine-learning based and included features providing different types of information about the mention pair. Most of these features are standard and well established features for CR. Selecting the most complete and informative feature set is not an easy task even when one single language is concerned, however, since features are designed and collected also depending on the type of the machine learning approach used in the CR system. Additionally, the heterogeneity of the data also needs to be analyzed. In the cases in which the feature sets consist of features of multiple different kinds (e.g. discrete, discrete ordered, counts, continuous values), some algorithms are more applicable than others. Many algorithms,

including support vector machines, linear regression, logistic regression, neural networks and nearest neighbor methods, can make better use of features that are scaled within a similar range. Approaches that make use of a distance function, such as nearest neighbor and support vector machines, are even more influenced by the type of features. One of the advantages of decision trees is that they can better handle heterogeneous data. In the following section we provide a short and selective overview of the type of system employed by each team within the CoNLL 2011 shared task (we keep the order of the systems that the task elicited and present each system separately for easier comparison across the various methods). We mainly focus on the learning algorithm the authors make use of and the type of features they include with respect to their approach. Appendix B.2 provides in a summarized form the system results released by the shared task.

7.1.1.1 *Features in State-of-the-art Systems Applied to English*

Sapena et al. [2011] use the RelaxCor system [Sapena et al., 2010], which was introduced within the SemEval-2 shared task. As the authors report, RelaxCor is a CR system based on constraint satisfaction. It represents the coreference phenomenon in a graph structure. The candidate mention pairs are connected and relaxation labeling is applied over a set of constraints so that the set of most compatible coreference relations is extracted. In order to automatically build the constraints, which are conjunctions of attribute-value pairs, an initial set of over one hundred features (that the authors call attributes) is used. Sapena et al. [2011] included attributes for distance and position of the mentions (e.g are the candidate mentions in the same sentence, are they in consecutive ones, the distance between the mentions, etc.), lexical features (e.g. string match of the full mentions, string match of their heads, etc.), morphological (e.g. number, gender, agreement information etc.), syntactic (e.g. NP definiteness, embeddedness, coordination, etc.) and semantic information (e.g. semantic class, speaker information, etc.)

Chang et al. [2011] introduce the Illinois-Coref system that uses Learning Based Java [Rizzolo and Roth, 2010]. Illinois-Coref, [Bengtson and Roth, 2008], is a system that also makes use of the mention-pair model and calculates a compatibility score for each candidate pair on the basis of the features for the pair extracted by the system. The features that Chang et al. [2011] use are the ones presented in [Bengtson and Roth, 2008]. Bengtson and Roth [2008] grouped the features in several categories: mention types (e.g. the mention type pair), string relations (e.g. string match, modifiers match, alias, etc.), semantic information (e.g. gender, number and speaker information as well as ontological information), relative location (distance, apposition or information if the mention is a relative pronoun, or not), learned (those are features that are learned in the process, e.g. anaphoricity and name modifiers predicted match), aligned modifiers (consisting of the relation between the aligned modifiers),

memorization (this is also a learned feature that represents a pair of nouns that is most often used to refer to the given entity as seen in previous examples) and predicted entity types (e.g. entity types match and entity type pair).

Cai et al. [2011] introduce COPA, which is the CR system the authors use for the CoNLL 2011 shared task. COPA represents each of the documents in the data as a hypergraph in which the vertices denote the mentions extracted from the text and the edges denote the relational features between those mentions. The learning in COPA is realized by computing hypergraph weights on the training data[1]. As Cai et al. [2011] report, the features used in the system are grouped in three different classes: negative, positive and weak features. The six negative features model the relations between mentions which do not corefer. Such features are: if the mentions do not agree in number or gender; if the mentions do not agree in their semantic class; if the mentions have the same syntactic heads and the anaphor has a pre-modifier which does not occur in the antecedent and does not contradict the antecedent; if the mentions are first person pronouns that occur in direct speech and are elicited from different speakers; if the two mentions are within the subject and object of the same verb and the anaphor is a non-possessive pronoun. As positive features, the authors extract 10 indicators: nominal/pronominal string match; alias information; head match; if the antecedent is a pronoun and the anaphor is not; in case both are pronouns, if the speaker of the second person pronoun is talking to the speaker of the first person pronoun; if the one of both mentions is the subject of a *speak* verb and other mentions are first person pronouns within direct speech; if the anaphor is a possessive pronoun and the antecedent is the subject of the sentence/subclause; if the mentions are of the same GPE named entity type; if the mentions are of the same Organization named entity type. As weak feature only 3 different indicators are included: if the mentions occur with a word meaning *to say* in a window of 2 words; if the mentions are subjects; if the mentions are synonyms.

With respect to their system that explored both decision trees and logic regression, Björkelund and Nugues [2011] make use of the feature set presented by Soon et al. [2001]. This set consists of 12 features that describe a candidate mention pair: sentence distance, if the antecedent is a pronoun, if the anaphor is a pronoun, string match, NP definiteness, NP demonstrativeness, number/semantic class/gender agreement, if the pair consists of proper names, alias and appositiveness. Björkelund and Nugues [2011] extend the Soon et al. [2001] set with additional features for which information was provided in the CoNLL 2011 dataset, mostly based on the syntactic dependencies included in the annotations.

Uryupina et al. [2011] also use decision trees for the resolution process. The authors present their 42 features in several groups without further details about the actual features. The different classes are as follow: 7 features are used

[1] Cai et al. [2011] used only 30% of the actual training dataset.

to classify the mention type; 8 - for string matching; 2 - for aliasing, 4 - for agreement, 12 - for syntactic information; 3 - to encode salience, 1 - to encode patterns extracted from the Web, 3 - for proximity, and 2 - for 1st and 2nd person pronouns.

The approach presented by Nogueira dos Santos and Lopes Carvalho [2011] makes use of entropy guided transformation learning and decision trees and random forest. Nogueira dos Santos and Lopes Carvalho [2011]'s system calculates 80 different features most of which were already introduced by Ng and Cardie [2002a] or Sapena et al. [2010]. The authors also divide the features in several important groups: lexical (e.g. tokens themselves, string match, length, edit distance, etc.), morphological (e.g. gender/number agreement, if the mentions are proper names, basic gender agreement, etc.), syntactic (e.g. POS information within a given context, predicate information, compatibility of pronouns, embeddedness, etc.), semantic (e.g. baseline system output, head token sense, NE type, semantic role information, speaker and alias information, etc.), distance and position (e.g. sentence distance, mention distance, if both mentions are pronouns - the distance in person names, apposition, etc.).

[Song et al., 2011] use maximum entropy as learning method. As the authors note, the features they employ are commonly used for CR: word features/lexical features, POS information, position within the sentence, semantic role, verb and entity type features, string match, definitiveness, demonstrativeness and pronoun information.

Stoyanov et al. [2011] use the Reconcile [Stoyanov et al., 2010] system for the competition which trains a linear classifier using the averaged perceptron algorithm [Freund and Schapire, 1999]. The authors include 61 different features that have been shown to be successful indicators of coreference on different datasets and tasks. However, no further clarification of the exact features is provided.

Lalitha Devi et al. [2011]'s contribution applies refined salience measure for the pronominal resolution process and conditional random fields for non-pronominal classification. For this reason, the authors separate the feature set in two parts: one for pronominal and one for nominal resolution. The first includes the tokens and their POS in a window of five words and the second follows the widely used feature set from Soon et al. [2001].

Kobdani and Schütze [2011]'s system SUCRE [Kobdani and Schütze, 2010] uses decision trees as the backbone of the learning process. Similar to most approaches presented above, the authors also base the feature set used by the SUCRE system on already well established and tested collections of features for English, such as the one from Bengtson and Roth [2008].

Zhou et al. [2011]'s support vector machine tree kernel also relies on features similar to those of the most widely used set from Soon et al. [2001]. As the authors report, their features are commonly used NLP processes covering: named entity information, semantic role, POS information, the verb and verb

frameset, string match, alias, distance, speaker information, gender/number information, semantic relation and a minimal tree - a partial syntactic analysis of the given mention rooted to the full document.

The system presented in [Charton and Gagnon, 2011] uses a multi-layer perceptron to extract the coreference links again based on 22 well established features: alias, similarity, token and sentence distance, if the mention is a NE, personal pronoun, or a noun phrase, its semantic type, the type of pronoun used, definiteness, demonstrativeness as well as gender and number information.

[Yang et al., 2011] also makes use of maximum entropy classification and the features the authors include in their approach re-use many of the features presented in other approaches above: definiteness and demonstrativeness of the mentions, number/gender information, entity type, is the mention subject/object as well as coordination of the phrase, the type of pronoun used in the mention and existence of prepositions in it, NE type, syntactic information, distance, string match, apposition and copular information, alias, semantic and speaker information.

Xiong et al. [2011] make use of maximum entropy classification for the resolution process. For their participation in the CoNLL 2011 shared task, the authors use the set of features presented in [Soon et al., 2001] that the majority number of systems also employ.

Li et al. [2011] use maximum entropy, integer linear programming and information gain. In their work, Li et al. [2011] implement a knowledge-rich approach including 65 different features from already predefined feature sets (sentence distance, minimum edit distance [Strube et al., 2002]), (string match, partial match, head word match [Daumé and Marcu, 2005]), (gender agreement, number agreement [Soon et al., 2001]), (same head, path [Yang et al., 2006]), (semantic class agreement, predicate [Ponzetto and Strube, 2006, Ng, 2007].

Chen et al. [2011] integrate multiple machine learning methods: maximum entropy, decision trees and support vector machines. The authors also note that the feature set in use consists mainly of the features in [Soon et al., 2001]: distance, if the antecedent or the anaphor are a pronoun, string match, definiteness and demonstrativeness of the anaphor, if the mentions are proper names as well as number, gender, alias and semantic information.

Kummerfeld et al. [2011] presented an unsupervised generative model. As the authors report, their approach makes use of a range of standard features, which are not further presented in their work. However, before the actual coreference classification is performed, pre-resolution filters are applied that constitute three reliable features of spurious mentions: apposition information, attributes signaled by copular verbs and single word mentions with one of the following POS tags: EX, IN, WRB, WP.

The UBIU baseline system used by Zhekova and Kübler [2011], as described in chapter 4.3, relies on memory-based learning for the resolution process.

Two different feature sets were employed: a base feature set, which is a subset of the set presented in [Rahman and Ng, 2009] and an extension set that calculates features for semantic relatedness of the mentions. The feature set described by Rahman and Ng [2009] is also largely a collection of features already existing in the well established and previously used feature sets - [Soon et al., 2001, Ng and Cardie, 2002a, Bengtson and Roth, 2008]. The base set includes features such as: lexical information about the head, if the mentions are pronominal or proper nouns, string match, number and distance information. The extended set considers information about the relation between the heads of both mentions such as: hyponymy, partial holonymy, partial meronymy for nouns and entailment, hypernymy and troponymy for verbs.

The system presented in [Irwin et al., 2011] uses the cluster-ranking algorithm introduced by Rahman and Ng [2009] and follows the description of the [Rahman and Ng, 2009] feature set with respect to the cluster-ranking model. The features provide information if the antecedent is a subject and its NE type, definiteness, demonstrativeness and NE type of the anaphor as well as if it is pronominal or not, distance, string match and schema and cluster-level features.

7.1.1.2 *UBIU's Base Set of Features for the* CoNLL 2011 *Shared Task*

As we shortly mentioned in the previous section, the participation of UBIU within the CoNLL 2011 shared task [Zhekova and Kübler, 2011] was also based on well established features. The overview of the systems participating in the task showed that there is a huge variation with respect to the number of features a system employs as well as the type of information these features may carry. Additionally, it is seldomly reported which effect the different features or a group of features have on the overall system performance. Our system used two different datasets and at this point it is important to show what our findings with respect to this division were.

The base feature set included altogether 14 different features and is listed in table 7.1. The intuition behind this minimalistic selection is to show the performance of the UBIU system at a baseline level. We targeted only English, but our selection was motivated by various reasons supporting the multilinguality and flexibility of application of the set:

- We included easily computable features that are not dependent on a specific layer of annotation apart from POS information (e.g. the mentions themselve, POS information and comparison, string match and distance). This decision enables the application of this set of features on any dataset for which POS information is provided, which is the most wide-spread and easily available annotation type across all languages.

- The base feature set does not include any language specific features (for example, features providing information about the pleonastic *it* in

#	Feature Description
1	m_j - the antecedent
2	m_k - the mention to be resolved
3	Y if m_j is a pronoun; else N
4	number - S(ingular) or P(lural)
5	Y if m_k is a pronoun; else N
6	C if the mentions are the same string; else I
7	C if one mention is a substring of the other; else I
8	C if both mentions are pronominal and are the same string; else I
9	C if the two mentions are both non-pronominal and are the same string; else I
10	C if both mentions are pronominal and are either the same pronoun or different only w.r.t. case; NA if at least one of them is not pronominal; else I
11	C if the mentions agree in number; I if they disagree; NA if the number for one or both mentions cannot be determined
12	C if both mentions are pronouns; I if neither are pronouns; else NA
13	C if both mentions are proper nouns; I if neither are proper nouns; else NA
14	sentence distance between the mentions

Table 7.1: The complete pool of features used as a base feature set including 14 different features. Used for the participation of UBIU in the CoNLL 2011 shared task [Pradhan et al., 2011].

English, which is a phenomenon that may not necessarily occur in every language).

The reported performance of the UBIU system on the development set of the CoNLL 2011 shared task when only the base feature set was made use of is shown in table 7.2. This performance shows that the base feature set reaches baseline performance with an averaged F-score of 43.01 over the three metrics considered by the shared task (MUC, B^3 and CEAFE).

The main motivation for a baseline feature set is that the latter can be universally employed for any dataset and language. Whenever needed this set can be further extended in a language dependent manner in order to improve

IM			MUC			B^3			CEAFE			Av.
R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	F_1
62.71	38.66	47.83	30.59	24.65	27.30	67.06	62.65	64.78	34.19	40.16	36.94	43.01

Table 7.2: UBIU's results achieved on the CoNLL 2011 shared task development set from the employment of the base feature set.

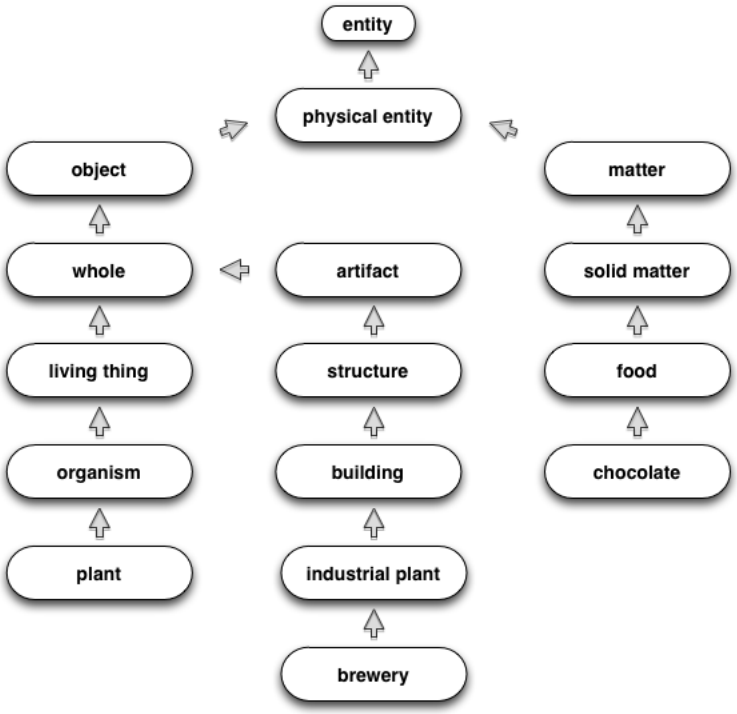

Figure 7.1: An example of the hyponymy relation in WordNet representing the tree for three different concepts: plant, brewery and chocolate.

the system performance. This was our main goal when we assembled an extension set of features that aimed to provide semantic information about the mention pair.

The CoNLL 2011 shared task allowed for the use and integration of WordNet[2] version 3.0 as part of the closed task. Being a large lexical database and ontology that includes conceptual-semantic and lexical relations, WordNet provides information about many different lexical relations that can be helpful during the coreference resolution process. Such a relation is, for example, hyponymy (denoting the inclusion of the semantic field of a token/phrase within that of another token/phrase: e.g. cheese is a hyponym of food, cat is a hyponym of animal, etc.), which is also graphically represented in figure 7.1.

All additional features that we extracted from WordNet are listed in table 7.3. Note, that we also include features representing verb relations - features #21-#26 in the table. This is so because UBIU also targeted verb coreference within

[2] http://wordnet.princeton.edu

#	Feature Description
15	C if both are nouns and m_k is hyponym of m_j; I if both are nouns but m_k is not a hyponym of m_j; NA otherwise
16	C if both are nouns and m_j is hyponym of m_k; I if both are nouns but m_j is not a hyponym of m_k; NA otherwise
17	C if both are nouns and m_k is a partial holonym of m_j; I if both are nouns but m_k is not a partial holonym of m_j; NA otherwise
18	C if both are nouns and m_j is a partial holonym of m_k; I if both are nouns but m_j is not a partial holonym of m_k; NA otherwise
19	C if both are nouns and m_k is a partial meronym of m_j; I if both are nouns but m_k is not a partial meronym of m_j; NA otherwise
20	C if both are nouns and m_j is a partial meronym of m_k; I if both are nouns but m_j is not a partial meronym of m_k; NA otherwise
21	C if both are verbs and m_k entails m_j; I if both are verbs but m_k does not entail m_j; NA otherwise
22	C if both are verbs and m_j entails m_k; I if both are verbs but m_j does not entail m_k; NA otherwise
23	C if both are verbs and m_k is a hypernym of m_j; I if both are verbs but m_k is not a hypernym of m_j; NA otherwise
24	C if both are verbs and m_j is a hypernym of m_k; I if both are verbs but m_j is not a hypernym of m_k; NA otherwise
25	C if both are verbs and m_k is a troponym of m_j; I if both are verbs but m_k is not a troponym of m_j; NA otherwise
26	C if both are verbs and m_j is a troponym of m_k; I if both are verbs but m_j is not a troponym of m_k; NA otherwise

Table 7.3: The supplemental features carrying semantic information about the mention pair that we extracted from WordNet version 3.0 and used as an addition to the base feature set in table 7.1.

	IM			MUC			B^3			CEAFE			Av.
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	F_1
B	62.71	38.66	47.83	30.59	24.65	27.30	67.06	62.65	64.78	34.19	40.16	36.94	43.01
E	62.72	39.09	48.16	30.63	24.94	27.49	66.72	62.76	64.68	34.19	39.90	36.82	43.00

Table 7.4: A comparison of the results achieved by the employment of the base (B) and the extended (E) feature sets in the CoNLL 2011 shared task.

the CoNLL 2011 shared task. However, the latter fact is not of much interest to the current discussion, since we want to concentrate more on the actual feature set UBIU used in the CoNLL 2011 shared task with respect to nominal coreference.

The important and interesting part of our observations with respect to system performance when this group of semantic features was added to the base features is the fact that contrary to our expectations, using extended semantic information does not improve the overall system performance. The system performance is shown in table 7.4 as a comparison to the performance achieved with the base feature set. Only on the mention level, we see a minimal gain in precision. But this does not translate into any improvement on the coreference level.

In general, the assumption is that deeper linguistic analysis leads to an improved capability of the system to identify correctly the coreference relations in the data. What our comparison showed is that this is not always the case. However, the latter might be because of several different reasons:

- MBL is known to be sensitive to large feature sets with less indicative features. This means that the semantic information might be more helpful in an MBL approach if it is represented by less, but more informative features.
- The supplementary feature set that we considered included also features for verb coreference, which might lead to additional noise for the learner.
- WordNet is a big lexical database, but unfortunately not big enough for tasks of this scale. It does not manage to provide the needed coverage, meaning that it does not include all tokens in the data and even less a complete set of the relations between the tokens we search for. This also leads to less informative or even misleading features.

The best performing system in the CoNLL 2011 shared task, [Lee et al., 2011], also reported that their semantic sieve that included information from WordNet, Wikipedia infoboxes and Freebase records had a detrimental effect on the overall performance of the system and was respectively not used in the final system participation. The authors assume that this unexpected performance could be changed via a different tuning for the sieve parameters. Unfortunately, other approaches, such as [Sapena et al., 2011] that use WordNet features, do not report the effect of the semantic features on the overall system performance for the CoNLL 2011 shared task dataset. This makes it harder to draw further conclusions on the use of WordNet for the task.

The fact that we discuss a set of features extracted from WordNet for coreference resolution with respect to only one language, namely English, is not a coincidence. Overall, semantic information is believed to increase the system performance for all languages and not only for English. However, ontologies that can provide information about the semantic relation for each of the mention pairs is not always easily available for a wide range of languages. Our comparison of the base and extended feature sets showed that even WordNet is not sufficient for English, which is the language best supported

by the project. The latter fact was confirmed by the implementation of the successor of the CoNLL 2011 shared task, namely the CoNLL 2012 shared task, for which Pradhan et al. [2012] reported that as a result of the lack of resources and state-of-the-art tools, as well as time constraints, some layers of information for the Chinese and Arabic portion of the data could not be provided. Additionally, the use of WordNet within the closed track of the task was also allowed only for the English language, since similar resources for Arabic and Chinese do not provide comparable coverage.

7.1.2 *Discussion*

The general overview in the preceding section shows that almost all state-of-the-art machine learning approaches to coreference resolution for English rely on approximately identical feature sets such as the ones presented by Soon et al. [2001], Ng and Cardie [2002a], Bengtson and Roth [2008]. Systems vary to a great extent in the number of features they include in their sets ranging between the 12 initial features of the Soon et al. [2001] set (e.g. [Lalitha Devi et al., 2011, Zhou et al., 2011]) and the implementation of over 100 different features as presented in [Sapena et al., 2011]. An interesting fact to note is that there is no specific correlation between the number of features a system uses and its overall coreference performance. Often, even if a given feature is implemented within a given system, the form that is selected to represent its values might not be very informative to the learning algorithm. For this reason, it is highly important that features and their value representation are thoroughly investigated and evaluated with respect to the learning method they are employed with. Altogether, the latter is not a complex but rather highly time consuming task, which leads to the fact that such evaluations are seldom carried out and reported within state-of-the-art work. This is confirmed by the approaches we reviewed in the previous section. Moreover, these approaches targeted only a single language. Increasing the number of languages included in the CR system leads to a drastic increase in the effort needed to evaluate the separate features with respect to each language and machine learning approach.

The overall lack of feature evaluation as well as the problem posed by multilinguality on this subtask of coreference resolution motivates our further explorations on the topic of feature selection within multilingual coreference resolution.

7.2 FEATURES AND THEIR EFFECT WITHIN THE CONLL 2012 DATASETS

This section will address the issue of multilingual feature selection and evaluation. We will employ the UBIU system in its implementation for the partic-

#	POS-based	Type	Feature Description
1	√	Lexical	m_j - the antecedent
2	√		m_k - the mention to be resolved
3	√		C if both mentions are the same string; else I
4	√		C if one mention is a substring of the other; else I
5	√		C if both mentions are pronominal and are the same string; else I
6	√		C if both are non-pronominal and are the same string; else I
7	√	Grammatical	C if m_j is a pronoun; else I
8	√	NP type	C if m_k is a pronoun; else I
9	√		the concatenated values of feature 7 and feature 8
10	√		C if both are pronouns; I if neither is a pronoun; else U
11	√		C if both are proper nouns; I if neither is; else U
12	-		D if m_j is in a definite mention; I otherwise
13	-		PR if m_j is premodified, PO if it is postmodified; UN otherwise
14	-		PR if m_k is premodified, PO if it is postmodified; UN otherwise
15	-		the concatenated values for feature 13 and 14
16	-	Grammatical	C if m_j is within the subject; I-within an object; U otherwise
17	-	function	C if m_k is within the subject; I-within an object; U otherwise
18	-		C if both are within ARG0-ARG4; I-within ARGM; else U
19	-		C if m_j is within ARG0-ARG4; I-within ARGM; else U
20	-		C if m_k is within ARG0-ARG4; I-within ARGM; else U
21	-		concatenated values for features 19 and 20
22	-		the predicate argument label for m_j
23	-		the predicate argument label for m_k
24	-	Grammatical	C if neither is embedded in a PP; I otherwise
25	-	heuristic	C if neither is embedded in a NP; I otherwise
26	√	Grammatical	C if both mentions agree in number; else I
27	-	agreement	C if both mentions agree in gender; else I
28	-	Semantic	C if both mentions have the same speaker; I if they do not
29	-		C if both mentions are the same named entity; I if they are not and U if they are not assigned a named entity
30	√	Positional	token distance between m_j and m_k
31	√		sentence distance between m_j and m_k
32	√	Other	normalized levenstein distance for both mentions
33	-		C if m_j has been classified as singleton; I otherwise

Table 7.5: The features used by the coreference classifier of the UBIU system within its participation in the CoNLL 2012 shared task. # lists the ID for each feature, column *POS-based* shows if the feature is (√) or is not (-) part of the POS-based feature set, column *Type* represents the separation of the feature group types and column *Feature Description* lists selection of values as well as description of the feature.

ipation at the CoNLL 2012 shared task [Zhekova et al., 2012] as well as the datasets provided by the task in order to evaluate the features that the system uses. We include the full feature set that was considered within the CoNLL 2012 shared task, which is also listed in table 7.5. We separate the features

of the full feature set in several aforementioned categories: Lexical features, Grammatical features describing the NP type, Grammatical features describing the function of the mention, Grammatical features extracted via heuristic rules, Grammatical features for mention agreement, Semantic features, Positional features and Other. These group types are given in table 7.5 in column *Type*. The column *POS-based* of the table denotes if the given feature is considered as being part of the POS-based feature set ($\surd$) or not (-).

The current section provides a thorough investigation of the performance of the POS-based feature set and the full feature set (section 7.3). Following, in section 7.4 we list a series of experiments in which we test the effect of each of the feature groups from table 7.5 to the overall system performance by removing this group from the full feature set. We show detailed system scores and offer a comprehensive discussion for each of the targeted groups.

7.3 EVALUATION OF PERFORMANCE OF THE POS-BASED AND FULL FEATURE SETS

7.3.1 *Full Feature Set*

The full feature set includes all features that we employed during our participation within the CoNLL 2012 shared task. We include all easily computable features plus additional ones that require deeper analysis, such as semantic or syntactic analysis of the data. We use various annotation layers provided by the shared task, which enriches the overall system knowledge with respect to the coreference problem. Our investigation will include a separate experiment for each of the existing feature groups in which we will remove the given group from the full feature set.

The results achieved by the MCR system on the CoNLL 2012 shared task datasets with the use of the full feature set are listed in table 7.6. All intermediate scores are reported from all evaluation metrics used by the scoring software provided by the task. Additionally, a TOTAL score is computed that is an average of the F-measures from the various evaluation metrics.

According to the figures presented in table 7.6 we can make several important observations. Several aspects will be discussed for each of the achieved intermediate results further on in the chapter in order to obtain an objective comparison between the various groups of features in the implementation of the UBIU multilingual coreference resolution system. We include the following characteristics into our consideration: overall performance – general comments about the performance of the feature set; recall vs. precision[3] – analyses of the difference in scores with respect to the changes of recall and precision. This aims at investigating if some groups tend to improve/decrease system performance only with respect to one of the metrics or with respect to both at

[3]This aspect is superfluous in this section, because we analyze the full feature set itself.

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
GM	MD	100	100	100	100	100	100	100	100	100
	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.70			54.12
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39

Table 7.6: The results achieved by the UBIU coreference resolution system on the CoNLL 2012 shared task datasets with the use of the full feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). We report the scores in terms of (P)recision, (R)ecall and (F)-measure from all evaluation metrics and TOTAL denotes their averaged F-measures.

the same time; cross-lingual differences – which aims to investigate the effect of the group of features on a cross-lingual level by a direct comparison among the results of all three languages; mention detection[3] – will discuss the changes that can be observed for mention detection depending on the selection of the feature set again with respect to the scores achieved by the full set of features.

overall performance – The full feature set leads to an overall performance that corresponds to the participation of the UBIU system within the CoNLL 2012 shared task. It ranges from a TOTAL score of 58.64 for English when *gold* mentions (GM) are used to 37.99 for Arabic when mentions are detected automatically (AM). As expected, the use of *gold* mentions leads to the best overall scores, followed by the employment of *gold* boundaries (GB) and finally, the worst performance is observed

when mentions are detected automatically. Also worthwhile mentioning is the fact that the decrease of the TOTAL scores across the three different settings (GM, GB, AM) is seen for all evaluation metrics, but the largest gap in general is reported by MUC. One explanation for this observation can be found in connection to the sensitivity of the MUC metric to the variation of fully detected coreference chains. The use of *gold* mentions is not as complex a task as the use of *auto* mentions for any coreference system. The GM set contains less mentions altogether and all the mentions in it have the correct mention boundaries and spans. This means that the CR system in use only needs to identify the correct chains. Increasing the number of mentions, namely using the GB set, introduces all phrases (including the singletons), which leads to a higher task complexity, directly mirrored by the lower scores for this setting. When UBIU uses the AM set of mentions, the difficulty is increased by the errors introduced by the mention detection procedure, such as wrong mention boundaries, superfluous mentions, etc. which we discussed thoroughly in chapter 5. As a result of the increased difficulty, the system does not manage to create complete coreference chains. The coreference links that are identified build rather shorter and incomplete clusters, which has highest significance for the MUC metric. This is one of the biggest drawbacks of this evaluation metric, which we described in section 2.3.4.2.

cross-lingual differences – It is interesting to note that within almost all settings (apart from GB for Arabic), the large differences across the TOTAL scores of all languages within a given setting are mainly due to the big gaps between the performance reported by the MUC metric. This fact indicates that the MUC metric also has a higher cross-lingual sensitivity than the rest of the evaluation metrics. We assume that the latter is again caused by MUC's higher scores for more complete chains or overmerged entities. Arabic has a significantly larger number of mentions in comparison to both English and Chinese, which makes it harder for the system to identify all coreference links for a given entity. Therefore, MUC's score for this language is correspondingly a lot lower than the scores for English or Chinese.

The results listed in table 7.6 are further included (highlighted in grey), for better clarity and comparison, in the tables presenting the performance of the various feature groups.

7.3.2 *POS-based Feature Set*

The motivation behind the POS-based feature set is relatively straightforward. In it we include only features that can be solely derived from the text itself and the POS annotation layer. Since POS information is generally widely available for

a large number of languages, we assume that this will be the minimal setting a system would be presented with. This allows us to evaluate a minimalistic feature set on three typologically very different languages, such as Arabic, Chinese and English. Such an investigation will show if a general expectation about a comparable performance of the system for any given language can be anticipated. A good system performance achieved on the POS-based feature set will additionally motivate the further exploration of information-poor multilingual approaches to coreference resolution. The set contains all the features marked with $\surd$ in column *POS-based* in table 7.5.

The results achieved by the UBIU coreference resolution system on the CoNLL 2012 shared task datasets with the use of the POS-based feature set are listed in table 7.7. The performance is again evaluated with respect to various aspects:

overall performance – The overall performance of the POS-based feature set is lower than the performance achieved via the use of the full set. For Arabic we note an average decrease of 2.72 percent points across all three settings. For English, the decline is 3.27 percent points and for Chinese – 0.70 percent points. However, the POS-based set solely consists of easily computable features that do not require deeper linguistic analysis of the data and thus fewer provided annotation layers. Altogether, the performance of the POS-based set shows that the trade-off between annotation effort and performance ranges in a decrease in performance from the full set between 0.45 percent points for Chinese in the *gold* boundaries (GB) setting and and 3.88 percent points for English in the *gold* mentions (GM) setting. This indicates that information-poor approaches can be used with an approximate decrease of performance of 2 percent points as calculated by the results across the three languages we target and all three evaluation settings. Depending on the application in which multilingual CR is used, this trade-off could be highly acceptable regarding the immense effort needed to provide training data with a wide range of linguistic annotations. The latter is an important finding of our investigation, since one of our main aims is to develop an approach that is easily adaptable and applicable to all languages. As a multilingual system, UBIU does not strive for best performance, but rather competitive overall performance that can be achieved for any given language and dataset with a minimum amount of linguistic annotations and annotation effort. With the use of the POS-based feature set, this is made easier without an unacceptable decrease of system performance.

recall vs. precision – With respect to the precision and recall that are reported by the evaluation metrics in comparison to the full feature set, there is again an interesting observation that is indicated by the figures in table 7.7. It is not always the case that a lower F-score, caused by the

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
GM	MD	100	100	100	100	100	100	100	100	100
full	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GM	MD	100	100	100	100	100	100	100	100	100
POS-based	MUC	34.85	79.12	48.39	71.16	77.44	74.17	52.91	76.97	62.71
	B^3	43.27	91.07	58.67	69.54	51.02	58.86	57.82	69.84	63.27
	$CEAF_M$	48.25	48.26	48.25	46.37	46.37	46.37	47.34	47.34	47.34
	$CEAF_E$	68.32	28.22	39.94	47.21	37.47	41.78	55.73	30.42	39.36
	BLANC	56.38	80.81	59.16	66.47	56.56	52.63	68.50	58.32	60.24
	TOTAL			50.88			54.76			54.58
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
full	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.70			54.12
GB	MD	31.57	100	47.99	68.35	100	81.20	50.77	100	67.35
POS-based	MUC	14.48	45.45	21.97	60.47	77.53	67.94	44.29	76.62	56.13
	B^3	32.71	84.23	47.12	60.64	60.89	60.77	52.28	76.93	62.26
	$CEAF_M$	34.70	34.70	34.70	47.01	47.02	47.02	47.35	47.35	47.35
	$CEAF_E$	55.98	20.48	29.99	53.50	31.34	39.52	57.93	27.29	37.10
	BLANC	51.93	57.91	52.00	67.49	57.94	57.36	68.89	63.41	65.51
	TOTAL			37.16			54.52			53.67
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
full	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39
AM	MD	10.28	71.48	17.98	57.22	59.06	58.13	43.36	65.38	52.14
POS-based	MUC	6.59	54.13	11.75	44.83	43.28	44.04	34.80	49.98	41.03
	B^3	33.05	95.87	49.15	61.79	53.71	57.47	54.71	72.17	62.24
	$CEAF_M$	32.72	32.72	32.72	38.68	38.68	38.68	43.67	43.67	43.67
	$CEAF_E$	54.15	18.35	27.41	31.04	32.44	31.72	42.06	29.23	34.49
	BLANC	51.01	71.63	50.13	61.59	54.12	53.94	62.78	62.39	62.58
	TOTAL			34.23			45.17			48.80

Table 7.7: UBIU's results on the CoNLL 2012 datasets with the use of the POS-based feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). We report the scores in terms of (P)recision, (R)ecall and (F)-measure from all evaluation metrics. TOTAL denotes their averaged F-measures. In grey the performance of the full feature set is listed.

use of the POS-based feature set, is formed by both lower recall and lower precision. The fact that both variations are present (for example, for Arabic in the GM setting – MUC, B^3 and BLANC report lower recall but higher precision; in other cases, as for English in the *auto* mentions (AM) setting, the same metrics show increase in recall, but decrease in precision) makes it harder to draw a conclusion about the actual effect of the change in the feature set. Moreover, because the evaluation metrics have a highly different nature, their results often disagree in the change of precision vs. recall. For our investigation, this change is important, because often enough there is a preference for either a higher precision or a higher recall depending on the NLP application the system is used in. However, for the lack of agreement between the metrics no meaningful conclusions can be drawn.

cross-lingual differences – Similar to the observations we had on the performance of the full set, the MUC metric is most sensitive to the changes of the resolution approach. In general, the overall tendencies of TOTAL system performance across all three languages remain the same with respect to the decrease of scores for all GM, GB and AM settings. Yet, we can find an interesting occurrence. Across the three settings an average of 2.72 percent points of decrease is observed for Arabic, 3.27 is the decrease for English and 0.70 percent points for Chinese. These figures show that the POS-based set has a different effect on the system performance depending on the language. While the deviation is not overly large, it shows that the lack of more linguistically informative features is more harmful to some languages, as English in this case, and less to others, as Chinese. From the results within this evaluation setting, it is not clear to what degree the various features that were left out (excluded from the full feature set) affect the language-specific performance. Even though we achieved a highly competitive performance with a minimalistic feature set, such as the POS-based set, we consider it important to know which group of features carries most indicative information for the MBL classifier. For this reason, we offer a more detailed analysis of the various types of features in section 7.4.

mention detection – With respect to MD, there are changes only for the GB and AM settings. This happens, since when *gold* mentions are used the system does not remove any singletons from the set of mentions after the resolution process (see chapter 5). The use of *gold* boundaries and the POS-based feature set leads to a decrease of 5.82 percent points for MD recall for Arabic, while for both English (with 1.41 percent points) and Chinese (with 0.73 percent points) the recall increases. Precision is stable for this setting, since this mirrors the *gold* annotations of the boundaries/spans of the used mentions. For the *auto* mentions (AM)

setting, as we previously introduced, neither the boundaries nor the complete set of mentions is defined by the annotations. Within that setting both precision and recall as well as the calculated F-measures for MD are lower, apart from the slight improvement in precision (0.93 percent points) for Arabic. The improvement in mention detection is important to us, because a better performance means that the system can better identify the actually coreferent mentions. Additionally, from the TOTAL score, we can conclude that for the GB setting for English and Chinese, UBIU better identifies the set of singletons, but does not improve on the correct coreference links between the set of coreferent mentions. Thus the MD scores are higher, but the TOTAL figures lower. However, this observation is not valid for the AM setting (the most realistic and objective setting, because the system needs to implement the full pipeline), which does not confirm any specific effect of the POS-based set on the mention detection capabilities of the multilingual coreference resolution procedure. This finding has an overall positive meaning for our work, because the POS-based feature set includes very simplistic and easily computable features that do not harm the capability of the system to identify the set of mentions participating in the coreference chains.

7.4 EVALUATION OF PERFORMANCE LOSS PER FEATURE GROUP

The comparison between the full and the POS-based feature sets that we presented in section 7.3 showed that the POS-based feature set leads to a very moderate decrease in system performance and that it could be used even when only POS information is provided as an annotation layer of the given dataset. However, our investigation did not lead to any further conclusive remarks about the effect of the various features on the system performance. Furthermore, we showed, that the tendencies vary across the languages, which means that a deeper analysis of feature selection is needed to carry out a better categorization of the various feature groups and their effect on the separate languages. Such an evaluation would provide a deeper knowledge on the informativeness of the features or feature types, which can be used by multilingual systems that aim to gain an optimal language dependent performance.

For this reason, the current section presents a distinct experimental setting for each of the separate groups of features listed in table 7.5. Each setting uses the full feature set from which the given group of features is excluded in order to determine its informativeness for the CR pipeline. Section 7.4.1 discusses the group of *lexical features*, section 7.4.2 observes the effect of the *grammatical NP type* features on the system performance. Following (in section 7.4.3), we show the performance when the set of *grammatical function* features is excluded from the full set of features, section 7.4.4 covers *grammatical heuristic*

features, section 7.4.5 – *grammatical agreement* features, section 7.4.6 discusses the *semantic feature* set, in section 7.4.7 we show the effect of the small set of *positional features* and finally in section 7.4.8 we present our observations when the class of features listed as *other* in table 7.5 is removed.

7.4.1 *Lexical*

The current section will investigate the effect of the lexical group of features. These are features #1 – #6 from table 7.5, which include the tokens of the anaphor and the antecedent, as well as string-match information about these tokens. Lexical features are an essential part of each state-of-the-art coreference resolution system, since they can always be easily computed from the underlying text in the datasets. We will evaluate the informativeness of this set of features by removing it from the full set of features and observing the change in overall system performance. For convenience, we will further call this set Full minus Lexical (F-L) feature set. The rest of the system settings and the data that is used are kept the same.

The results that the UBIU multilingual coreference resolution system achieves with the F-L set are listed in table 7.8. Again, we report the figures in the same manner as for the full and POS-based feature sets and keep the conventions that we previously introduced. The performance of the MCR system using the full feature set is also provided for better comparison (highlighted in grey in table 7.8). Similar to our discussion in the previous section, we will consider the following aspects within the analysis of the results presented in table 7.8: overall performance, recall vs. precision, cross-lingual differences and mention detection. This will allow us to observe the effect of this type of features from different perspectives and gain a more objective general evaluation.

overall performance – Regarding the overall evaluation we can again note interesting changes in TOTAL scores. There is an overall increase in the scores of 0.22 percent points for Arabic as well as a decrease of 2.47 percent points for English and 1.83 percent points for Chinese. While the POS-based feature set showed a detrimental effect on the system performance, the use of the F-L set, shows slightly different tendencies. It leads to a decrease in performance for English and Chinese, while for Arabic the scores improve. This shows that lexical features for Arabic are more noisy rather than informative for the memory-based learner, because their exclusion from the full feature set leads to better TOTAL system performance. The latter is presumably caused by the morphological richness of the Arabic language and the fact that every root in classical Arabic may lead to a large number of different word forms and senses. The detrimental effect of the F-L group of features for Arabic is an important finding, since lexical features are a well-established part of

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
GM	MD	100	100	100	100	100	100	100	100	100
full	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GM	MD	100	100	100	100	100	100	100	100	100
F-L	MUC	38.26	77.53	51.23	65.11	76.22	70.23	47.68	74.36	58.10
	B^3	44.85	88.83	59.60	59.87	61.92	60.88	53.24	72.95	61.56
	$CEAF_M$	49.47	49.48	49.47	50.55	50.55	50.55	47.52	47.52	47.52
	$CEAF_E$	68.18	29.80	41.48	53.46	36.41	43.32	57.33	29.35	38.82
	BLANC	57.47	78.82	60.64	67.61	62.21	63.96	67.77	61.77	63.90
	TOTAL			52.48			57.79			53.98
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
full	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.70			54.12
GB	MD	38.66	100	55.76	63.52	100	77.69	47.50	100	64.41
F-L	MUC	20.03	49.94	28.59	53.65	74.90	62.52	41.10	75.52	53.23
	B^3	34.98	80.25	48.72	51.33	67.58	58.35	49.15	78.51	60.46
	$CEAF_M$	37.17	37.17	37.17	46.04	46.04	46.04	45.88	45.88	45.88
	$CEAF_E$	56.65	22.44	32.15	53.54	28.01	36.78	57.73	26.09	35.94
	BLANC	54.01	59.18	55.00	64.70	61.44	62.68	66.48	64.69	65.52
	TOTAL			40.33			53.27			52.21
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
full	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39
AM	MD	18.81	71.88	29.82	56.22	63.12	59.47	42.88	71.53	53.62
F-L	MUC	12.43	52.66	20.11	41.65	44.55	43.05	34.00	51.91	41.09
	B^3	36.71	91.56	52.41	55.72	59.69	57.64	51.37	72.25	60.05
	$CEAF_M$	35.86	35.86	35.86	39.13	39.14	39.14	42.12	42.12	42.12
	$CEAF_E$	52.68	20.72	29.75	32.70	30.06	31.32	42.77	27.33	33.35
	BLANC	52.66	72.07	53.32	58.77	57.02	57.73	60.44	61.14	60.78
	TOTAL			38.29			45.78			47.48

Table 7.8: UBIU's results on the CoNLL 2012 datasets with the use of the F-L feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). We report the scores in terms of (P)recision, (R)ecall and (F)-measure from all evaluation metrics. TOTAL denotes their averaged F-measures. In grey the performance of the full feature set is listed.

all state-of-the-art feature sets for coreference resolution (see section 7.1). Moreover, previous work in the field of coreference resolution [Björkelund and Nugues, 2011] indicates that lexical features carry information that is highly valuable to the CR resolver. Yet, the authors based their study on the CoNLL 2011 shared task dataset, which included only English data and thus there is no indication about the helpfulness of this set of features on a multilingual level. Our results show that lexical features may be informative for some languages (e.g. English and Chinese), but not for all (e.g. Arabic). We assume that languages that are highly morphologically rich, as Arabic, will also lead to similar performance.

recall vs. precision – There are no particular cases that lead to abnormal scores for precision or recall which shows that the F-L feature group does not lead to any specific behaviour of the system.

cross-lingual differences – For Arabic, the F-L set leads to higher TOTAL scores across all three evaluation settings (GM – increase of 0.19 percent points, GB – 0.17 percent points and AM – 0.30). This tendency is not kept for English and Chinese, for which the TOTAL scores are lower across all evaluation settings. Similar to our observations on the POS-based set, English shows highest deviations from the performance of the full feature set. As a whole, the cross-lingual comparison of the results shows, that the F-L feature set, and respectively all other groups of features, should be reevaluated for each separate language and system setting in order for the actual informativeness of the features in the given context to be determined. This is also an important overall finding of our investigation, because it shows that only information-poor approaches can achieve language independent behaviour. Information-rich approaches and especially those that aim at optimal system performance for all targeted languages need to be optimized via a language dependent feature selection that depicts the proper groups of features per given language.

mention detection – When we look at the results for mention detection we can again analyze the differences between GB and AM for the two feature sets. The results in table 7.8 indicate another issue interesting to us. The overall better coreference performance for Arabic is also mirrored in the mention detection figures for this language when the GB scores are considered, leading to an improvement from an F-measure of 54.43% to 55.76% (gaining 1.33 percent points) and an improvement from an F-measure of 27.83% to 29.82% (gaining 1.99 percent points) for the AM setting. For English, MD decreases with 2.59 percent points for the GB setting and with 1.53 percent points for the AM setting. Similar to English, Chinese also shows a detrimental effect of the feature set on mention detection with respect to both settings – 2.29 percent points lower for GB and 1.57 percent points lower for AM. This fact can be an indicator

that mention detection is tightly correlated to coreference resolution and that the performance of both depend on each other. The fact that the MCR system better identifies the coreference links allows for a more accurate removal of singletons from the final output. However, this relation does not always hold, as can be seen for example for English and Chinese and the GB setting for the POS-based feature set. Thus, we will continue our observation of this correlation for all feature sets we are further on going to explore in order to be able to reach a more conclusive judgment about the topic.

7.4.2 *Grammatical NP type*

The next set of features that we would like to examine is the full set without the set marked as grammatical NP type in table 7.5 (further referred to as F-GNPt), which includes features #7 through #15 from the table. These are features that determine the type of noun phrase that both the antecedent and the anaphor are – if they are pronominal, proper nouns, definite, pre-modified or post-modified, or not modified at all, etc. Grammatical NP type features are also a substantial part of state-of-the-art feature sets for coreference resolution. They are easily computable and could be achieved solely by the POS annotation layer. For this reason, grammatical NP type features are also partially included in our POS-based set of features. However, in the POS-based set we do not consider features #12 through #15, because they require language-specific knowledge and adaptation, which is against our motivation for an easily implementable and language independent POS-based feature set.

In table 7.9, we list the detailed scores from all evaluation metrics for the evaluation setting of the F-GNPt set. The results should determine whether or not this group is informative for the multilingual coreference resolver and to what degree it is helpful across the three languages that we target. The figures in the table again show information that is valuable for our investigation. We examine the following aspects:

overall performance – The overall performance of UBIU is substantially affected by the loss of the features that contain information about the grammatical type of the anaphor and the antecedent. The scores vary considerably with a decrease of 1.71 percent points for Arabic and an increase of 3.08 percent points for English and 3.65 percent points for Chinese. These figures show that the F-GNPt feature set is highly important for the coreference resolver and could lead to a higher amplitude deviation of the scores in comparison to the performance achieved by the full feature set. What is specifically surprising here, is the fact that for English and Chinese the results improve with the use of the F-GNPt feature set. This was not what we had expected, as this group of features is generally believed to be very important for coreference resolution and

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
GM	MD	100	100	100	100	100	100	100	100	100
full	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GM	MD	100	100	100	100	100	100	100	100	100
F-GNPt	MUC	33.55	75.37	46.43	64.42	80.52	71.58	49.05	80.67	61.01
	B^3	42.16	88.98	57.22	59.61	72.54	65.45	53.90	82.48	65.19
	$CEAF_M$	46.74	46.76	46.75	56.86	56.87	56.86	53.61	53.61	53.61
	$CEAF_E$	66.11	27.48	38.82	62.84	38.30	47.60	63.67	31.18	41.86
	BLANC	56.81	77.62	59.67	70.92	67.88	69.20	69.89	69.58	69.73
	TOTAL			49.78			62.14			58.28
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
full	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.70			54.12
GB	MD	35.19	100	52.06	68.94	100	81.62	53.93	100	70.07
F-GNPt	MUC	18.05	49.94	26.52	58.41	81.12	67.92	47.37	82.27	60.12
	B^3	34.45	82.80	48.66	53.97	77.21	63.53	52.42	84.62	64.73
	$CEAF_M$	36.71	36.71	36.71	54.76	54.76	54.76	53.07	53.07	53.07
	$CEAF_E$	56.44	21.51	31.15	64.59	34.00	44.55	64.14	30.13	41.00
	BLANC	53.60	62.08	54.58	69.37	68.30	68.81	69.33	70.62	69.95
	TOTAL			39.52			59.91			57.77
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
full	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39
AM	MD	13.05	71.21	22.06	63.38	62.91	63.14	48.83	68.63	57.06
F-GNPt	MUC	9.07	54.54	15.55	49.73	49.48	49.61	40.80	56.16	47.26
	B^3	34.54	94.16	50.55	61.37	62.41	61.88	56.95	74.90	64.71
	$CEAF_M$	33.90	33.90	33.90	45.72	45.72	45.72	47.76	47.76	47.76
	$CEAF_E$	53.16	19.07	28.07	36.37	36.61	36.49	44.63	31.87	37.18
	BLANC	51.97	73.49	52.02	64.86	62.53	63.56	65.43	66.94	66.15
	TOTAL			36.02			51.45			52.61

Table 7.9: UBIU's results on the CoNLL 2012 datasets with the use of the F-GNPt feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). We report the scores in terms of (P)recision, (R)ecall and (F)-measure from all evaluation metrics. TOTAL denotes their averaged F-measures. In grey the performance of the full feature set is listed.

is thus often included in state-of-the-art feature sets. Another curious observation is the big difference between the scores for Arabic. While the use of the F-GNPt set leads to a decrease of scores of 2.51 percent points in the GM setting, for the GB setting the reduction is only 0.64 percent points. The surprising results are hard to explain since all of the features in this particular set are linguistically well motivated and carry information that is important for the problem. One possible reason would be the fact that features #7 through #11 are redundant for the memory-based classifier, since this information is contained in the lexical features (as the tokens for the anaphor and the antecedent) and languages such as English and Chinese that are not as morphologically rich as Arabic do not need explicit features carrying this information. MBL is known to be very sensitive to less informative features and thus morphologically rich languages (e.g. Arabic) that have a large number of lexical surface forms do not provide very informative lexical features (which we also showed in the F-L section for this language) and can make better use of the features included in the F-GNPt feature set.

This information once again confirms our observations that the informativeness of the separate sets of features are closely related to the given evaluation setting, type of system, language selection and learning algorithm.

recall vs. precision – The variation between the decrease in the GM and GB setting for Arabic is also intriguing in regard to the recall and precision figures reported by the evaluation metrics. For the GM setting, the decrease is confirmed by all evaluation metrics used and via both precision and recall figures. This gives a clear and mutually confirmed indication that within this setting the F-GNPt set leads to a decrease in system performance across the metrics and for both precision and recall. This agreement is not as present in the GB and AM settings in which the CEAF variants both report lower precision and recall, but the rest of the metrics indicate rather lower recall and higher precision. The latter finding is significant for our work, because the F-GNPt feature set seems to carry important information mostly with respect to recall, thus further use of the F-GNPt set for Arabic could be targeted when there is a preference for higher precision than recall.

cross-lingual differences – The cross-lingual comparison of the figures in table 7.9 was already discussed in the overall performance part of this section. We depicted the division of languages on the basis of their morphological richness. This is an important cross-lingual observation, which is linguistically motivated and corresponds to the general expectations accompanied with the capabilities every MBL learner offers. These findings should be further confirmed across a wider range of languages

and language families, but they constitute an important part of our investigation.

mention detection – With respect to mention detection, we can observe a predictable performance, namely that a correlation exists between the results achieved by the system for both mention detection and coreference resolution. The figures in table 7.9 show that for this feature set an increase in MD scores translates into an increase in CR performance. As we already discussed in the previous section, this is not always the case and there are no general tendencies that can be observed in the system performance.

7.4.3 *Grammatical Function*

The set of features providing information about the grammatical function of the anaphor and the antecedent is the next group that we will investigate. We removed this group from the full feature set, which we further refer to as F-Gf. F-Gf consists of features #16 through #23 from table 7.5. In order to have access to such information, the data needs to be annotated with additional levels of linguistic analysis, such as predicate-argument structure, predicate frame set, etc. Deeper linguistic analysis is not easily available for every language, which is the reason why we do not include features that are dependent on such annotation layers in our POS-based feature set. The problem of availability of annotation layers and respectively of state-of-the-art computational linguistic tools to provide such annotations is also well demonstrated by the current evaluation setting for the F-Gf feature set. While the *gold* data set for Arabic contained predicate-argument information, the *auto* set did not include such annotations [Pradhan et al., 2012]. Our experiments, though, make use of the *auto* layers of the data and for this reason we will not be able to provide results for Arabic in this particular experiment.

The system scores achieved by the F-Gf feature set for the two targeted languages, English and Chinese, are presented in detail in table 7.10. The four aspects that we consider in our examination follow below.

overall performance – The scores that UBIU achieves when the F-Gf feature set is used show that the information that this group carries is helpful to the multilingual memory-based coreference resolver. However, a very curious fact is depicted by the scores for English with respect to the three evaluation settings. According to the figures in table 7.10, the F-Gf feature set does not show itself to be beneficial when *gold* mentions are used in the pipeline – in other words, the lack of this set leads to an increase in scores by 0.76 percent points. Within the GB and AM settings, however, the scores for the given language decrease – 2.86 percent points for GB and 3.38 percent points for AM. This occurrence

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
GM	MD	100	100	100	100	100	100	100	100	100
full	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GM	MD	*	*	*	100	100	100	100	100	100
F-Gf	MUC	*	*	*	67.13	77.38	71.89	53.62	76.51	63.05
	B^3	*	*	*	62.87	61.42	62.13	58.34	68.95	63.20
	$CEAF_M$	*	*	*	52.32	52.32	53.32	47.63	47.63	47.63
	$CEAF_E$	*	*	*	54.78	38.47	45.20	55.22	30.75	39.50
	BLANC	*	*	*	70.03	62.42	64.44	68.78	58.57	60.58
	TOTAL			*			59.40			54.79
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
full	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.70			54.12
GB	MD	*	*	*	70.56	100	82.74	52.77	100	69.08
F-Gf	MUC	*	*	*	63.37	77.53	69.74	46.82	76.85	58.19
	B^3	*	*	*	62.87	57.05	59.82	53.68	74.61	62.43
	$CEAF_M$	*	*	*	45.58	45.59	45.59	47.36	47.36	47.36
	$CEAF_E$	*	*	*	51.19	32.26	39.58	57.04	27.97	37.53
	BLANC	*	*	*	66.19	56.78	54.45	68.95	61.35	63.77
	TOTAL			*			53.84			53.86
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
full	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39
AM	MD	*	*	*	62.18	59.82	60.98	45.91	64.45	53.62
F-Gf	MUC	*	*	*	49.89	44.53	47.06	37.68	49.22	42.69
	B^3	*	*	*	65.11	50.49	56.88	56.83	68.62	62.17
	$CEAF_M$	*	*	*	38.89	38.89	38.89	43.01	43.01	43.01
	$CEAF_E$	*	*	*	30.03	34.93	32.30	39.75	30.03	34.22
	BLANC	*	*	*	62.80	53.80	52.54	62.82	58.01	59.67
	TOTAL			*			45.53			48.35

Table 7.10: UBIU's results on the CoNLL 2012 datasets with the use of the F-Gf feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). We report the scores in terms of (P)recision, (R)ecall and (F)-measure from all evaluation metrics. TOTAL denotes their averaged F-measures. In grey the performance of the full feature set is listed. The evaluation settings for Arabic, marked with *, are the ones for which no annotations for the given feature set was provided.

indicates that this group of features is more helpful when an increased number of mentions is used by the system. This is not confirmed for Chinese, but could be explained by the overall lower number of mentions for this language. In fact, for English the difference in scores is quite big, since it varies between an increase of performance of 0.76 percent points for the GM setting and a decrease in performance of 3.38 percent points for the AM setting. Even though the overall results for the two languages as an average across the settings shows that the F-Gf group has a positive effect on the performance, the results are not as clear and indicative as for the F-GNPt group, especially with the lack of data for Arabic. One explanation for the positive score for English could be the fact that when *gold* mentions are used the performance of the system for some feature groups is significantly different than within the other two settings. This is an easily predictable behaviour, because the GM setting contains a significantly different set of mentions than the GB or AM settings. Furthermore, we already have seen several confirmations of this assumption in our previous results. For example, the outcome for Arabic and Chinese in the POS-based setting with a decrease per setting as follows: GM – 1.41 percent points, GB – 3.00 percent points and AM – 3.76 percent points for Arabic and GM – 1.06 percent points, GB – 0.45 percent points and AM – 0.59 percent points for Chinese. The F-L feature group also proved to lead to a different performance for the GM setting with respect to what the system achieved for the GB and AM settings. The decrease amounted to 0.85 percent points for GM, 3.43 percent points for GB and 3.13 percent points for AM. Such a distinctively different behaviour clearly shows that the GM setting does not always lead to similar system performance than the one achieved by GB and AM. This fact indicates that a pipeline, such as the one we employ in our work, cannot be optimized on one setting only and that the behaviour of the system on a specific feature set for the GM setting cannot be predicted on the basis of system output achieved from the GB and AM settings.

recall vs. precision – When recall and precision are concerned, the improvements in TOTAL scores for English within the GM setting, which means that the system does not profit from the F-Gf feature set in that setting, show similar tendencies. According to most metrics, the lack of the F-Gf group in the feature set tends to increase precision rather than recall. This is not the case for Chinese and the GM setting, which also leads to the overall lower TOTAL score for this language, being reduced from 55.64% to 54.79% (0.85 percent points). This outcome indicates that knowing the grammatical function of either the anaphor or the antecedent is more harmful rather than helpful with respect to recall for English when the gold mentions are used.

cross-lingual differences – Unlike the F-GNPt group, the results of the F-Gf feature set do not show a big variation across the languages. For English (which has the higher decrease) the scores are on average lower with 1.83 percent points in comparison with the use of the full feature set, while for Chinese the reduction is only 0.72 percent points. These figures once again confirm our previous observations that each feature group has a different effect on the system performance across the various languages.

mention detection – Mention detection is once more to be compared only for the GB and AM settings. The MD scores are interesting further, since the general correlation between the MD and coreference performance is not kept for both languages and settings. Namely, while the CR scores for English and Chinese for the GB setting are decreased, the overall MD F-measures are higher. This is counterintuitive, but we have observed such a behaviour for the POS-based feature set already. This indicates that the correlation is not confirmed by all of the evaluated feature sets and settings. However, we will continue our observations on this particular correlation for the rest of the feature groups in order to determine if such exceptions do occur more often and what the causes for the changing tendencies are.

7.4.4 *Grammatical Heuristic*

Section 7.4.4 will discuss the effect of a smaller feature group than the ones we analyzed in the previous sections – the grammatical heuristic set of features. The full set without the grammatical heuristic set of features is further referred to as F-Gh. The grammatical heuristic group consist of only two separate features that test if both mentions are embedded in a prepositional phrase or in a noun phrase. In order to extract features of this type, we need a syntactic parse of the data to be provided. The information is then further collected via heuristic rules that are not language specific but rather dependent on the given annotation layer.

All results achieved by the use of the F-Gh group as a feature set are listed in table 7.11. The following section offers our analysis and detailed discussion with respect to the new outcome and the overall conclusions that we can draw out of the general system performance.

The figures that table 7.11 offers lead to some very surprising conclusions, which are hardly categorizable in our predefined four aspects. The performance that UBIU achieves when the F-Gh feature set is used is very interesting, because there is absolutely no change (apart from the 0.01 percent points decrease for English in the GB setting) in performance for any of the languages and any of the three targeted settings. This shows that the two features within this set do not carry important information for the memory-based learner,

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
GM	MD	100	100	100	100	100	100	100	100	100
full	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GM	MD	100	100	100	100	100	100	100	100	100
F-Gh	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
full	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.70			54.12
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
F-Gh	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.69			54.12
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
full	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
F-Gh	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39

Table 7.11: UBIU's results on the CoNLL 2012 datasets with the use of the F-Gh feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). We report the scores in terms of (P)recision, (R)ecall and (F)-measure from all evaluation metrics. TOTAL denotes their averaged F-measures. In grey the performance of the full feature set is listed.

which is confirmed by all nine experimental settings for the evaluation of this group. The outcome assertively indicates that these two features can be left out for any other language that can be targeted within a similar system architecture and resolution procedure. We would like to note that the number of features within a given feature set does not necessarily stand in a correlation with the informativeness of that set. In other words, even a set with one feature can lead to a considerable change in scores if this feature carries information valuable for the coreference resolver.

7.4.5 *Grammatical Agreement*

The next two features in table 7.5, #26 and #27, build the feature group for grammatical agreement. These features describe if the mentions agree in either number or gender. In fact, they are often considered as important features and included in state-of-the-art feature sets. Unfortunately, number and gender information is not always easily available when it is not included in the POS tags or if the language does not follow easily derivable rules for their identification. In our approach, we also use agreement information only if it is included in the POS annotation layer (e.g. NN – is the POS tag for a singular noun and NNS – the POS tag for plurals) or if we could use clues such as definite articles for Arabic.

As in our previous evaluation settings, the results that the multilingual coreference resolution system achieves via the employment of the full set without the grammatical agreement features (F-Ga) are listed in detail in table 7.12 on page 212. Similar to the F-Gh set, F-Ga also consists of only two features. However, the performance that UBIU achieves with this set also leads to a new and informative outcome. Our analysis of the outcome is listed below:

> **overall performance** – The general system performance confirms our previous findings that feature sets are informative to a different degree to the memory-based learner across the languages and that for some languages, such as Arabic in this case, the feature set is not informative at all. What the figures in table 7.12 show is that for Arabic, the feature set is not helpful, leading to 0.00 percent points change in TOTAL scores for all three evaluation settings. However, this is not the case for English and Chinese, for which the decrease in scores is considerably larger. It is also important to note that unlike the F-Gh features, the number/gender information is extracted in a different way for each of the three languages because no annotation layer was included that provided this information for all data sets consistently. This means that the informativeness of the features could be improved by including consistent number/gender annotations either as a separate annotation layer or combined with the POS tagset.

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
GM	MD	100	100	100	100	100	100	100	100	100
full	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GM	MD	100	100	100	100	100	100	100	100	100
F-Ga	MUC	37.88	77.89	50.97	65.99	75.46	70.41	50.58	75.72	60.65
	B^3	44.46	89.32	59.37	62.17	54.96	58.34	55.52	72.17	62.76
	$CEAF_M$	49.06	49.08	49.07	47.22	47.22	47.22	48.96	48.96	48.96
	$CEAF_E$	68.07	29.54	41.20	52.20	37.23	43.47	57.82	30.72	40.12
	BLANC	57.59	79.94	60.85	65.10	56.99	56.38	68.52	60.82	63.19
	TOTAL			52.29			55.16			55.14
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
full	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.70			54.12
GB	MD	37.39	100	54.43	65.77	100	79.35	50.71	100	67.29
F-Ga	MUC	19.31	49.56	27.80	56.60	74.96	64.50	44.93	76.33	56.57
	B^3	34.69	80.98	48.57	55.00	60.85	57.78	52.05	76.00	61.79
	$CEAF_M$	37.20	37.20	37.20	44.48	44.48	44.48	47.09	47.09	47.09
	$CEAF_E$	56.81	22.25	31.98	53.36	29.86	38.29	57.13	27.28	36.92
	BLANC	54.14	60.30	55.23	64.25	57.06	57.08	68.54	62.81	64.95
	TOTAL			40.16			52.43			53.46
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
full	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39
AM	MD	17.34	70.55	27.83	58.24	65.01	61.44	45.25	71.43	55.40
F-Ga	MUC	11.80	52.62	19.27	45.23	46.64	45.93	37.55	53.80	44.23
	B^3	36.59	91.89	52.34	57.90	55.59	56.72	54.16	71.11	61.48
	$CEAF_M$	35.78	35.78	35.78	38.95	38.95	38.95	43.53	43.53	43.53
	$CEAF_E$	52.74	20.51	29.54	32.77	31.45	32.10	42.18	28.43	33.96
	BLANC	52.48	71.88	53.01	59.74	54.21	54.47	63.29	59.76	61.16
	TOTAL			37.99			45.63			48.87

Table 7.12: UBIU's results on the CoNLL 2012 datasets with the use of the F-Ga feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). We report the scores in terms of (P)recision, (R)ecall and (F)-measure from all evaluation metrics. TOTAL denotes their averaged F-measures. In grey the performance of the full feature set is listed.

recall vs. precision – For Arabic there is no change in TOTAL scores, as well as no change in the ratio of precision and recall between the full and the current feature set. When the results for English are observed, we can note that the decrease in performance is distributed across both precision and recall across all evaluation metrics, while this is not the case for Chinese (e.g. recall increases for MUC and B^3 in the GM and GB settings). The latter shows that while the two agreement features are clearly helpful for English, also confirmed by the highest variation in scores, for Chinese their informativeness is limited and in cases harms recall in favor of precision. Additionally these features seem not to be informative for Arabic at all.

cross-lingual differences – The cross-lingual differences can be observed again mainly with respect to the level with which the feature group is informative to the given language. Across all settings, the F-Ga group leads to 0.00 percent points change of TOTAL scores for Arabic, for English the reduction is the highest observed across all previous and further evaluations of the different feature sets – 3.68 percent points, while for Chinese the lack of these two features lead to an overall reduction of 0.56 percent points.

mention detection – For Arabic there is no change of scores in any of the settings and similar to previous settings we can compare only recall for the GB setting and the complete scores for the AM setting for English and Chinese. What the figures in table 7.12 show is that for English and the GB setting there is a decrease in both mention detection and coreference performance, while for Chinese in the same setting mention detection shows higher scores for the new feature set. Interestingly enough, the changes for the AM setting are not identical. While recall is decreased for English, precision seems to profit from the F-Ga set, but this improvement does not lead to an overall higher F-measure for the mention detection score. For Chinese, we can observe the opposite change, while MD recall increases for this language in the AM setting, precision is slightly decreased. The deviations lead to an overall higher F-measure for MD for Chinese, but lower coreference score. These findings once more show that MD is important to the resolution process, but is not the only factor that affects the system performance. In other words, the CR pipeline can be influenced independently by either MD or the actual ability of the system to identify the coreference links between the mentions that were considered. A well performing coreference resolution system, multilingual or not, is one that performs well with respect to both aspects.

7.4.6 *Semantic*

The set of semantic features also contains only two separate features. Feature #28 shows if both mentions have the same speaker, while feature #29 indicates if the mentions are the same named entity type or not. Similar to the predicate-argument information, speaker and named entity annotations were not provided within the *auto* set of the Arabic data. For this reason, we cannot evaluate the performance of the full set minus the semantic features (F-S) across all three languages, but again only with respect to English and Chinese. However, in general both features are considered important for the coreference resolution problem, specifically feature #29, the comparison between named entities, because it is often the case that mentions that do not agree on their NE class are seldom labeled as coreferent.

The detailed results for both English and Chinese are given in table 7.13 on page 215. Note that the slots filled with * denote the lack of annotations and correspondingly lack of different output (in comparison to the use of the full feature set) for the Arabic language.

The discussion of scores in the current section will once more focus on the four aspects we consider important and most indicative for the actual effect of the used feature sets on the system performance.

overall performance – UBIU's overall performance for English and Chinese provides strong evidence that for a system, such as the one we employ, namely one that uses memory-based learning, the two semantic features lead to an improvement of the overall performance when excluded from the full feature set (i.e. the F-S feature set leads to better system performance). In other words, the presence of the features in the feature set in use has a detrimental effect on the system performance: For English we have an increase in performance with 0.37 percent points for the GM setting, 0.45 percent points for GB and 0.20 percent points for AM; for Chinese, the GM setting improves with 0.18 percent points, GB with 0.03 percent points and AM with 0.14 percent points.

recall vs. precision – There are no specific tendencies with respect to precision and recall in connection to the increase of TOTAL scores for both English and Chinese. All metrics behave differently across the settings and use of mentions, which makes it hard to draw any sensible conclusions for the effect of the given feature set.

cross-lingual differences – Unlike other feature groups, such as F-GNPt or F-Ga, the deviation in scores is relatively small and in ranges similar for both languages – 0.34 percent points is the average across all settings for English and 0.12 percent points – for Chinese. This consistency and similarity shows higher certainty that the use of these two semantic

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
GM	MD	100	100	100	100	100	100	100	100	100
full	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GM	MD	*	*	*	100	100	100	100	100	100
F-S	MUC	*	*	*	67.45	76.91	71.87	49.48	76.10	59.97
	B^3	*	*	*	63.10	60.40	61.72	54.32	74.66	62.89
	$CEAF_M$	*	*	*	52.08	52.08	52.08	49.98	49.98	49.98
	$CEAF_E$	*	*	*	53.70	38.50	44.85	59.21	30.70	40.44
	BLANC	*	*	*	70.10	62.49	64.52	68.92	63.82	65.83
	TOTAL			*			59.01			55.82
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
full	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.70			54.12
GB	MD	*	*	*	67.26	100	80.42	49.75	100	66.45
F-S	MUC	*	*	*	58.26	77.78	66.62	44.20	77.75	56.36
	B^3	*	*	*	56.21	68.26	61.65	51.55	78.53	62.24
	$CEAF_M$	*	*	*	50.98	50.98	50.98	48.14	48.14	48.14
	$CEAF_E$	*	*	*	57.77	31.98	41.17	58.59	27.28	37.23
	BLANC	*	*	*	69.45	63.35	65.33	69.14	65.04	66.76
	TOTAL			*			57.15			54.15
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
full	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39
AM	MD	*	*	*	60.38	64.51	62.38	44.65	72.00	55.12
F-S	MUC	*	*	*	48.18	48.76	48.47	37.29	55.11	44.48
	B^3	*	*	*	60.45	58.51	59.46	53.78	72.77	61.85
	$CEAF_M$	*	*	*	42.99	42.99	42.99	44.20	44.20	44.20
	$CEAF_E$	*	*	*	34.21	33.68	33.95	43.28	28.36	34.27
	BLANC	*	*	*	64.50	58.92	60.67	63.75	62.08	62.85
	TOTAL			*			49.11			49.53

Table 7.13: UBIU's results on the CoNLL 2012 datasets with the use of the F-S feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). We report the scores in terms of (P)recision, (R)ecall and (F)-measure from all evaluation metrics. TOTAL denotes their averaged F-measures. In grey the performance of the full feature set is listed. The evaluation settings for Arabic, marked with *, are the ones for which no annotations for the given feature set was provided.

features for other languages will have a detrimental effect on the system scores.

mention detection – In almost all settings for which mention detection can be compared, except GB for English, the increase in coreference scores is not a result of improved mention detection. This fact confirms that the F-S feature set has a direct effect on the capability of the UBIU multilingual coreference resolution system to detect the coreference links between the mention pairs, but not in the expected positive direction (note that only a decrease in scores would have shown that the features are informative to the learner).

7.4.7 *Positional*

Positional features are also often considered as easily implementable and not dependent on various layers of annotation. For this reason, the two positional features are again included in our POS-based set of features. Feature #30 calculates the token distance between the two mentions and feature #31 indicates the sentence distance between them.

Once again we tested the effect of the given feature set for the performance of the UBIU multilingual coreference resolution system by excluding this set from the full feature set listed in table 7.5. We refer to the resulting set as F-P further on in our work.

The following paragraphs examine the changes of the overall performance of the system. We investigate if there are abnormal differences in recall and precision figures, the cross-lingual conclusions that we can draw from the numbers and as well our observations with respect to the correlation between mention detection and coreference performance.

overall performance – The two positional features that we make use of in the F-P feature set lead to an outcome both highly interesting and very similar to previous results, such as the output for the F-GNPt feature set for example. In general, the outcome shows that there is a big difference between the influence of the feature set across the languages and as well that for Arabic a striking gap across the three evaluation settings can be observed. For Arabic the decrease in results (meaning that positional features are informative for the Arabic CR learner) astonishingly reaches 10.00 percent points when *gold* mentions are used, while for GB and AM the difference is less substantial – 2.70 percent points and 2.09 percent points respectively. For English and Chinese the change leads to an improvement in TOTAL scores (meaning that the F-P group is rather uninformative and even harmful for the overall performance) within a range of 0.06 percent points for AM and Chinese and 1.10 percent points for English and the GB evaluation setting. This outcome once more

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
GM	MD	100	100	100	100	100	100	100	100	100
full	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GM	MD	100	100	100	100	100	100	100	100	100
F-P	MUC	20.66	74.88	32.38	67.50	77.20	72.03	49.96	76.62	60.48
	B^3	36.26	93.38	52.24	63.77	60.03	61.85	54.96	74.26	63.17
	$CEAF_M$	39.87	39.89	39.88	51.75	51.75	51.75	49.83	49.83	49.83
	$CEAF_E$	60.76	21.43	31.68	53.90	38.41	44.86	59.12	30.70	40.42
	BLANC	54.43	76.14	56.07	69.96	61.56	63.38	69.65	63.33	65.65
	TOTAL			42.45			58.77			55.91
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
full	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.70			54.12
GB	MD	14.59	100	25.47	67.83	100	80.83	49.83	100	66.51
F-P	MUC	12.22	87.65	21.45	59.24	78.37	67.48	44.42	78.19	56.65
	B^3	32.64	98.40	49.02	57.40	68.32	62.39	51.53	78.91	62.34
	$CEAF_M$	35.42	35.42	35.42	51.59	51.59	51.59	48.38	48.38	48.38
	$CEAF_E$	58.50	18.34	27.93	58.01	32.52	41.67	59.35	27.61	37.69
	BLANC	52.90	89.03	53.49	70.12	63.82	65.88	68.75	66.18	67.34
	TOTAL			37.46			57.80			54.48
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
full	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39
AM	MD	12.00	61.89	20.10	61.96	64.50	63.21	45.52	72.02	55.78
F-P	MUC	8.52	48.10	14.48	49.92	48.85	49.38	38.43	55.02	45.25
	B^3	35.63	93.21	51.56	61.65	57.13	59.31	54.49	71.44	61.82
	$CEAF_M$	33.96	33.96	33.96	42.97	42.97	42.97	43.80	43.80	43.80
	$CEAF_E$	51.39	19.32	28.09	33.33	34.32	33.82	42.37	28.49	34.07
	BLANC	51.61	68.99	51.43	64.98	59.03	60.86	64.02	61.08	62.33
	TOTAL			35.90			49.27			49.45

Table 7.14: UBIU's results on the CoNLL 2012 datasets with the use of the F-P feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). We report the scores in terms of (P)recision, (R)ecall and (F)-measure from all evaluation metrics. TOTAL denotes their averaged F-measures. In grey the performance of the full feature set is listed.

confirms the fact that feature optimization should be approached on a language dependent basis and that different system settings, such as the use of diverse mention sets as in the GM, GB, AM settings, changes of the machine learning algorithm, etc. should also be respectively anticipated.

recall vs. precision – Observing the figures for Arabic, we can see that when *gold* mentions are used, the loss of information seems to be highly influential in comparison to the case where only *gold* boundaries are made use of and even less for *auto* mentions. The recall and precision figures in this case are also very informative, because we can see that in general for Arabic the drastic change in scores is a result of majorly decreased recall and not this much decreased or in cases even increased precision figures. A very good example of this ratio is presented by the MUC metric across all evaluation settings, for which recall is in general decreased more than precision. Moreover, for the GB evaluation setting a surprising improvement of precision by 38.09 percent points (from 49.56% when the full feature set is used to 87.65% when F-P is employed) does not manage to lead to an overall higher F-measure with a recall that is decreased only with 7.09 percent points. These results are important to keep in mind, since increase in precision in this evaluation setting and especially in the GB case, means that positional features are important for Arabic not only, but mostly, with respect to precision.

cross-lingual differences – Our cross-lingual analysis of the scores concentrates on the high difference in averaged TOTAL scores. While Arabic's performance decreases with 4.93 percent points across all settings, the figures for English and Chinese improve with 0.53 percent points and 0.23 percent points on average. This huge difference between Arabic on the one side and English and Chinese on the other, confirms a fact that we could observe in other evaluation settings as well, such as F-L, F-GNPt and even F-Ga. System performance for English and Chinese is shown to lead to similar or closer scores than the ones the system achieves for the Arabic language. All three languages are typologically very different, but again one reason for this immense difference may be the large variation in the number of mentions per data set overall. Arabic has an exceedingly noun-phrase rich syntactic structure, which increases the complexity of the task for the memory-based learner. Thus, we assume that this is the main cause for the overall lower system performance reported for this language. Additionally, not all annotation layers were provided for Arabic, which may put different weights on the various feature collections and their direct effect to the MCR process.

mention detection – With respect to mention detection there are no new insights that this specific setting can give us. For Arabic the decrease in coreference scores is also connected to decrease in mention detection

figures, while for English the change is connected to an increase in both results. For Chinese and the GB setting the improvement in coreference scores is not accompanied by growth of mention detection figures, while for the AM evaluation setting both F-scores are increased.

7.4.8 *Other*

The last set of features that we want to evaluate is the one listed as *other* in table 7.5. This is a small set of two separate features: feature #32 indicating the normalized levenstein distance between the two tokens in the mention pair and feature #33, which is a learned feature, showing if the antecedent has already been classified by the system as singleton. The first feature is easily computable for any mention pair and does not require any specific annotation layer, nor more complex calculations. However, the extraction or computation of feature #33 is a complex problem on its own, since a separate classifier needs to be trained in order to label each mention as either being a singleton or not [Zhekova et al., 2012]. The singleton classifier is trained, based on the features listed in table 7.15, which we used in the participation of the UBIU multilingual coreference resolution system in the CoNLL 2012 shared task. The full set without feature #32 and #33 is further referred to as F-O.

The results that the multilingual coreference resolution system UBIU achieves when the F-O feature set is made use of are listed in table 7.16. The following paragraphs include our interpretation of the figures.

overall performance – The highest decrease in performance is observed for Arabic and the GM setting with a change of 1.85 percent points. However, the GB and AM settings for this language are less harmed by the lack of information provided by the features. On the contrary, English and Chinese seem to profit from the lack of the feature group within the GM setting and vice versa, the system reaches decreased scores for the latter two languages and the GB and AM setting. This shows once more, similar to the F-P setting for Arabic and F-Gf setting for English, that the evaluation setting for which *gold* mentions are used, seems to lead to most controversial results across the languages and feature groups. One very logical explanation for this phenomenon can be found in the drastically reduced number of mentions the system needs to work with, which poses a different level of difficulty and information need.

recall vs. precision – With respect to recall and precision, there are no new insights that we can gain from this evaluation setting.

cross-lingual differences – In general, the two features are shown to be informative mostly for Arabic for which the system reaches an average TOTAL decrease of 0.87 percent points. A lot smaller is the decrease for

#	Feature Description
1	the depth of the mention in the syntax tree
2	the length of the mention
3	the head token of the mention
4	the POS tag of the head
5	the NE of the head
6	the NE of the mention
7	PR if the head is premodified, PO if it is not; UN otherwise
8	D if the head is in a definite mention; I otherwise
9	the predicate argument corresponding to the mention
10	left context token on position token -3
11	left context token on position token -2
12	left context token on position token -1
13	left context POS tag of token on position token -3
14	left context POS tag of token on position token -2
15	left context POS tag of token on position token -1
10	right context token on position token +1
11	right context token on position token +2
12	right context token on position token +3
13	right context POS tag of token on position token +1
14	right context POS tag of token on position token +2
15	right context POS tag of token on position token +3
16	the syntactic label of the mother node
17	the syntactic label of the grandmother node
18	a concatenation of the labels of the preceding nodes
19	C if the mention is in a PP; else I

Table 7.15: The features used by the singleton classifier that extracts a feature for the coreference classification indicating if the given mention is potentially singleton mention or not. This set we also used in our participation at the CoNLL 2012 shared task [Zhekova et al., 2012].

English, namely 0.08 percent points and for Chinese the average decrease across all TOTAL scores is only 0.01 percent points. It is not surprising that Arabic profits the most from the current feature group, since the richer noun-phrase structure of this language leads to a proportionally higher number of singleton mentions and, thus, not coreferent mention pairs, as we noted in the overall performance analysis. However, it is

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
GM	MD	100	100	100	100	100	100	100	100	100
full	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GM	MD	100	100	100	100	100	100	100	100	100
F-O	MUC	34.90	76.37	47.91	67.78	76.69	71.96	49.74	76.50	60.29
	B^3	42.70	89.48	57.82	63.39	59.41	61.33	54.73	74.49	63.10
	$CEAF_M$	47.27	47.28	47.27	51.50	51.50	51.50	49.75	49.75	49.75
	$CEAF_E$	67.13	28.20	39.72	52.96	38.57	44.63	59.07	30.61	40.32
	BLANC	56.65	78.99	59.49	69.87	61.99	63.92	68.79	62.78	64.99
	TOTAL			50.44			58.67			55.69
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
full	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.70			54.12
GB	MD	36.79	100	53.79	66.67	100	80.00	49.94	100	66.61
F-O	MUC	18.81	48.90	27.17	57.57	77.12	65.92	44.44	77.39	56.46
	B^3	34.49	81.13	48.41	55.85	67.60	61.17	51.68	77.95	62.15
	$CEAF_M$	36.96	36.96	36.96	50.15	50.15	50.15	47.96	47.96	47.96
	$CEAF_E$	56.65	22.08	31.78	56.77	31.28	40.33	58.34	27.35	37.25
	BLANC	53.97	60.06	55.00	69.43	63.14	65.12	69.03	64.80	66.56
	TOTAL			39.86			56.54			54.08
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
full	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39
AM	MD	17.12	71.26	27.61	60.62	64.75	62.62	44.84	71.75	55.19
F-O	MUC	11.25	51.43	18.47	47.85	48.47	48.15	37.69	54.92	44.71
	B^3	36.10	91.87	51.83	59.89	58.16	59.02	54.23	72.11	61.91
	$CEAF_M$	35.33	35.33	35.33	42.50	42.50	42.50	43.98	43.98	43.98
	$CEAF_E$	52.87	20.36	29.40	34.09	33.52	33.80	42.84	28.43	34.18
	BLANC	52.29	70.84	52.66	63.98	58.80	60.47	63.92	60.74	62.07
	TOTAL			37.54			48.79			49.37

Table 7.16: UBIU's results on the CoNLL 2012 datasets with the use of the F-O feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). We report the scores in terms of (P)recision, (R)ecall and (F)-measure from all evaluation metrics. TOTAL denotes their averaged F-measures. In grey the performance of the full feature set is listed.

surprising that this improvement is not higher for the GB and AM settings for Arabic, unlike the change for English and Chinese, since only these settings actually include singletons in their mention sets. The reason for this unexpected outcome can be found in the actual accuracy of the singleton classifier. If it labels coreferent mentions as singletons, these are excluded from the set of mentions considered during the coreference resolution process and thus directly harm the overall outcome. Our scores in this setting, similar to our results from the participation of UBIU in the CoNLL 2012 shared task [Zhekova et al., 2012] indicate that singleton classification is a highly complex problem on its own, but can have a great impact on MCR and thus needs to be further improved and made use of on a language independent level. The outcomes show that proper exclusion of singleton mentions from the training/test data can reduce the bias of the memory-based learner towards increased tendency to label the output mentions as singletons as well. Furthermore, when working with less mentions altogether, the system provides more complete coreference chains, which is often confirmed by the increased results of the MUC metric.

mention detection – Similar to our observations with respect to recall and precision, there is no new or unexpected knowledge that we can gain from the current evaluation setting.

7.5 EXCLUDING THE SETS WITH DETRIMENTAL EFFECT

The previous section presented a thorough evaluation of the separate feature groups and their effect on the overall coreference resolution performance. We observed this effect by excluding them from the full coreference resolution set presented in table 7.5 and discussed the four different aspects: overall performance, recall vs. precision, cross-lingual differences and mention detection. We showed that some of the groups had a detrimental effect to the system performance and some carried informative and thus helpful information for the resolver. Before we summarize the findings of the previous sections and conclude the chapter, however, there is another evaluation that we would like to carry out.

One of the important findings of this research indicated that feature groups have a different effect on the separate languages. The latter means that language independent feature sets could be created only for information-poor approaches, for which a small number of language independent features are used, such as our POS-based evaluation. Feature sets, such as the one listed in table 7.5, include a number of features or feature groups that cannot be defined as language independent, because they were shown to have substantially distinct behaviour across the three languages we targeted. As a consequence, we gain the ability to optimize the feature set used by the system on a language

dependent basis. Even though our investigation focuses on the exploration and development of multilingual approaches that allow for language independent or at least as close to language independent behaviour as possible, we regard language dependent optimization as important. The knowledge we will gain with such an approach will be helpful for systems that target best performance for all languages they target.

For this reason, the current section creates language dependent feature sets that exclude the feature groups from the full feature set on a per-language basis. We remove the sets that were shown to have detrimental or no effect for the given language within our investigation. The full set of features without this language specific collection of sets is further referred to as F-Det. We assume that such a language specific optimization of the feature set will improve the system capability to detect coreference relations maximally for each of the targeted languages.

For Arabic, we exclude the F-L, F-Gh and F-Ga feature groups, while for the English and Chinese language dependent sets we do not consider the F-GNPt, F-Gh, F-S and F-P groups. Table 7.17 lists detailed results achieved by the multilingual coreference resolution system with the use of the F-Det language dependent feature set.

overall performance – The overall performance reported by UBIU shows a very positive outcome. For all languages, the F-Det language specific set reaches better performance than the one achieved by the full feature set, which shows that the features we exclude are not informative for the memory-based learner. For Arabic the scores improve with 0.19 percent points for the GM setting, with 0.17 percent points for GB and 0.30 percent points for AM. For English, the GM setting reaches a 2.84 percent points higher score than the full set, GB improves with 4.15 percent points and AM with 3.14 percent points. The improvements for Chinese are again very close to the performance for English, namely 2.72 percent points of increase for GM, 4.45 percent points for GB and 3.45 percent points for AM. This shows, that feature sets can be optimized on a per-language basis. Moreover, as we showed in some of the evaluation settings, the best optimization should as well include other factors, such as for example the GM, GB or the AM setting, which also show a variation in the performance tendencies. Our findings also indicate that such an investigation could and should be further broken down to an optimization performed by excluding separate features and not feature groups. However, the latter issue raises the question whether the trade-off between optimization/improvement and the time/effort that needs to be invested in it for a language dependent approach is appropriate with respect to the achieved improvement. This is so because a possible optimization on per-feature bases for every language, every mention set (GM, GB, AM), every type of system or learning approach

		AR			EN			ZH		
		R	P	F_1	R	P	F_1	R	P	F_1
GM	MD	100	100	100	100	100	100	100	100	100
full	MUC	37.88	77.89	50.97	67.69	76.59	71.86	49.64	76.24	60.13
	B^3	44.46	89.32	59.37	63.29	59.40	61.28	54.68	74.30	63.00
	$CEAF_M$	49.06	49.08	49.07	51.49	51.50	51.49	49.69	49.69	49.69
	$CEAF_E$	68.07	29.54	41.20	52.91	38.53	44.59	58.98	30.59	40.29
	BLANC	57.59	79.94	60.85	69.85	62.03	63.98	68.93	62.89	65.11
	TOTAL			52.29			58.64			55.64
GM	MD	100	100	100	100	100	100	100	100	100
F-Det	MUC	38.26	77.53	51.23	60.35	80.78	69.09	47.25	83.63	60.38
	B^3	44.85	88.83	59.60	55.99	76.89	64.80	52.12	86.67	65.09
	$CEAF_M$	49.47	49.48	49.47	56.70	56.71	56.71	53.82	53.82	53.82
	$CEAF_E$	68.18	29.80	41.48	65.11	35.94	46.32	65.78	30.54	41.71
	BLANC	57.47	78.82	60.64	70.51	70.48	70.50	68.15	74.59	70.79
	TOTAL			52.48			61.48			58.36
GB	MD	37.39	100	54.43	67.06	100	80.28	50.04	100	66.70
full	MUC	19.31	49.56	27.80	57.89	77.18	66.16	44.48	77.37	56.49
	B^3	34.69	80.98	48.57	56.05	67.46	61.23	51.74	77.90	62.18
	$CEAF_M$	37.20	37.20	37.20	50.35	50.35	50.35	48.00	48.00	48.00
	$CEAF_E$	56.81	22.25	31.98	56.94	31.58	40.63	58.34	27.38	37.27
	BLANC	54.14	60.30	55.23	69.50	63.10	65.10	69.17	64.87	66.66
	TOTAL			40.16			56.70			54.12
GB	MD	38.66	100	55.76	67.43	100	80.55	52.64	100	68.97
F-Det	MUC	20.03	49.94	28.59	57.13	82.65	67.56	46.68	85.22	60.32
	B^3	34.98	80.25	48.72	53.36	81.07	64.36	51.82	88.36	65.33
	$CEAF_M$	37.17	37.17	37.17	56.09	56.09	56.09	54.00	54.00	54.00
	$CEAF_E$	56.65	22.44	32.15	66.83	33.54	44.66	66.30	30.09	41.39
	BLANC	54.01	59.18	55.00	70.80	72.49	71.59	68.60	76.72	71.80
	TOTAL			40.33			60.85			58.57
AM	MD	17.34	70.55	27.83	61.06	64.65	62.80	44.87	71.69	55.19
full	MUC	11.80	52.62	19.27	48.20	48.47	48.34	37.76	54.94	44.76
	B^3	36.59	91.89	52.34	60.17	58.03	59.08	54.33	72.05	61.95
	$CEAF_M$	35.78	35.78	35.78	42.65	42.65	42.65	44.00	44.00	44.00
	$CEAF_E$	52.74	20.51	29.54	34.11	33.86	33.98	42.82	28.46	34.19
	BLANC	52.48	71.88	53.01	64.07	58.83	60.50	63.98	60.66	62.03
	TOTAL			37.99			48.91			49.39
AM	MD	18.81	71.88	29.82	61.35	63.64	62.47	48.54	69.55	57.17
F-Det	MUC	12.43	52.66	20.11	47.64	50.36	48.96	40.60	57.83	47.71
	B^3	36.71	91.56	52.41	59.98	65.29	62.52	56.36	76.67	64.96
	$CEAF_M$	35.86	35.86	35.86	46.44	46.44	46.44	48.07	48.07	48.07
	$CEAF_E$	52.68	20.72	29.75	38.35	35.84	37.05	46.13	31.91	37.72
	BLANC	52.66	72.07	53.32	65.46	65.07	65.26	64.54	67.21	65.75
	TOTAL			38.29			52.05			52.84

Table 7.17: UBIU's results on the CoNLL 2012 datasets with the use of the F-Det feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). We report the scores in terms of (P)recision, (R)ecall and (F)-measure from all evaluation metrics. TOTAL denotes their averaged F-measures. In grey the performance of the full feature set is listed.

parameter setting is an immense amount of computational work. Our evaluation showed that such an optimization is a huge effort that leads to an overall improvement of 2.38 percent points as an average on all three languages. This figure may change depending on the targeted languages, such as Arabic for example, which gained only 0.22 percent points.

recall vs. precision – For Arabic, for which the F-Det accomplishes the smallest improvement in scores, the F-score, as well as precision and recall figures are identical with the performance of the system achieved by the F-L feature set. However, for the English and Chinese languages, and especially based on the results reported by the MUC and B^3 metrics, the main improvement in results is based on the slight reduction of recall, but increase in precision across all evaluation settings. This indicates that the excluded feature groups for those languages allow for a better recall, but on account of that for worse precision in the detection of the coreference links.

cross-lingual differences – As we already pointed out, the improvement for Arabic is only 0.22 percent points, which is basically derived from the exclusion of the F-L feature set, as the other two sets do not show themselves to have an effect on the performance. However, for English and Chinese the performance of the F-Det set is an accumulation of the various excluded feature sets per language. English reaches a performance of 3.38 percent points higher than the use of the full feature set. Judged by the previously seen scores, this change is mostly due to the effect of the F-GNPt feature group. The changes are similar for Chinese – an increase of 3.54 percent points, again mainly based on the absence of the F-GNPt group of features.

mention detection – With respect to mention detection, we received one last confirmation that in general MD features are tightly bound to the overall CR performance. Using the F-Det feature set shows a correlation between both in almost all cases apart from English and the AM setting, for which MD performance decreases with 0.33 percent points (due to decreased precision figures), while the overall CR performance improves by 3.14 percent points. This again indicates that even though both aspects are very closely related, they also have an independent effect on the combined system performance.

7.6 SUMMARY AND CONCLUSION

For a better overview and clarity of the results from all evaluation settings, we provide a table (table 7.18) that lists all TOTAL scores across the various experimental settings and languages as well as the calculated deviation of the given set of employed features from the use of the full feature set.

Feature Set	Set.	AR		EN		ZH		FST		
		TOTAL	Dev	TOTAL	Dev	TOTAL	Dev	AR	EN	ZH
Full	GM	52.29	-	58.64	-	55.64	-	-	-	-
	GB	40.16	-	56.70	-	54.12	-			
	AM	37.99	-	48.91	-	49.39	-			
POS-based	GM	50.88	-1.41	54.76	-3.88	54.58	-1.06	-2.72	-3.27	-0.70
	GB	37.16	-3.00	54.52	-2.18	53.67	-0.45			
	AM	34.23	-3.76	45.17	-3.74	48.80	-0.59			
F-L	GM	52.48	+0.19	57.79	-0.85	53.98	-1.66	+0.22	-2.47	-1.83
	GB	40.33	+0.17	53.27	-3.43	52.21	-1.91			
	AM	38.29	+0.30	45.78	-3.13	47.48	-1.91			
F-GNPt	GM	49.78	-2.51	62.14	+3.50	58.28	+2.64	-1.71	+3.08	+3.17
	GB	39.52	-0.64	59.91	+3.21	57.77	+3.65			
	AM	36.02	-1.97	51.45	+2.54	52.61	+3.22			
F-Gf	GM	*	*	59.40	+0.76	54.79	-0.85	*	-1.83	-0.72
	GB	*	*	53.84	-2.86	53.86	-0.26			
	AM	*	*	45.53	-3.38	48.35	-1.04			
F-Gh	GM	52.29	0.00	58.64	0.00	55.64	0.00	0.00	0.00	0.00
	GB	40.16	0.00	56.69	-0.01	54.12	0.00			
	AM	37.99	0.00	48.91	0.00	49.39	0.00			
F-Ga	GM	52.29	0.00	55.16	-3.48	55.14	-0.50	0.00	-3.68	-0.56
	GB	40.16	0.00	52.43	-4.27	53.46	-0.66			
	AM	37.99	0.00	45.63	-3.28	48.87	-0.52			
F-S	GM	*	*	59.01	+0.37	55.82	+0.18	*	+0.34	+0.12
	GB	*	*	57.15	+0.45	54.15	+0.03			
	AM	*	*	49.11	+0.20	49.53	+0.14			
F-P	GM	42.45	-10.00	58.77	+0.13	55.91	+0.27	-4.93	+0.53	+0.23
	GB	37.46	-2.70	57.80	+1.10	54.48	+0.36			
	AM	35.90	-2.09	49.27	+0.36	49.45	+0.06			
F-O	GM	50.44	-1.85	58.67	+0.03	55.69	+0.05	-0.87	-0.08	-0.01
	GB	39.86	-0.30	56.54	-0.16	54.08	-0.04			
	AM	37.54	-0.45	48.79	-0.12	49.37	-0.02			
F-Det	GM	52.48	+0.19	61.48	+2.84	58.36	+2.72	+0.22	+3.38	+3.54
	GB	40.33	+0.17	60.85	+4.15	58.57	+4.45			
	AM	38.29	+0.30	52.05	+3.14	52.84	+3.45			

Table 7.18: A summary of UBIU's TOTAL scores on the CoNLL 2012 shared task datasets with the use of all combinations of the feature set on the *gold* mentions (GM), the *gold* boundaries (GB) and on *auto* mentions (AM). The scores are the TOTAL figures that denote the averaged F-measures of all metrics per setting. Column *Dev* gives the calculated setting deviation from the full set. The column *FST* lists the average group deviation from the full feature set. The evaluation settings for Arabic, marked with *, are the ones for which no annotations for the given feature set was provided.

To sum up, our investigation of the effect of the various groups of features to the overall system performance when a memory-based multilingual coreference resolution system, such as UBIU, is made use of, showed several aspects important for our work:

- When the full feature set is used, the system performs competitively against state-of-the-art MCR approaches, as also reported by the CoNLL 2012 shared task proceedings [Pradhan et al., 2012]. UBIU was ranked 6th out of 16 participating systems on the CoNLL 2012 shared task (which is an international and multilingual enterprise). This indicates that our system has a good estimate for the MCR problem overall. The fact that there were so few participants and even less managed to submit scores for all three languages confirms our observations that multilingual coreference resolution is a highly challenging task. The latter was also one of the main conclusions of the SemEval-2 shared task where UBIU was one of only two systems that submitted scores for all six targeted languages.

- An employment of the POS-based feature set leads to an approximate decrease of 2 percent points across the three languages and evaluation settings we investigated. This is a trade-off that is acceptable when no further annotations but POS information are provided within the used datasets. Our results indicated that missing syntactic information can be partially approximated by this type of local annotations. This shows that information-poor approaches should be given further attention and that less resourced languages will not be deeply affected by the lack of annotated data and state-of-the-art NLP tools. The most valuable conclusion with respect to the use of the POS-based feature set is that it is possible to employ a fully language independent pipeline within the UBIU MCR system. We recall that we provided language independent solutions based on the POS annotation layer for each of the main subtasks of CR discussed in our work: mention detection, mention head detection and now, feature selection.

- Each feature set that we examined showed itself to have a language-specific behaviour. Thus if optimal performance for each of the targeted languages is needed, a language dependent optimization, such as the one we presented by assembling and testing the F-Det feature set should be aimed at.

- The informativeness of the feature sets cannot be determined on a general basis, because it appears that it is tightly connected with the employed evaluation setting. For this reason, feature sets should be additionally optimized for each of the settings separately.

- Our experiments showed that there are feature sets that show uncertain or surprising results because some of the outcomes lead to a decreased system performance. In general, all the features we used are well established features in various state-of-the-art approaches, so such detrimental outcomes were unexpected and thus need to be further investigated. For example, an evaluation that examines the effect of each of the features within the feature groups that lead to such surprising outcomes could show if the behaviour was due to one, multiple or all features present in the given group. However, we note again that this is an exceedingly time-consuming task that goes beyond the scope of this work.

- Similar to our findings in chapter 5, mention detection and proper coreference resolution were shown to be tightly connected. While mention detection directly affects the overall CR performance, tuning the coreference resolver also shows an effect on MD results. The reason is that improved CR performance means more correctly found links and thus less wrongly excluded mentions from the final output. In certain cases, such as the evaluation of the F-Gf feature set and the GB setting for English and Chinese, the correlation between mention detection and coreference results does not hold, which confirms that MD is not the only factor that has a significant effect on the final system performance. The proper and optimal selection of features can lead to an influential change of the informativeness of the feature vectors for the memory-based learner. In our work so far, we found evidence for both theses: that mention detection is highly important and can directly affect the coreference performance and vice versa that the actual coreference performance can influence the mention detection scores to a great extent. This shows that an improvement of the full coreference process can be achieved only when all important aspects (mention detection and feature selection, as well as mention head detection, discussed in chapter 6) are improved independently from each other, so that possible interference is ruled out.

- There are features and feature sets that do not show themselves to be valuable to any of the three targeted languages, such as the F-Gh feature set for example. The latter is definitely not informative for any of the evaluation settings, from which we cannot draw specific conclusions, but which shows higher confidence that this feature set will also be uninformative for any other potentially targeted languages. Additionally, feature sets, such as F-Ga, seem to be less informative for some languages (e.g. Arabic), but more so for others (e.g. English). Such sets are informative with less confidence on a cross-lingual basis, but can and should be used for other new languages as well, as they do not indicate to have a detrimental effect on the scores either.

- There is no guarantee that features that are believed to be helpful for state-of-the-art systems and are normally included in their feature sets will be informative and thus beneficial for different types of systems on which the effect of those features has not been tested. We showed that well established features for various CR approaches did not lead to the expected results within the UBIU system. Thus, feature optimization needs to be not only language specific but as well learner/system specific.

- The three evaluation settings (GM, GB and AM) do not always seem to show similar tendencies. The bigger gaps in TOTAL scores that we observed for the POS-based feature set and the F-L feature set for English, the F-GNPt set and Arabic, the F-Gf feature set and English, the huge gap for Arabic in the F-P evaluation setting as well as the last feature set, F-O, for all targeted languages, indicates that the use of *gold* mentions may lead to a different outcome than the use of *gold* boundaries or *auto* mentions in specific cases. We assume that the deviations that the GM setting reaches, contrary to the GB and AM, are the result of the highly reduced number of mentions with which the multilingual coreference resolution system has to work with. The latter means that the memory-based learner is confronted with an easier task and, thus, a different amount and type of information is needed. This fact also indicates that the performance of a MBL-based CR system in the GM setting cannot be predicted with high certainty on the bases of output achieved in either the GB or the AM settings.

- Morphologically rich languages, such as Arabic, do not provide informative lexical features. The increased number of lexical surface forms poses a higher complexity for the MBL learner and features of this type should be avoided for this class of languages.

- As we showed, for the F-L and F-GNPt feature sets, MBL proves to be able to partially induce the information provided by the grammatical NP type feature set from the information that the lexical features carry with respect to languages that are not morphologically complex. Thus, these two sets seem to convey redundant information for this type of learner. Our results suggest that morphologically rich languages should include the features presented in the grammatical NP type feature set, while the memory-based learner can make a better use of lexical features for the rest of the languages.

- Singleton classification is still not optimally used within UBIU and therefore more attention needs to be devoted to this matter. We noted that this aspect has not yet been largely improved on a language independent manner and should be observed regarding this aspect. There are several

ways that the detection of singleton mentions can be incorporated within a system.

- First, singletons could be filtered out before the formation of the feature vectors that the memory-based classifier uses. This changes the amount of data that the classifier needs to deal with, but also introduces direct error propagation, because the wrongly classified singletons will not be considered by the learner either.
- Second, instead of directly reducing the data, the information that the singleton classifier provides can be used in a different way (as we showed in our work), namely as additional information to the feature set, or in other words – as a separate feature, that the classifier can use during classification. This makes it possible to keep the original proportion and ratios in the data and to let the classifier make more objective conclusions based on the new information.
- Furthermore, a third option for the use of the singleton classifier in a MCR system, such as UBIU, can be employed. Once the set of coreferent mentions has been identified by the memory-based learner, there will be mentions that were found to be potentially coreferent by the singleton classifier but not included in any coreference chain by the coreference classifier. During post-processing, though, all mentions that are not part of a coreference chain are removed. At that step, one can use the information provided by the singleton classifier and remove only the mentions that are not a part of a coreference chain and have been positively classified by the singleton classifier. This may lead to an improvement in scores, because the system has not managed to identify the correct chain to include those mentions, but has managed to correctly recognize that they are potentially coreferent, for which it will partially be rewarded by some of the evaluation metrics.

The current chapter presented a detailed evaluation of the effect of various feature groups on the full coreference pipeline. We carried out an analysis on three languages that are typologically highly different, which has not been done before in such detail and for this many languages simultaneously with respect to coreference resolution. We concluded with multiple important findings that can be further used for the development of multilingual coreference resolution systems independent of the machine-learning approach that is employed. Our results raised further questions that can be investigated in the future, such as the fact that instead of evaluating feature groups, an even more detailed and exact analysis can be carried out on a per-feature basis. Additionally, new multilingual features, that can replace ontological information, or in other words, features carrying world knowledge, should also be explored and better integrated in the coreference resolution pipeline.

Part IV

FUTURE WORK AND CONCLUSION

CHAPTER

8

DISCUSSION, FUTURE WORK AND CONCLUSION

The current work described different aspects that have a significant influence on the challenging task of Coreference Resolution (CR), when this task is advanced to a new, more complex level – multilinguality. We have covered the most important aspects of multilingual coreference resolution, as well as the main problems that Multilingual Coreference Resolution (MCR) systems, using the mention-pair model as a coreference model and memory-based learning for the resolution process, need to solve in order to tackle this highly demanding task. Our investigation covered datasets in eight different languages (Arabic, Catalan, Chinese, Dutch, English, German, Italian and Spanish) concerning the three most important steps in the MCR pipeline of the UBIU system – mention detection, mention head detection and feature selection.

In this chapter, we would like to review and discuss our findings (see section 8.1), summarize the open problems for multilingual coreference resolution (see section 8.2) and propose new ways and directions that can further be investigated (section 8.3). This will give us the possibility to abstract away from the details and summarize the newly gained knowledge from a different perspective. With this abstraction and overview, we aim to provide the chance for other approaches to the problem (employing other coreference models or machine learning methods, as well as targeting a different set of languages) to easily apply our advances in a new environment. Section 8.4 introduces the

first international enterprise that aims at such an advance. In section 8.5, we conclude our work.

8.1 DISCUSSION

In chapters 2 through 4, we introduced the complex task of coreference resolution by describing both rule-based as well as machine learning approaches to the problem. We discussed the difficulties that they face: the manual effort needed to develop rules that gain wide coverage and at the same time are efficient and accurate on the one hand, as well as the necessity of large corpora, annotated with multiple layers of linguistic information, on the other. Additionally, we delineated the various attempts to improve on both methods or merge them in a hybrid approach to coreference resolution. However, we also showed that performance improvement is not the only important direction to explore in the field. Coreference resolution has been furthermore extended to accommodate emerging and modern needs for robust and accurate CR performance for more than one given language and the ability to easily adapt already existing methods to further, unexplored and less resourced languages. Multilinguality of the CR pipeline, the main topic of our work, was introduced within the framework of the UBIU multilingual coreference resolution system that uses the mention-pair model as a coreference model as well as memory-based learning for the resolution procedure.

With respect to the selected framework, we showed that there are numerous predicaments that a CR pipeline of this kind faces when multilinguality is introduced. Along the way we examined various questions that we sum up below together with the conclusions we drew based on the evidence our experimental work provided:

WHICH ANNOTATION LAYERS ARE SUFFICIENT FOR THE DEVELOPMENT OF A MULTILINGUAL APPROACH? Our findings showed that across all three important subtasks of MCR, part-of-speech information is sufficient to develop a minimalistic, but language independent approach based on machine learning techniques.

With respect to *mention detection*, we showed that the Mention Detection based on IOB Annotation (mdIOBA), a machine learning method, provides the highly important flexibility for easy adaptation to any language that is newly introduced to a system for which only Part of Speech (POS) information needs to be provided. We also showed that mdIOBA reaches a performance that is competitive to rule-based approaches and that additional annotation layers are not necessarily needed but can be beneficial for mdIOBA's performance.

With respect to *mention head detection*, part-of-speech and mention head information was the minimal annotation requirement. With it, the Mention Head Detection Machine Learning Based (mhdML) approach was capable of

tackling the multilinguality problem and it led to system results that indicated that mhdML can be used reliably without much annotation effort. We showed that the heuristic approach used by many state-of-the-art systems does not provide a competitive performance for every targeted language, but can be used when no mention head information is provided and the targeted language has a consistent directionality.

The last aspect we discussed, *feature selection*, also showed itself to be a problem that can be efficiently resolved by machine learning techniques that solely use part-of-speech information. The POS-based feature set did not achieve optimal performance, such as the one reported by the F-Det set (which is the language specific feature set constructed by excluding the groups of features with detrimental effect for every language from the full feature set from table 7.5 on page 192). However, the considerably smaller reduction reached by the POS-based set with respect to the scores gained by the full feature set (2.23 percent points of overall system performance across all three evaluated languages) seems to be a reasonable trade off against the needed effort to provide further annotation layers. Once multilinguality does not pose a problem to the CR pipeline, system optimization in a language independent manner would be a reasonable direction for future investigation.

WHICH ANNOTATION LAYERS ARE BENEFICIAL FOR A RELIABLE, ROBUST AND WELL PERFORMING APPROACH? In all three important subtasks of the problem, we showed that additional annotation layers can either enhance the performance of machine learning methods or provide the background for the development of rule-based approaches.

According to our investigation, constituency and dependency structures were highly beneficial for the *mention detection* procedure. We showed that rule-based approaches can be successfully and easily implemented, when such syntactic annotations are provided. Additionally, the system performance indicated that the Memory-Based Learning (MBL) method we employed for this task (mdIOBA) improved when syntactic information was included in its feature set. This improvement showed that mdIOBA can profit from additional information and thus it is a competitive and robust solution to the mention detection task.

Mention head information is not provided in standard dataset distributions for which we showed that it is the key to the development of competitive and robust machine learning methods for *mention head extraction*. We propose that the addition of this layer of annotation is included in standard linguistic annotation distributions for the coreference resolution task.

With respect to *feature selection*, we showed that the complexity and importance of this subtask of coreference resolution is immense. We also demonstrated that the identification of appropriate features on a language dependent basis, extracted from various annotation layers, is not a straightforward task,

but rather a highly time-consuming matter. Moreover, our investigation of the feature selection procedure did not elicit a preference for any specific annotation layer. Our results indicated that for feature selection any additional information is beneficial and can be used depending on the particular setting and language in use. We also showed that some of the information provided by various annotation layers can be encoded within the information the POS annotations carry (depending on the POS tagset, number, even gender information can be extracted from the tags).

WHICH LAYER PROVIDES THE MOST INDICATIVE INFORMATION FOR MULTILINGUAL COREFERENCE RESOLUTION? UNDER WHAT CIRCUMSTANCES CAN THIS INFORMATION BE EMPLOYED? As we already noted, POS is the only annotation layer that is crucial to the MCR pipeline, which also makes it the most indicative information repository. This is so because only this annotation layer is provided across all targeted languages, and can be easily integrated into a language independent approach to CR. The optimal circumstances under which this annotation layer can be best employed from a MCR framework would be to make use of a language independent POS tagset. The latter can provide equal granularity (division of main lexical categories, such as *nouns* into subcategories, such as *proper nouns* and *common nouns*) for all different languages as well as an equal possibility and encoding for morphological features, if such are present in the language. Universal and language independent coarse POS tagsets have already been investigated in various approaches [Yarowsky and Ngai, 2001, Xi and Hwa, 2005, Das and Petrov, 2011, Petrov et al., 2012], which could be easily made use of in multilingual enterprises, such as MCR.

Additionally, syntactic information also proved to play an important role in the MCR pipeline. However, even more effort needs to be invested into the preparation of this layer for every targeted language. Moreover, for a multilingual and flexible pipeline, this layer of annotation needs to correspond to the annotation scheme used for all targeted languages and annotation schemes. We showed that annotation schemes for coreference may have a highly divergent overlap with the underlying syntactic structure of the language and thus lead to decreased system performance.

HOW IMPORTANT IS THE RELIABILITY OF THE ANNOTATION LAYERS ACROSS THE VARIOUS LANGUAGES? Within our Mention Detection (MD) investigation, we researched in more detail how important the reliability of the annotation layers is across the various languages. Our analysis indicated that there is not only a wide variation between the annotation schemes used across the languages, but as well a big variation between the quality of the provided information. This issue can be an essential problem for any of the procedures

integrated in the multilingual coreference resolution pipeline and can also significantly affect the system performance on a language dependent level.

Additionally, especially with respect to Arabic, we demonstrated that not only the quality, but as well the lack of annotations can be even more crucial to the development process. In some of the evaluation settings of feature selection that we presented in chapter 7, we were not able to provide system output for this language, because of missing annotation layers. These findings were also one more important reason for the exploration of an information-poor approach to multilingual coreference resolution.

These facts indicate that the biggest predicament for MCR is the lack and unreliability of commensurate resources across languages as well as the use of divergent annotation schemes for the coreference annotation.

WHICH PROBLEMS CAN OCCUR WHEN MD IS APPROACHED FOR MORE THAN ONE LANGUAGE AND WHAT WOULD THE PREREQUISITES FOR OBJECTIVE EVALUATION OF MENTION DETECTION METHODS BE? Our in-depth MD analysis delineated several big problems for this particular subtask of MCR:

- When rule-based approaches are developed, language specific knowledge is needed in order to assemble rules that are accurate and well performing.
- With respect to machine learning and MCR, we demonstrated that the higher consistency across the annotation schemes and better quality of the actual annotations directly translates to an improvement of the system's performance.
- We also showed that the datasets provided by the two shared tasks (SEMEVAL-2 and CONLL-2012) differed in the presence or absence of singletons in them. This can lead to serious ramifications with respect to evaluation (as in the CONLL-2012 shared task and our intrinsic evaluation where we showed that mention detection cannot be objectively evaluated when singletons are not present in the key data) and thus a harder comparison between the various languages. For this reason, we propose that singleton mentions are always included in standard dataset distributions.

WHICH OF THE FEATURES USED IN MONOLINGUAL APPROACHES ARE APPLICABLE IN MULTILINGUAL OR EVEN LANGUAGE INDEPENDENT METHODS? In chapter 7, we evaluated the different feature groups and their informativeness to the memory-based learner in the attempt to find an answer to this question. Our investigation showed that feature selection is a highly complex task for which general conclusions on a multilingual level can hardly be drawn. As we showed, various groups seemed to have a beneficial effect for

some languages and a detrimental one for others (e.g. the lexical, grammatical NP type or positional groups of features). Some could not be evaluated for all languages for lack of annotations (e.g. the grammatical function and the semantic groups) and one did not show itself to have an effect for any of the targeted languages (grammatical heuristic). Furthermore, we discussed the fact that features can only be extracted when the corresponding annotation layers are provided for all languages. Keeping this in mind, we concluded that feature selection and optimization should be performed for every language and specific system separately depending on the annotations that are provided in the specific dataset.

Additionally, we showed that our POS-based feature set allows for a completely language independent behaviour of the system, because it can be easily employed when POS information is provided for all languages with only 2.23 percent points decrease in system performance in comparison to the initial full feature set.

IS THERE A DIFFERENCE BETWEEN THE INFORMATIVENESS AND IMPORTANCE BETWEEN THE TYPES OF FEATURES CONSIDERED BY THE PIPELINE AND WHICH TYPE CARRIES THE MOST DESCRIPTIVE AND HELPFUL INFORMATION FOR THE COREFERENCE RESOLVER? DOES THAT TREND CHANGE ACROSS LANGUAGES AND WHICH IS THE SETTING MOST HELPFUL TO ALL TARGETED LANGUAGES? The evaluation of the feature groups that we conducted brought strong evidence that all feature groups differ in their importance and informativeness for the coreference resolver. However, we also showed that none of the groups carried equally descriptive and helpful information across all languages (apart from the grammatical heuristic group, which was not helpful to any of the languages). For this reason, in a multilingual approach features need to be selected on a language dependent manner in order for optimal system performance to be achieved. In our work, we also assembled such language dependent feature sets (F-Det) that included only the feature groups per language that were informative to the learner for that given language. The F-Det language dependent feature sets lead to an average increase of system performance of 2.38 percent points with respect to the performance achieved by the full feature set.

ARE THERE OTHER LAYERS OF ANNOTATIONS OR EXTERNAL SOURCES OF KNOWLEDGE THAT CAN ENHANCE THE RESOLUTION PROCESS AND WHAT KIND OF LIMITATIONS DO THOSE SOURCES HAVE? In section 7.1, we discussed the various layers of annotation that state-of-the-art systems use and presented an evaluation of the use of ontological information. Our investigation indicated that ontological information in the form and coverage

provided by WordNet is not sufficient for tasks with the scale of coreference resolution and even less of multilingual CR.

We also discussed the fact that additional layers of annotation, such as ontological information, mainly pose limitations to multilinguality and coverage. This is a highly important finding for our work, because it confirms that multilingual approaches need to be developed in an exceedingly minimalistic and information-poor manner, so that they are not dependent on sources and annotation layers that cannot be provided for all languages targeted by the system.

Besides all the important issues we described above, the main question that our work aimed to answer captures the general idea of multilinguality in coreference resolution. We showed that multilingual MCR faces various difficulties within the selected framework. Moreover, we demonstrated that multilinguality can be separated into two different levels: working successfully on a defined set of more than one language (as for example the eight languages that we used in our investigation) or achieving a completely language independent manner of processing. For this reason, the next question proved to be the most valuable question of our exploration.

CAN COREFERENCE RESOLUTION BE PERFORMED IN A CLOSE TO LANGUAGE INDEPENDENT MANNER WITHIN THE FRAMEWORK OF THE UBIU MULTILINGUAL COREFERENCE RESOLUTION SYSTEM? We showed a detailed evaluation of various approaches to the important subtasks of CR with respect to a mention-pair coreference resolution model and a resolution process based on memory-based learning techniques. The three subtasks of CR (mention detection, mention head detection and feature selection) were shown to be highly challenging when multilinguality was taken as a key feature for the system performance. Our work described in greater detail all approaches to the subtasks and provided an evaluation of their performance, dependability on various annotation layers and most importantly flexibility for new languages.

We found evidence that each of the given subtasks can be approached by information-poor machine learning approaches that rely solely on POS information – mention detection can be easily applied to new languages when the mdIOBA method is made use of; mention head detection also showed itself to be a task that machine learning in the form of the mhdML approach could easily solve when mention heads are also part of the coreference annotation layer; with respect to feature selection, we proposed a POS-based set of features that can be applied to any new language when this layer of annotation is provided. In other words, for each of the main tasks of MCR, we searched for and accordingly developed a well performing machine learning solution. The latter relies solely on POS information which provided us with a pipeline

that is competitive to the state-of-the-art approaches to coreference resolution. Our approaches can be applied to any language for which part-of-speech annotations are provided, with the latter being the most widely used and accessible annotation layer across all datasets.

The investigation we carried out, reveals information, not only important to approaches based on the discussed framework, but to all coreference resolution enterprises that target more than one language at a time and especially languages that do not have an abundant collection of linguistic tools and resources to enrich the employed datasets with additional annotation layers. Our findings were supported by detailed evaluations, analysis and comparison across the various methods, languages and settings. Additionally, for each of the subtasks we proposed a language independent solution that constructed a MCR pipeline that can be used as a guideline for systems of similar kinds. We did not focus on achieving best system performance, but rather on exploring the possibilities for recasting the task on a multilingual level. The system results we reported, confirmed that language independent approaches do not always reach best performance but are still highly competitive with the state-of-the-art. Another drawback of the language independent pipeline we assembled is the need for mention head information in order for an ML classifier to be trained.

Being such an exceptionally complex task, multilingual coreference resolution offers numerous possibilities for further research and advancement with respect to all its subtasks. Some of these challenging opportunities are presented in the following section.

8.2 OPEN PROBLEMS FOR MULTILINGUAL CR

Multilinguality is one necessary step for CR which needs to be payed more attention to in the coming years. In our work we discussed various open problems, such as the lack of uniformity of the annotation standards and schemes across the languages and even the lack of some annotation layers for less resourced languages. We believe that this issue will not be easily solved in the next decade, since the coordination and standardization of diverse annotation layers is a complex task that requires a lot of manual exploration and effort.

Multilinguality is a factor that also challenges the competitiveness of the CR pipeline. The lack of language specific adaptation of the employed approaches reduces the overall system performance. This is one disadvantage of MCR that will be the main subject of investigation in the field in the following years. However, the general performance of the MCR pipeline could be improved by the introduction of easily available world knowledge that is attainable on a language independent manner (e.g. by a direct extraction or computation from the World Wide Web (WWW)). In section 8.3.1, we propose a way in which this can be accomplished.

In our work, we presented a truly language independent set of features that can be collected for the coreference resolver. Yet, this set is applicable for information-poor approaches to the task. We believe that the creation of a language independent feature set for information-rich approaches to MCR is and will be impossible. Additional features can be added if standardized annotation schemes are employed. However, information-rich approaches normally include knowledge about highly language specific phenomena. Features representing such knowledge will be helpful only for the language for which they are developed and will distort the informativeness of the feature set for the languages for which they do not apply. An approach that offers only a partial solution to the problem is discussed in section 8.3.2.

8.3 FUTURE DIRECTIONS FOR MULTILINGUAL CR

The current section proposes two ways in which MCR could be directed so that enhanced and still multilingual system behaviour can be achieved. One such direction is the exploration of other forms and techniques of representation of world knowledge within the information used by the system (section 8.3.1). In section 8.3.2, we suggest a possible generalization of the features or feature groups that is not based on the specific language in use, but rather on the language family that it is part of.

8.3.1 *World Knowledge and Coreference Resolution*

In chapter 7, we discussed the applicability of ontological information to the MCR task and noted that the mere language dependability and variation in coverage renders ontologies as hardly suitable for this type of enterprise. However, ontological information has been shown to carry important world knowledge for numerous NLP tasks, such as word sense disambiguation, information extraction, question answering, etc. Nevertheless, world knowledge can also be represented and structured in other forms and extracted from different sources.

term co-occurrence
co-occurrence

Keeping this in mind, we propose that **term co-occurrence** or as well only **co-occurrence**, with the WWW as a search space, is examined as a potential possibility to provide a large-scale source of world knowledge for multilingual coreference resolution. Term co-occurrence can be used to enrich the features describing each mention pair with information about the **semantic proximity**, or in other words the **semantic similarity**, of the syntactic heads of the mentions.

semantic proximity
semantic similarity

For example, let us consider the two sentences given in example (46).

(46) Mary baked a dessert for the party. The best pie I have every tried!

The possible mention pairs that can be constructed from this example are listed below:

1. Mary dessert
2. Mary theater
3. Mary pie
4. dessert party
5. dessert pie
6. party pie

Our assumption is that with the use of term co-occurence within the WWW search space, the semantic proximity between the tokens of the given mention pairs can be established. The knowledge that can be gained merely describes the correlation between the occurrences of these terms together in a document in the web. With respect to coreference resolution, this information can be directly included into the feature representation of the mention pair depicting real counts of their co-occurence. These figures do not necessarily mean that the mentions are either coreferent or not. However, mentions that never or rarely occur together (such as *dessert* and *party*) will seldom be actually coreferent. On the other hand, the co-occurence of *dessert* and *pie* will be higher, telling us that there is a higher chance that this mention pair represents coreferent mentions.

8.3.2 *Coreference Resolution Across Language Families*

All three subtasks of the multilingual coreference resolution process showed themselves to have various language specific issues, such as the immensely increased number of mentions for the Arabic language, the inconsistent head directionality for English or the distinct behaviour of Arabic in the feature selection investigation that we conducted. For all these problems, we proposed, implemented and evaluated language independent solutions, which are based on the POS tagsets that the datasets included. However, the introduced POS-based feature set was shown to lead to an overall reduction of system performance of 2.23 percent points. We also showed that feature selection is a highly time-consuming task and thus in state-of-the-art research, systems are mostly optimized for only one language. However, in the multilingual environment in which we situated our framework, we are interested in a possibility that would allow us to optimize the system performance in a more efficient manner than the language specific approach that we presented in chapter 7.

typology
typological classification

Typology or **typological classification** makes use of morphological, phonological, syntactic, and semantic similarities of various languages of the world in order to group them in language types. In general, language families include languages that are typologically similar. For this reason a well-motivated direction for future work is the investigation of the hypothesis that typological

information is beneficial for the design of multilingual coreference resolution solutions. This should potentially provide a deeper insight about the intersection and dependencies between the MCR task and various language family properties.

However, an examination of multilingual coreference resolution system optimization around the typological differences of various language families can only be conducted if several typologically distant language families are included in the observation. Furthermore, each family must be represented by more than one family member so that a generalization over the family can be made. Unfortunately, there are, as yet, no resources of that scale and size in order for an investigation of the hypothesis to be carried out.

8.4 ADVANCES IN MULTILINGUAL COREFERENCE RESOLUTION

The current work showed that nowadays flexibility and multilinguality of NLP areas, such as CR, are important issues that can be tackled when machine learning approaches are explored. We believe that the new direction of the field will provide a bridge between language specific methods that were previously used and innovative and adaptable solutions that can be applied across languages. Various topics need to be addressed, such as the availability and uniformity of resources and annotation schemes on a multilingual level as well as the introduction of world knowledge from cross-lingual resources.

The advancement of MCR is also the main topic of The first international workshop on Advances in Multilingual Coreference resolution (AMCR 2013)[1], which will be held on September 12th/13th, 2013 at the International Conference on Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria. We believe that this enterprise will provide the possibility for in depth exploration of multilinguality in CR and will enable the exchange of knowledge and experience on a cross-lingual level.

8.5 CONCLUSION

The current work presented a thorough exploration of the possibility to recast the complex coreference resolution task at a new, even more demanding level – multilinguality. We carried out an examination of the issues that arise when multilingual coreference resolution is considered in the framework of the UBIU multilingual coreference resolution system. The framework we employed uses the mention-pair coreference model and machine learning in the form of memory-based learning for the resolution process.

Our analysis covered the three aspects essentially important to the framework: mention detection, mention head detection and feature selection. We reviewed the problems that multilinguality faces on each of these levels and

[1] http://cl.indiana.edu/~zhekova/amcr/2013

proposed language independent solutions for each of them within the pipeline. Additionally, we evaluated all methods and approaches across the datasets distributed from the two multilingual shared tasks that have been organized so far: SemEval-2 and CoNLL-2012.

We showed that multilinguality can be achieved with an information-poor approach that requires only the POS annotation layer. We discussed the improvements or the changes to the pipeline that can be approached when other layers of annotations are provided. Furthermore, we compared the proposed rule-based or heuristic methods to language independent machine learning solutions. Eight different languages (Arabic, Catalan, Chinese, Dutch, English, German, Italian and Spanish) were included in the analysis. According to our knowledge, an exploration of multilinguality with respect to coreference resolution in such detail has not been presented so far in state-of-the-art research.

The most important finding of the current work is the fact that coreference resolution can be taken to a multilingual and even language independent level. On the way, we described the necessary steps needed for this transition within the used framework. The newly gained knowledge can also be used for CR problems employing other types of frameworks when there are areas of common ground, such as mention detection for example, which is a main part of most types of state-of-the-art coreference resolution systems.

Part V

APPENDIX

APPENDIX

A

DATA EXCERPTS

A.1 SEMEVAL-2

A.1.1 *Catalan*

ID	TOKEN	LEMMA	POS	FEAT	HEAD	DEPREL	NE	PRED	APRED	COREF
1	El	el	d	postype=article|gen=m|num=s	2	spec	_	_	_	(34|(8
2	paquet	paquet	n	postype=common|gen=m|num=s	15	suj	_	_	arg1-tem	_
3	de	de	s	postype=preposition	2	sp	_	_	_	_
4	Ducados	Ducados	n	postype=proper	3	sn	(org	_	_	(2
5	,	,	f	punct=comma	7	f	(org	_	_	_
6	la	el	d	postype=article|gen=f|num=s	7	spec	_	_	_	_
7	marca	marca	n	postype=common|gen=f|num=s	4	sn	_	_	_	_
8	lider	lider	a	postype=qualificative|gen=c|num=s	7	s.a	_	_	_	_
9	en	en	s	postype=preposition	7	sp	_	_	_	_
10	tabac	tabac	n	postype=common|gen=m|num=s	9	sn	_	_	_	(12
11	negre	negre	a	postype=qualificative|gen=m|num=s	10	s.a	_	_	_	12)
12	d'	de	s	postype=preposition	7	sp	_	_	_	_
13	Altadis	Altadis	n	postype=proper	12	sn	(org)	_	_	(1)
14	,	,	f	punct=comma	7	f	org)|org)	_	_	2)|34)
15	passa	passar	v	postype=main|num=s|person=3|mood=indicative|tense=present	0	sentence	_	passar.b2	_	_
16	de	de	s	postype=preposition	15	cc	_	_	arg3-ein	_
17	les	el	d	postype=article|gen=f|num=p	19	spec	(number	_	_	(35
18	220	220	z	_	17	z	_	_	_	_
19	pessetes	pesseta	z	postype=currency	16	sn	_	_	_	_
20	actuals	actual	a	postype=qualificative|gen=c|num=p	19	s.a	number)	_	_	35)
21	a	a	s	postype=preposition	15	creg	_	_	arg4-efi	_
22	les	el	d	postype=article|gen=f|num=p	23	spec	(number	_	_	(36
23	225	225	z	_	21	sn	number)	_	_	36)
24	.	.	f	punct=period	15	f	_	_	_	8)

Table A.1: Excerpt from the Catalan SemEval-2 training data set (*gold* annotations).

A.1.2 *Dutch*

ID	TOKEN	PLEMMA	PPOS	PFEAT	PHEAD	PDEPREL	PNE	PPRED	COREF
1	Tijdens	Tijdens	VZ	position=initial	7	mod	_	_	_
2	de	de	LID	type=definite|case=standard	3	det	_	_	_
3	voorstelling	voorstelling	N	type=common|num=singular|degree=basis|gender=masc|case=standard	1	obj1	_	_	_
4	van	van	VZ	position=initial	3	mod	_	_	_
5	de	de	LID	type=definite|case=standard	6	det	_	_	(84
6	resultaten	resultaat	N	type=common|num=plural|degree=basis	4	obj1	_	_	84)
7	belicht	belichten	WW	mood=finiteform|tense=present|number=singular	0	ROOT	_	_	_
8	professor	professor	N	type=common|num=singular|degree=basis|gender=masc|case=standard	7	su	_	_	(8
9	Michel	Michel	SPEC	_	8	app	PER	_	_
10	Poulain	Poulain	SPEC	_	9	mwp	MISC	_	8)
11	(	(	LET	_	10	punct	_	_	_
12	UCL	UCL	N	type=name|num=singular|degree=basis|gender=masc|case=standard	8	mod	ORG	_	(9)
13	)	)	LET	_	12	punct	_	_	_
14	de	de	LID	type=definite|case=standard	15	det	_	_	_
15	mogelijkheid	mogelijkheid	N	type=common|num=singular|degree=basis|gender=masc|case=standard	7	obj1	_	_	_
16	om	om	VZ	position=initial	15	vc	_	_	_
17	de	de	LID	type=definite|case=standard	18	det	_	_	_
18	pensioenleeftijd	pensioenleeftijd	N	type=common|num=singular|degree=basis|gender=masc|case=standard	21	ld	_	_	_
19	op	op	VZ	position=final	21	svp	_	_	_
20	te	te	VZ	position=initial	16	body	_	_	_
21	trekken	trekken	WW	mood=infinitive|position=free	20	body	_	_	_
22	.	.	LET	_	21	punct	_	_	_

Table A.2: Excerpt from the Dutch SEMEVAL-2 training data set (*auto* annotations).

A.1.3 *English*

ID	TOKEN	LEMMA	POS	FEAT	HEAD	DEPREL	NE	PRED	APRED	COREF
1	Between	_	IN	IN	13	TMP	_	_	argM-tmp	_
2	now	_	RB	RB	1	PMOD	_	_	_	_
3	and	_	CC	CC	2	COORD	_	_	_	_
4	election	_	NN	NN	5	NMOD	(date	_	_	(176
5	day	_	NN	NN	3	CONJ	date)	_	_	176)
6	,	_	,	,	13	P	_	_	_	_
7	a	_	DT	DT	8	NMOD	_	_	_	(204
8	team	_	NN	NN	13	SBJ	_	_	arg0	_
9	of	_	IN	IN	8	NMOD	_	_	_	_
10	NBC	_	NNP	NNP	11	NAME	(org	_	_	(202|(34
11	News	_	NNP	NNP	12	NMOD	org)	_	_	34)
12	correspondents	_	NNS	NNS	9	PMOD	_	_	_	202)|204)
13	will	_	MD	MD	0	sentence	_	_	argM-mod	_
14	review	_	VB	VB	13	VC	_	review.01	_	_
15	what	_	WP	WP	19	OBJ	_	_	_	_
16	the	_	DT	DT	18	NMOD	_	_	_	(9
17	presidential	_	JJ	JJ	18	NMOD	_	_	_	_
18	candidates	_	NNS	NNS	19	SBJ	_	_	_	9)
19	say	_	VBP	VBP	22	SUB	_	say.01	_	_
20	to	_	TO	TO	19	ADV	_	_	_	_
21	s	_	PRP	PRP	20	PMOD	_	_	_	(237)
22	if	_	IN	IN	14	OBJ	_	_	arg1	_
23	they	_	PRP	PRP	24	SBJ	_	_	_	(9)
24	are	_	VBP	VBP	22	SUB	_	be.03	_	_
25	telling	_	VBG	VBG	24	VC	_	tell.01	_	_
26	the	_	DT	DT	27	NMOD	_	_	_	(245
27	truth	_	NN	NN	25	OBJ	_	_	_	245)
28	.	_	.	.	13	P	_	_	_	_

Table A.3: Excerpt from the English SemEval-2 training data set (*gold* annotations).

A.1.4 *German*

ID	TOKEN	LEMMA	POS	FEAT	HEAD	DEPREL	NE	PRED	COREF
1	Mein	mein	PPOSAT	_	2	DET	_	_	(22|3)
2	Freund	Freund	NN	cas=n|num=sg|gend=masc	5	SUBJ	_	_	_
3	zum	zu	APPRART	_	2	PP	_	_	_
4	Beispiel	Beispiel	NN	cas=d|num=sg|gend=neut	3	PN	_	_	(23)|22)
5	wuerde	werden	VAFIN	_	0	ROOT	_	_	_
6	schreiben	schreiben	VVINF	_	5	AUX	_	_	_
7	:	:	$.	_	6	-PUNCT-	_	_	_
8	Sehr	sehr	ADV	_	9	ADV	_	_	(17
9	geehrter	geehrt	ADJA	cas=n|num=sg|gend=masc	10	ATTR	_	_	_
10	Herr	Herr	NN	cas=n|num=sg|gend=masc	0	ROOT	_	_	_
11	Stroebele	Stroebele	NE	cas=n|num=sg|gend=masc	10	APP	_	_	17)
12	,	,	$,	_	11	-PUNCT-	_	_	_
13	Ihre	ihre	PPOSAT	_	14	DET	_	_	(9|(17)
14	Partei	Partei	NN	cas=n|num=sg|gend=fem	15	SUBJ	_	_	9)
15	hat	haben	VAFIN	_	0	ROOT	_	_	_
16	von	von	APPR	_	25	PP	_	_	_
17	mir	mir	PPER	cas=d|num=sg|gend=*|per=1	16	PN	_	_	(22)
18	ein	ein	ART	cas=a|num=sg|gend=neut	19	DET	_	_	(24
19	Mandat	Mandat	NN	cas=a|num=sg|gend=neut	25	OBJA	_	_	_
20	fuer	fuer	APPR	_	19	PP	_	_	_
21	eine	eine	ART	cas=a|num=sg|gend=fem	22	DET	_	_	(25
22	Friedenspolitik	Friedenspolitik	NN	cas=a|num=sg|gend=fem	20	PN	_	_	_
23	ohne	ohne	APPR	_	22	PP	_	_	_
24	Krieg	Krieg	NN	cas=a|num=sg|gend=masc	23	PN	_	_	(26)|25)|24)
25	erhalten	erhalten	VVPP	_	15	AUX	_	_	_
26	.	.	$.	_	25	-PUNCT-	_	_	_

Table A.4: Excerpt from the German SEMEVAL-2 training data set (*gold* annotations).

A.1.5 *Italian*

ID	TOKEN	LEMMA	POS	FEAT	HEAD	DEPREL	NE	PRED	COREF
1	Storia	storiare	VI	_	16	RMOD	_	_	(22)
2	Il	il	RS	_	3	DET	_	_	(23
3	primo	primo	SS	_	1	SUBJ	_	_	_
4	ad	ad	E	_	3	RMOD	_	_	_
5	appellarsi	appellare/si	VF+E	_	4	PN	_	_	_
6	ad	ad	E	_	5	RMOD	_	_	_
7	un	un	RS	_	8	DET	_	_	(1
8	concilio	concilio	SS	_	6	PN	_	_	_
9	che	che	CCHE	_	10	SUBJ	_	_	_
10	dirimesse	dirimere	VI	_	8	RELCL	_	_	_
11	il	il	RS	_	13	DET	_	_	(24
12	suo	suo	DS	_	13	RMOD	_	_	(25)
13	contrasto	contrasto	SS	_	10	DOBJ	_	_	_
14	col	con	ES	_	10	RMOD	_	_	(26
15	papa	papa	SS	_	14	PN	_	_	26)|24)|1)|23)
16	fu	essere	VI	_	0	ROOT	_	_	_
17	Lutero	lutero	SPN	_	16	PRED	_	_	(25)
18	,	,	XPW	_	17	SEPARATOR	_	_	_
19	già	già	SPN	_	17	APPOSITION	_	_	_
20	nel	in	ES	_	16	RMOD	_	_	(27
21	1518	1518	N	_	20	PN	_	_	27)
22	:	colon	XPS	_	16	END	_	_	_

Table A.5: Excerpt from the Italian SemEval-2 training data set (*auto* annotations).

A.1.6 *Spanish*

ID	TOKEN	LEMMA	POS	FEAT	HEAD	DEPREL	NE	PRED	COREF
1	Respecto_al	respecto_al	s	postype=preposition|gen=m|num=s|contracted=yes	9	ao	_	_	_
2	resultado	resultado	n	postype=common|gen=m|num=s	1	sn	_	_	(19
3	de	de	s	postype=preposition	2	sp	_	_	_
4	las	el	d	postype=article|gen=f|num=p	5	spec	_	_	(14
5	elecciones	eleccian	n	postype=common|gen=f|num=p	3	sn	_	_	_
6	autonamicas	autonamico	a	postype=qualificative|gen=f|num=p	5	s.a	_	_	14)|19)
7	,	,	f	punct=comma	1	f	_	_	_
8	Villalobos	Villalobos	n	postype=proper	9	suj	_	_	(1)
9	senala	senalar	v	postype=main|gen=c|num=s|person=3|mood=indicative|tense=past	0	sentence	_	senalar.a31	_
10	que	que	p	postype=relative|gen=c|num=c	12	conj	_	_	_
11	_	_	p	_	12	suj	_	_	(19)
12	demuestra	demostrar	v	postype=main|gen=c|num=s|person=3|mood=indicative|tense=present	9	cd	_	demostrar.a2	_
13	que	que	p	postype=relative|gen=c|num=c	17	conj	_	_	_
14	"	"	f	punct=quotation	17	f	_	_	_
15	Andalucia	Andalucia	n	postype=proper	17	suj	_	_	(15)
16	no	no	r	postype=negative	17	mod	_	_	_
17	es	ser	v	postype=semiauxiliary|gen=c|num=s|person=3|mood=indicative|tense=present	12	cd	_	ser.c2	_
18	propiedad	propiedad	n	postype=common|gen=f|num=s	17	atr	_	_	_
19	del	del	s	postype=preposition|gen=m|num=s|contracted=yes	18	sp	_	_	_
20	PSOE	PSOE	n	postype=proper	19	sn	_	_	(38)
21	,	,	f	punct=comma	23	f	_	_	_
22	_	_	p	_	23	suj	_	_	(15)
23	es	ser	v	postype=semiauxiliary|gen=c|num=s|person=3|mood=indicative|tense=present	18	S	_	ser.c2	_
24	propiedad	propiedad	n	postype=common|gen=f|num=s	23	atr	_	_	_
25	de	de	s	postype=preposition	24	sp	_	_	_
26	los	el	d	postype=article|gen=m|num=p	27	spec	_	_	(39
27	andaluces	andaluz	a	postype=qualificative|gen=m|num=p	25	sn	_	_	39)
28	"	"	f	punct=quotation	17	f	_	_	_
29	.	.	f	punct=period	9	f	_	_	_

Table A.6: Excerpt from the Spanish SemEval-2 training data set (*auto* annotations).

A.2 CONLL 2012

A.2.1 *Arabic*

Word#	Word	POS	ParseBit	PredLemma	PFID	WS	SA	NE	Coref
0	650#DEFAULT#650#650	NOUN_NUM	(TOP(FRAG(NP*	DEFAULT	-	-	-	(CARDINAL)	(10
1	بِـ#بِ#junodiy #jndyAF#junodiy +AF	NOUN+CASE_INDEF_ACC	(NP*	junodiy	-	-	-	*	-
2	أَمِيرْكِيّا#>myrkyAF#>amiyrokiy +AF	ADJ+CASE_INDEF_ACC	*))	>amoriykiy	-	-	-	(NORP)	-
3	لِـ#<ilaY#<iY#<ilaY	PREP	(PP*	<ilaY	-	-	-	*	-
4	فِلَسْطِينَ#fiyliyb iyn#Alfiylybyn#Al+fiyliyb iyn	DET+NOUN_PROP	(NP*))	fiyliyb iyn	-	-	-	(GPE)	(5))
5	مِنْ#min#mn#min	PREP	(PP*	min	-	-	-	*	-
6	يَوْمِ#yawom#Alywm#Al+yawom+i	DET+NOUN+CASE_DEF_GEN	(NP*))	yawom	-	1	-	(DATE)	-
7	فِي#fiy#fy#fiy	PREP	(PP*	fiy	-	-	-	*	-
8	بِضْعِ#baEovap#bEvp#biEov+ap+K	NOUN+NSUFF_FEM_SG+CASE_INDEF_GEN	(NP(NP*	baEovap	-	1	-	*	(2
9	سَرِيعٍ#tadoriybiy #drybyp#tadoriybiy +ap+K	ADJ+NSUFF_FEM_SG+CASE_INDEF_GEN	*)	tadoriybiy	-	-	-	*	-
10	الـ#clitics#il#Il-	PREP	(PP*	clitics	-	-	-	*	-
11	لِقَدْ#qaDA'#AlqDA'#-Al+qaDA'+i	DET+NOUN+CASE_DEF_GEN	(NP(NP*)	qaDA'	-	-	-	*	-
12	لَ#EalaY#ElY#EalaY	PREP	(PP*	EalaY	-	-	-	*	-
13	,#DEFAULT#'#"	PUNC	(NP*	DEFAULT	-	-	-	(ORG*	(3
14	أَبُو#>abuw#>bw#>abuw	NOUN_PROP	*	>abuw	-	-	-	*	-
15	سَعِيدٍ#say Af#syAf#say Af	NOUN_PROP	*	say Af	-	-	-	*	-
16	,#DEFAULT#'#"	PUNC	*))))))	DEFAULT	-	-	-	*)	3)(2)

Table A.7: Excerpt from the Arabic CoNLL 2012 training data set (*gold* annotations).

A.2.2 *English*

Word#	Word	POS	ParseBit	PredLemma	PFID	WS	SA	NE	PredArgs	PredArgs	Coref
0	It	PRP	(TOP(S(NP*)	-	-	-	-	Speaker#1 *	*	(ARG1*)	(22)
1	is	VBZ	(VP*	-	03	-	-	Speaker#1 *	(V*)	*	-
2	composed	VBN	(VP*	-	01	2	-	Speaker#1 *	*	(V*)	-
3	of	IN	(PP*	-	-	-	-	Speaker#1 *	*	(ARG2*	-
4	a	DT	(NP(NP*	-	-	-	-	Speaker#1 *	*	*	(24
5	primary	JJ	*	-	-	-	-	Speaker#1 *	*	*	-
6	stele	NN	*)	-	-	-	-	Speaker#1 *	*	*	24)
7	,	,	*	-	-	-	-	Speaker#1 *	*	*	-
8	secondary	JJ	(NP*	-	-	-	-	Speaker#1 *	*	*	(13
9	steles	NNS	*)	-	-	-	-	Speaker#1 *	*	*	13)
10	,	,	*	-	-	-	-	Speaker#1 *	*	*	-
11	a	DT	(NP*	-	-	-	-	Speaker#1 *	*	*	-
12	huge	JJ	*	-	-	-	-	Speaker#1 *	*	*	-
13	round	NN	*	-	-	-	-	Speaker#1 *	*	*	-
14	sculpture	NN	(NML(NML*)	-	-	-	-	Speaker#1 *	*	*	-
15	and	CC	*	-	-	-	-	Speaker#1 *	*	*	-
16	beacon	NN	(NML*	-	-	-	-	Speaker#1 *	*	*	-
17	tower	NN	*)))	-	-	-	-	Speaker#1 *	*	*	-
18	,	,	*	-	-	-	-	Speaker#1 *	*	*	-
19	and	CC	*	-	-	-	-	Speaker#1 *	*	*	-
20	the	DT	(NP*	-	-	-	-	Speaker#1 (WORK_OF_ART*	*	*	-
21	Great	NNP	*	-	-	-	-	Speaker#1 *	*	*	-
22	Wall	NNP	*)	-	-	-	-	Speaker#1 *)	*	*	-
23	,	,	*	-	-	-	-	Speaker#1 *	*	*	-
24	among	IN	(PP*	-	-	-	-	Speaker#1 *	*	*	-
25	other	JJ	(NP*	-	-	-	-	Speaker#1 *	*	*	-
26	things	NNS	*))))))	-	-	-	-	Speaker#1 *	*	*)	-
27	.	.	*))	-	-	-	-	Speaker#1 *	*	*	-

Table A.8: Excerpt from the English CoNLL 2012 training data set (*auto* annotations).

A.2.3 *Chinese*

Word#	Word	POS	ParseBit	PredLemma	PFID	WS	SA	NE	PredArgs	Coref
0	这	DT	(TOP(IP(NP(DP*	-	-	-	Speaker#1	*	(ARG0*	(3
1	场	M	(CLP*))	-	-	-	Speaker#1	*	*	-
2	战役	NN	(NP*))	-	-	-	Speaker#1	*	*)	3)
3	打破	VV	(VP(VP*	-	01	-	Speaker#1	*	(V*)	-
4	了	AS	*	-	-	-	Speaker#1	*	*	-
5	日军	NN	(NP(NP*)	-	-	-	Speaker#1	(ORG)	(ARG1*	(32)
6	对	P	(DNP(PP*	-	-	3	Speaker#1	*	*	-
7	敌后	JJ	(NP(ADJP*)	-	-	-	Speaker#1	*	*	-
8	根据地	NN	(NP*)))	-	-	-	Speaker#1	*	*	-
9	的	DEG	*)	-	-	-	Speaker#1	*	*	-
10	封锁	NN	(NP*)))	-	-	-	Speaker#1	*	*)	-
11	、	PU	*	-	-	-	Speaker#1	*	*	-
12	振奋	VV	(VP*	-	01	-	Speaker#1	*	*	-
13	全	DT	(NP(NP(DP*)	-	-	2	Speaker#1	*	*	(66|(68
14	国	NN	(NP*))	-	-	-	Speaker#1	*	*	66)
15	人民	NN	(NP*))	-	-	-	Speaker#1	*	*	68)
16	抗日	NN	(NP*	-	-	-	Speaker#1	(GPE)	*	-
17	精神	NN	*)))	-	-	-	Speaker#1	*	*	-
18	、	PU	*	-	-	-	Speaker#1	*	*	-
19	影响	VV	(VP*	-	01	-	Speaker#1	*	*	-
20	了	AS	*	-	-	-	Speaker#1	*	*	-
21	世界	NN	(NP(DNP(NP(NP(NP*)	-	-	-	Speaker#1	(EVENT*	*	-
22	人民	NN	(NP*))	-	-	-	Speaker#1	*	*	-
23	反法西斯	JJ	(ADJP*)	-	-	-	Speaker#1	*	*	-
24	战争	NN	(NP*))	-	-	-	Speaker#1	*)	*	-
25	的	DEG	*)	-	-	-	Speaker#1	*	*	-
26	形势	NN	(NP*)))))	-	-	-	Speaker#1	*	*	-
27	。	PU	*))	-	-	-	Speaker#1	*	*	-

Table A.9: Excerpt from the Chinese CoNLL 2012 training data set (*gold* annotations).

APPENDIX

B

OFFICIAL SHARED TASK RESULTS

B.1 SEMEVAL-2

B.1.1 *Catalan*

	Mention detection			CEAF			MUC			B^3			BLANC		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
closed×gold															
RelaxCor	100	100	100	70.5	70.5	70.5	29.3	77.3	42.5	68.6	95.8	79.9	56.0	81.8	59.7
SUCRE	100	100	100	68.7	68.7	68.7	54.1	58.4	56.2	76.6	77.4	77.0	72.4	60.2	63.6
TANL-1	100	96.8	98.4	66.0	63.9	64.9	17.2	57.7	26.5	64.4	93.3	76.2	52.8	79.8	54.4
UBIU	75.1	96.3	84.4	46.6	59.6	52.3	8.8	17.1	11.7	47.8	76.3	58.8	51.6	57.9	52.2
closed×regular															
SUCRE	75.9	64.5	69.7	51.3	43.6	47.2	44.1	32.3	37.3	59.6	44.7	51.1	53.9	55.2	54.2
TANL-1	83.3	82.0	82.7	57.5	56.6	57.1	15.2	46.9	22.9	55.8	76.6	64.6	51.3	76.2	51.0
UBIU	51.4	70.9	59.6	33.2	45.7	38.4	6.5	12.6	8.6	32.4	55.7	40.9	50.2	53.7	47.8
open×gold															
open×regular															

Table B.1: The official results from the SEMEVAL-2 shared task for Catalan [Recasens et al., 2010].

B.1.2 *Dutch*

	Mention detection			CEAF			MUC			B^3			BLANC		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
closed×gold															
SUCRE	100	100	100	58.8	58.8	58.8	65.7	74.4	69.8	65.0	69.2	67.0	69.5	62.9	65.3
closed×regular															
SUCRE	78.0	29.0	42.3	29.4	10.9	15.9	62.0	19.5	29.7	59.1	6.5	11.7	46.9	46.9	46.9
UBIU	41.5	29.9	34.7	20.5	14.6	17.0	6.7	11.0	8.3	13.3	23.4	17.0	50.0	52.4	32.3
open×gold															
open×regular															

Table B.2: The official results from the SEMEVAL-2 shared task for Dutch [Recasens et al., 2010].

B.1.3 *English*

	Mention detection			CEAF			MUC			B^3			BLANC		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
closed×gold															
RelaxCor	100	100	100	75.6	75.6	75.6	21.9	72.4	33.7	74.8	97.0	84.5	57.0	83.4	61.3
SUCRE	100	100	100	74.3	74.3	74.3	68.1	54.9	60.8	86.7	78.5	82.4	77.3	67.0	70.8
TANL-1	99.8	81.7	89.8	75.0	61.4	67.6	23.7	24.4	24.0	74.6	72.1	73.4	51.8	68.8	52.1
UBIU	92.5	99.5	95.9	63.4	68.2	65.7	17.2	25.5	20.5	67.8	83.5	74.8	52.6	60.8	54.0
closed×regular															
SUCRE	78.4	83.0	80.7	61.0	64.5	62.7	57.7	48.1	52.5	68.3	65.9	67.1	58.9	65.7	61.2
TANL-1	79.6	68.9	73.9	61.7	53.4	57.3	23.8	25.5	24.6	62.1	60.5	61.3	50.9	68.0	49.3
UBIU	66.7	83.6	74.2	48.2	60.4	53.6	11.6	18.4	14.2	50.9	69.2	58.7	50.9	56.3	51.0
open×gold															
Corry-B	100	100	100	77.5	77.5	77.5	56.1	57.5	56.8	82.6	85.7	84.1	69.3	75.3	71.8
Corry-C	100	100	100	77.7	77.7	77.7	57.4	58.3	57.9	83.1	84.7	83.9	71.3	71.6	71.5
Corry-M	100	100	100	73.8	73.8	73.8	62.5	56.2	59.2	85.5	78.6	81.9	76.2	58.8	62.7
RelaxCor	100	100	100	75.8	75.8	75.8	22.6	70.5	34.2	75.2	96.7	84.6	58.0	83.8	62.7
open×regular															
BART	76.1	69.8	72.8	70.1	64.3	67.1	62.8	52.4	57.1	74.9	67.7	71.1	55.3	73.2	57.7
Corry-B	79.8	76.4	78.1	70.4	67.4	68.9	55.0	54.2	54.6	73.7	74.1	73.9	57.1	75.7	60.6
Corry-C	79.8	76.4	78.1	70.9	67.9	69.4	54.7	55.5	55.1	73.8	73.1	73.5	57.4	63.8	59.4
Corry-M	79.8	76.4	78.1	66.3	63.5	64.8	61.5	53.4	57.2	76.8	66.5	71.3	58.5	56.2	57.1

Table B.3: The official results from the SEMEVAL-2 shared task for English [Recasens et al., 2010].

B.1.4 *German*

	Mention detection			CEAF			MUC			B^3			BLANC		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
closed×gold															
SUCRE	100	100	100	72.9	72.9	72.9	74.4	48.1	58.4	90.4	73.6	81.1	78.2	61.8	66.4
TANL-1	100	100	100	77.7	77.7	77.7	16.4	60.6	25.9	77.2	96.7	85.9	54.4	75.1	57.4
UBIU	92.6	95.5	94.0	67.4	68.9	68.2	22.1	21.7	21.9	73.7	77.9	75.7	60.0	77.2	64.5
closed×regular															
SUCRE	79.3	77.5	78.4	60.6	59.2	59.9	49.3	35.0	40.9	69.1	60.1	64.3	52.7	59.3	53.6
TANL-1	60.9	57.7	59.2	50.9	48.2	49.5	10.2	31.5	15.4	47.2	54.9	50.7	50.2	63.0	44.7
UBIU	50.6	66.8	57.6	39.4	51.9	44.8	9.5	11.4	10.4	41.2	53.7	46.6	50.2	54.4	48.0
open×gold															
BART	94.3	93.7	94.0	67.1	66.7	66.9	70.5	40.1	51.1	85.3	64.4	73.4	65.5	61.0	62.8
open×regular															
BART	82.5	82.3	82.4	61.4	61.2	61.3	61.4	36.1	45.5	75.3	58.3	65.7	55.9	60.3	57.3

Table B.4: The official results from the SemEval-2 shared task for German [Recasens et al., 2010].

B.1.5 *Italian*

	Mention detection			CEAF			MUC			B^3			BLANC		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
closed×gold															
SUCRE	98.4	98.4	98.4	66.0	66.0	66.0	48.1	42.3	45.0	76.7	76.9	76.8	54.8	63.5	56.9
closed×regular															
SUCRE	84.6	98.1	90.8	57.1	66.2	61.3	50.1	50.7	50.4	63.6	79.2	70.6	55.2	68.3	57.7
UBIU	46.8	35.9	40.6	37.9	29.0	32.9	2.9	4.6	3.6	38.4	31.9	34.8	50.0	46.6	37.2
open×gold															
open×regular															
BART	42.8	80.7	55.9	35.0	66.1	45.8	35.3	54.0	42.7	34.6	70.6	46.4	57.1	68.1	59.6
TANL-1	90.5	73.8	81.3	62.2	50.7	55.9	37.2	28.3	32.1	66.8	56.5	61.2	50.7	69.3	48.5

Table B.5: The official results from the SemEval-2 shared task for Italian [Recasens et al., 2010].

B.1.6 *Spanish*

	Mention detection			CEAF			MUC			B^3			BLANC		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
closed×gold															
RelaxCor	100	100	100	66.6	66.6	66.6	14.8	73.8	24.7	65.3	97.5	78.2	53.4	81.8	55.6
SUCRE	100	100	100	69.8	69.8	69.8	52.7	58.3	55.3	75.8	79.0	77.4	67.3	62.5	64.5
TANL-1	100	96.8	98.4	66.9	64.7	65.8	16.6	56.5	25.7	65.2	93.4	76.8	52.5	79.0	54.1
UBIU	73.8	96.4	83.6	45.7	59.6	51.7	9.6	18.8	12.7	46.8	77.1	58.3	52.9	63.9	54.3
closed×regular															
SUCRE	74.9	66.3	70.3	56.3	49.9	52.9	35.8	36.8	36.3	56.6	54.6	55.6	52.1	61.2	51.4
TANL-1	82.2	84.1	83.1	58.6	60.0	59.3	14.0	48.4	21.7	56.6	79.0	66.0	51.4	74.7	51.4
UBIU	51.1	72.7	60.0	33.6	47.6	39.4	7.6	14.4	10.0	32.8	57.1	41.6	50.4	54.6	48.4
open×gold															
open×regular															

Table B.6: The official results from the SEMEVAL-2 shared task for Spanish [Recasens et al., 2010].

B.2 CONLL 2011

Participant	Official Score					
	CT/PM	OT/PM	CT/GB	OT/GB	CT/GM	OT/GM
[Lee et al., 2011]	57.79	58.31	60.74	61.36	-	73.05
[Sapena et al., 2011]	55.99	-	-	-	-	-
[Chang et al., 2011]	55.96	-	56.62	-	73.83	-
[Björkelund and Nugues, 2011]	54.53	-	56.91	-	-	-
[Nogueira dos Santos and Lopes Carvalho, 2011]	53.41	-	55.50	-	-	-
[Song et al., 2011]	53.05	-	49.77	-	-	-
[Stoyanov et al., 2011]	51.92	-	53.55	-	-	-
[Lalitha Devi et al., 2011]	51.90	-	-	-	-	-
[Kobdani and Schütze, 2011]	51.04	-	53.92	-	-	-
[Zhou et al., 2011]	50.92	-	-	-	-	-
[Charton and Gagnon, 2011]	50.36	-	-	-	-	-
[Yang et al., 2011]	49.99	-	-	-	-	-
[Xiong et al., 2011]	49.38	-	-	-	-	-
[Li et al., 2011]	48.46	-	-	-	-	-
[Chen et al., 2011]	48.07	-	50.25	-	-	-
[Kummerfeld et al., 2011]	47.10	-	-	-	-	-
[Zhekova and Kübler, 2011]	40.43	-	44.27	-	-	-
[Irwin et al., 2011]	31.88	35.84	-	-	-	-
[Cai et al., 2011]	-	55.71	-	-	-	-
[Uryupina et al., 2011]	-	54.32	-	-	-	-
[Klenner and Tuggener, 2011]	-	51.77	-	-	-	-

Table B.7: The official results from the CoNLL 2011 shared task [Pradhan et al., 2011] for all targeted settings and tracks: CT(closed track), OT(open track), GM(*gold* mentions), GB(*gold* boundaries), PM(predicted mentions)

B.3 CONLL 2012

B.3.1 *Predicted Mentions (Official)*

Participant	Open			Closed			Official	Final model	
	English	Chinese	Arabic	English	Chinese	Arabic	Score	Train	Dev
[Fernandes et al., 2012]				63.37	58.49	54.22	58.69	√	√
[Björkelund and Farkas, 2012]				61.24	59.97	53.55	58.25	√	√
[Chen and Ng, 2012]		63.53		59.69	62.24	47.13	56.35	√	×
[Stamborg et al., 2012]				59.36	56.85	49.43	55.21	√	√
[Uryupina et al., 2012]				56.12	53.87	50.41	53.47	√	√
[Zhekova et al., 2012]				48.70	44.53	40.57	44.60	√	√
[Li, 2012]				45.85	46.27	33.53	41.88	√	√
[Yuan et al., 2012]		61.02		58.68	60.69		39.79	√	√
[Xu et al., 2012]				57.49	59.22		38.90	√	×
[Martschat et al., 2012]				61.31	53.15		38.15	√	×
[Zhang et al., 2012]				59.24	51.83		37.02	-	-
yang[1]				55.29			18.43	√	×
[Chang et al., 2012]				60.18	45.71		35.30	√	×
[Li et al., 2012]				48.77	51.76		33.51	√	√
[Shou and Zhao, 2012]				58.25			19.42	√	×
[Xiong and Liu, 2012]	59.23	44.35	44.37				0.00	√	√

Table B.8: The official results from the CoNLL 2012 shared task [Pradhan et al., 2012] for the system runs using predicted mentions.

[1] This participant did not submit a final task paper.

B.3.2 *Gold Mention Boundaries (Supplementary)*

Participant	Open			Closed			Suppl.	Final model	
	English	Chinese	Arabic	English	Chinese	Arabic	Score	Train	Dev
[Fernandes et al., 2012]				63.16	61.48	53.90	59.51	√	√
[Björkelund and Farkas, 2012]				60.75	62.76	53.50	59.00	√	√
[Chen and Ng, 2012]		70.00		60.33	68.55	47.27	58.72	√	×
[Stamborg et al., 2012]				57.35	54.30	49.59	53.75	√	√
[Zhekova et al., 2012]				49.30	44.93	40.24	44.82	√	√
[Li, 2012]				43.04	43.28	31.46	39.26	√	√
[Yuan et al., 2012]				59.50	64.42		41.31	√	√
[Xu et al., 2012]				56.47	64.08		40.18	√	×
[Chang et al., 2012]				60.89			20.30	√	√

Table B.9: The official results from the CoNLL 2012 shared task [Pradhan et al., 2012] for the system runs using gold mentions.

B.3.3 *Gold Mentions (Supplementary)*

Participant	Open			Closed			Suppl.	Final model	
	English	Chinese	Arabic	English	Chinese	Arabic	Score	Train	Dev
[Fernandes et al., 2012]				69.35	66.36	63.49	66.40	√	√
[Björkelund and Farkas, 2012]				68.20	69.92	59.14	65.75	√	√
[Chen and Ng, 2012]		78.98		70.46	77.77	52.26	66.83	√	×
[Stamborg et al., 2012]				68.66	66.97	53.35	62.99	√	√
[Zhekova et al., 2012]				59.06	51.44	55.72	55.41	√	√
[Li, 2012]				51.40	59.93	40.62	50.65	√	√
[Yuan et al., 2012]				69.88	76.05		48.64	√	√
[Xu et al., 2012]				63.46	69.79		44.42	√	×
[Chang et al., 2012]				77.22			25.74	√	√

Table B.10: The official results from the CoNLL 2012 shared task [Pradhan et al., 2012] for the system runs using gold boundaries.

BIBLIOGRAPHY

Steven Abney. Parsing by Chunks. In Robert Berwick, Steven Abney, and Caroll Tenney, editors, *Principle-Based Parsing*, pages 257–278. Kluwer Academic Publishers, Dordrecht, 1991.

José Abraços and José Gabriel Lopes. Extending DRT with a Focusing Mechanism for Pronominal Anaphora and Ellipsis Resolution. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, (COLING '94), pages 1128–1132, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

David W. Aha, editor. *Lazy Learning*. Kluwer Academic Publishers, Norwell, MA, USA, 1997.

David W. Aha. Feature Weighting for Lazy Learning Algorithms. In Huan Liu and Hiroshi Motoda, editors, *Feature Extraction, Construction and Selection: a Data Mining Perspective*, volume SECS 453, pages 13–32. Kluwer Academic, Boston, 1998.

David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-Based Learning Algorithms. *Machine Learning*, 6(1):37–66, January 1991.

Sophia Ananiadou, John McNaught, and Paul Thompson. *The English Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30683-9. Available online at http://www.meta-net.eu/whitepapers.

Chinatsu Aone and Scott William Bennett. Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 122–129, Cambridge, Massachusetts, USA, June 1995. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/P95-1017.pdf.

Chinatsu Aone and Douglas McKee. Language-Independent Anaphora Resolution System for Understanding Multilingual Texts. In *Proceedings of 31st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1993.

Giuseppe Attardi, Maria Simi, and Stefano Dei Rossi. TANL-1: Coreference Resolution by Parse Analysis and Similarity Clustering. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 108–111, Uppsala, Sweden, July 2010. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/S10-1022.pdf.

Olga Babko-Malaya, Ann Bies, Ann Taylor, Szuting Yi, Martha Palmer, Mitch Marcus, Seth Kulick, and Libin Shen. Issues in Synchronizing the English Treebank and PropBank. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 70–77, Sydney, Australia, July 2006. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W/W06/W06-0609.pdf.

Amit Bagga and Breck Baldwin. Algorithms for Scoring Coreference Chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566, 1998.

A. Barss. *Anaphora: A Reference Guide*. Explaining Linguistics. John Wiley & Sons, 2008.

Andrew Barss. *Anaphora: a Reference Guide*. Explaining Linguistics; 3. Blackwell, Malden, MA [u.a.], 1. edition, 2003.

Eric Bengtson and Dan Roth. Understanding the Value of Features for Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 294–303, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Anders Björkelund and Richárd Farkas. Data-driven Multilingual Coreference Resolution using Resolver Stacking. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 49–55, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4503.pdf.

Anders Björkelund and Pierre Nugues. Exploring Lexicalized Features for Coreference Resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 45–50, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1905.pdf.

Diana Blagoeva, Svetla Koeva, and Vladko Murdarov. Българският език в дигиталната епоха – *The Bulgarian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30167-4. Available online at http://www.meta-net.eu/whitepapers.

Ondřej Bojar, Silvie Cinková, Jan Hajič, Barbora Hladká, Vladislav Kuboň, Jiří Mírovský, Jarmila Panevová, Nino Peterek, Johanka Spoustová, and Zdeněk Žabokrtský. *Čeština v digitálním věku – The Czech Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30705-8. Available online at http://www.meta-net.eu/whitepapers.

Lars Borin, Martha D. Brandt, Jens Edlund, Jonas Lindh, and Mikael Parkvall. *Svenska språket i den digitala tidsåldern – The Swedish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30831-4. Available online at http://www.meta-net.eu/whitepapers.

António Branco. *Anaphora: Analysis, Algorithms and Applications: 6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007, Lagos, Portugal, March 29 - 30, 2007; selceted papers*. Lecture Notes in Computer Science; 4410, Lecture notes in Artificial Intelligence. Springer, Berlin [u.a.], 2007.

António Branco, Amália Mendes, Sílvia Pereira, Paulo Henriques, Thomas Pellegrini, Hugo Meinedo, Isabel Trancoso, Paulo Quaresma, Vera Lúcia Strube de Lima, and Fernanda Bacelar. *A língua portuguesa na era digital – The Portuguese Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-29592-8. Available online at http://www.meta-net.eu/whitepapers.

Samuel Broscheit, Massimo Poesio, Simone Paolo Ponzetto, Kepa Joseba Rodriguez, Lorenza Romano, Olga Uryupina, Yannick Versley, and Roberto Zanoli. BART: A Multilingual Anaphora Resolution System. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 104–107, Uppsala, Sweden, July 2010. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/S10-1021.pdf.

Aljoscha Burchardt, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit,

and Martin Volk. *Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-27165-6. Available online at http://www.meta-net.eu/whitepapers.

Jie Cai and Michael Strube. Evaluation Metrics For End-to-End Coreference Resolution Systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 28–36, Tokyo, Japan, September 2010. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W/W10/W10-4305.pdf.

Jie Cai, Eva Mujdricza-Maydt, and Michael Strube. Unrestricted Coreference Resolution via Global Hypergraph Partitioning. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 56–60, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1907.pdf.

Nicoletta Calzolari, Bernardo Magnini, Claudia Soria, and Manuela Speranza. *La Lingua Italiana nell'Era Digitale – The Italian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30775-1. Available online at http://www.meta-net.eu/whitepapers.

Kai-Wei Chang, Rajhans Samdani, Alla Rozovskaya, Nick Rizzolo, Mark Sammons, and Dan Roth. Inference Protocols for Coreference Resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 40–44, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1904.pdf.

Kai-Wei Chang, Rajhans Samdani, Alla Rozovskaya, Mark Sammons, and Dan Roth. Illinois-Coref: The UI System in the CoNLL-2012 Shared Task. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 113–117, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4513.pdf.

Eric Charton and Michel Gagnon. Poly-co: a Multilayer Perceptron Approach for Coreference Detection. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 97–101, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1915.pdf.

Chen Chen and Vincent Ng. Combining the Best of Two Worlds: A Hybrid Approach to Multilingual Coreference Resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 56–63, Jeju Island, Korea, July 2012.

Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4504.pdf.

Weipeng Chen, Muyu Zhang, and Bing Qin. Coreference Resolution System using Maximum Entropy Classifier. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 127–130, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1921.pdf.

Jinho D. Choi and Martha Palmer. Robust Constituent-to-Dependency Conversion for English. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 55–66, 2010.

Noam Chomsky. Remarks on Nominalization. In R. Jacobs and P. Rosenbaum, editors, *Reading in English Transformational Grammar*, pages 184–221. Ginn and Co., Waltham, 1970.

Michael John Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, Philadelphia, PA, USA, 1999.

Scott Cost and Steven Salzberg. A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. *Machine Learning*, 10(1):57–78, 1993.

W. Daelemans and A. Van Den Bosch. *Memory-Based Language Processing*. Studies in Natural Language Processing. Cambridge University Press, 2005.

Walter Daelemans, Sabine Buchholz, and Jorn Veenstra. Memory-Based Shallow Parsing. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL 1999)*, pages 53–60, 1999.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. TiMBL: Tilburg Memory Based Learner – version 6.1 – Reference Guide. Technical Report ILK 07-07, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University, 2007.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. TiMBL: Tilburg Memory Based Learner, version 6.3,Reference Guide. Technical Report ILK 10-01, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University, 2010.

Dipanjan Das and Slav Petrov. Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/P11-1061.pdf.

Hal Daumé, III and Daniel Marcu. 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*, HLT '05, pages 97–104, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

Koenraad De Smedt, Gunn Inger Lyse, Anje Müller Gjesdal, and Gyri S. Losnegaard. *Norsk i den digitale tidsalderen (bokmålsversjon) – The Norwegian Language in the Digital Age (Bokmål Version)*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012a. ISBN 978-3-642-31388-2. Available online at http://www.meta-net.eu/whitepapers.

Koenraad De Smedt, Gunn Inger Lyse, Anje Müller Gjesdal, and Gyri S. Losnegaard. *Norsk i den digitale tidsalderen (nynorskversjon) – The Norwegian Language in the Digital Age (Nynorsk Version)*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012b. ISBN 978-3-642-31432-2. Available online at http://www.meta-net.eu/whitepapers.

Kees van Deemter and Rodger Kibble. On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics*, 26(4):629–637, 2000.

Pascal Denis and Jason Baldridge. Specialized Models and Ranking for Coreference Resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 660–669, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/D08-1069.pdf.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. *Proceedings of LREC 2004*, pages 837–840, 2004.

Guido Dunker and Carla Umbach. Verfahren zur Anaphernresolution in KIT-FAST [Experiments in Anaphora Resolution in KIT-FAST]. Technical Report KIT-2, Technische Universität Berlin, 1993.

Jacob Eisenstein and Randall Davis. Gesture Improves Coreference Resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Companion Volume: Short Papers on XX*, pages 37–40, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

Pradheep Elango. Coreference Resolution: A Survey. Technical report, University of Wisconsin Madison, 2005.

Simon Eszter, Lendvai Piroska, Németh Géza, Olaszy Gábor, and Vicsi Klára. *A magyar nyelv a digitális korban – The Hungarian Language in the Digital Age*.

META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30378-4. Available online at http://www.meta-net.eu/whitepapers.

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4502.pdf.

Ingrid Fischer, Bernd Geistert, and Günther Görz. Chart-based Incremental Semantics Construction with Anaphora Resolution Using λ-DRT. *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium*, pages 235–244, 1995.

B.A. Fox. *Discourse Structure and Anaphora: Written and Conversational English*. Cambridge Studies in Linguistics. Cambridge University Press, 1993.

Kari Fraurud. Pronoun Resolution in Unrestricted Text. *Nordic Journal of Linguistics*, 11:47–68, 1988.

Yoav Freund and Robert E. Schapire. Large Margin Classification Using the Perceptron Algorithm. *Machine Learning*, 37(3):277–296, 1999.

Carmen García-Mateo and Montserrat Arza Rodríguez. *O idioma galego na era dixital – The Galician Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30798-0. Available online at http://www.meta-net.eu/whitepapers.

Maria Gavrilidou, Maria Koutsombogera, Anastasios Patrikakos, and Stelios Piperidis. Η Ελληνική Γλώσσα στην Ψηφιακή Εποχή – *The Greek Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-28935-4. Available online at http://www.meta-net.eu/whitepapers.

Jesús Giménez and Lluís Màrquez. SVMTool: A General POS Tagger Generator Based on Support Vector Machines. In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC 2004)*, Lisbon, Portugal, 2004.

Diana Grigorova. Zero Pronoun Resolution in Bulgarian. In *Proceedings of the 12th International Conference on Computer Systems and Technologies*, CompSysTech '11, pages 399–404, New York, NY, USA, 2011. ACM.

Ralph Grishman and Beth Sundheim. Design of the MUC-6 evaluation. In *MUC6 '95: Proceedings of the 6th Conference on Message Understanding*, pages 1–11, Morristown, NJ, USA, 1995. Association for Computational Linguistics.

Jeanette K. Gundel, Michael Hegarty, and Kaja Borthen. Cognitive Status, Information Structure, and Pronominal Reference to Clausally Introduced Entities. *Journal of Logic, Language and Information*, 12(3):281–299, 2003.

Aria Haghighi and Dan Klein. Unsupervised Coreference Resolution in a Nonparametric Bayesian Model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/P07-1107.

Aria Haghighi and Dan Klein. Simple Coreference Resolution with Rich Syntactic and Semantic Features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 1152–1161, Singapore, August 2009. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/D/D09/D09-1120.pdf.

Johan Hall and Joakim Nivre. A Dependency-Driven Parser for German Dependency and Constituency Representations. In *Proceedings of the Workshop on Parsing German*, pages 47–54, Columbus, Ohio, June 2008. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W/W08/W08-1007.pdf.

M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English (English Language)*. Longman Pub Group, 1976.

Sanda M. Harabagiu and Steven J. Maiorano. Multilingual Coreference Resolution. In *Proceedings of ANLP 2000*, Seattle, WA, 2000.

Sanda M. Harabagiu, Razvan C. Bunescu, and Steven J. Maiorano. Text and Knowledge Mining for Coreference Resolution. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2001. Available online at http://aclweb.org/anthology-new/N/N01/N01-1008.pdf.

Sven Hartrumpf. Coreference Resolution with Syntactico-Semantic Rules and Corpus Statistics. In *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*, pages 137–144, Toulouse, France, July 2001. Available online at http://www.aclweb.org/anthology/W01-0717.

Iris Hendrickx, G. Bouma, F. Coppens, Walter Daelemans, Véronique Hoste, G. Kloostermans, A.-M. Mineur, J. van der Vloet, and J.-L. Verschelde. *Coreference Resolution for Extracting Answers for Dutch*. Marrakech, 2008.

Inmaculada Hernáez, Eva Navas, Igor Odriozola, Kepa Sarasola, Arantza Diaz de Ilarraza, Igor Leturia, Araceli Diaz de Lezana, Beñat Oihartzabal, and Jasone Salaberria. *Euskara Aro Digitalean – The Basque Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit

(Series Editors). Springer, 2012. ISBN 978-3-642-30795-9. Available online at http://www.meta-net.eu/whitepapers.

Donald Hindle and Mats Rooth. Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19:103–120, March 1993.

Erhard W. Hinrichs, Sandra Kübler, and Karin Naumann. A Unified Representation for Morphological, Syntactic, Semantic, and Referential Annotations. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 13–20, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W/W05/W05-0303.

Lynette Hirschman and Nancy Chinchor. MUC-7 Coreference Task Definition. In *Proceedings of MUC-7*. Science Applications International Corporation, 1997.

J.R. Hobbs. *Coherence and Coreference*. Technical note. SRI International, 1978.

Véronique Hoste. *Optimization Issues in Machine Learning of Coreference Resolution*. PhD thesis, University of Antwerp, 2005.

Véronique Hoste and Guy De Pauw. KNACK-2002: a Richly Annotated Corpus of Dutch Written Text. In *Proceedings of the Third Conference on International Language Resources and Evaluation (LREC 2006)*, pages 1432–1437, 2006.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Companion Volume: Short Papers*, pages 57–60, New York City, USA, June 2006. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/N/N06/N06-2015.pdf.

Joseph Irwin, Mamoru Komachi, and Yuji Matsumoto. Narrative Schema as World Knowledge for Coreference Resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 86–92, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1913.pdf.

Ray Jackendoff. X-bar-Syntax: A Study of Phrase Structure. *Linguistic Inquiry Monograph 2*, 1977.

John Judge, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P. Scannell, and Elaine Uí Dhonnchadha. *An Ghaeilge sa Ré Dhigiteach – The Irish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans

Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30557-3. Available online at http://www.meta-net.eu/whitepapers.

Mijail Kabadjov. *Anaphora Resolution and Discourse-new Classification: A Comprehensive Evaluation*. VDM Verlag, Saarbrücken, Germany, Germany, 2010.

Carmen Klaussner and Desislava Zhekova. Lexico-Syntactic Patterns for Automatic Ontology Building. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 109–114, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee. Available online at http://aclweb.org/anthology/R11-2017.

Dan Klein and Chris Manning. Maxent Models, Conditional Estimation, and Optimization. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2003) Tutorial, 2003.

Manfred Klenner and Don Tuggener. An Incremental Model for Coreference Resolution with Restrictive Antecedent Accessibility. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 81–85, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1912.pdf.

Hamidreza Kobdani and Hinrich Schütze. SUCRE: A Modular System for Coreference Resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 92–95, Uppsala, Sweden, July 2010. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/S10-1018.pdf.

Hamidreza Kobdani and Hinrich Schütze. Supervised Coreference Resolution with SUCRE. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 71–75, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1910.pdf.

Janet Kolodner. *Case-Based Reasoning*. Morgan Kaufmann, San Mateo, CA, 1993.

Jan G. Kooij. *Ambiguity in Natural Language: an Investigation of Certain Problems in its Linguistic Description.* North-Holland Pub. Co., Amsterdam,, 1971.

Kimmo Koskenniemi, Krister Lindén, Lauri Carlson, Martti Vainio, Antti Arppe, Mietta Lennes, Hanna Westerlund, Mirka Hyvärinen, Imre Bartis, Pirkko Nuolijärvi, and Aino Piehl. *Suomen kieli digitaalisella aikakaudella – The Finnish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-27247-9. Available online at http://www.meta-net.eu/whitepapers.

Simon Krek. *Slovenski jezik v digitalni dobi – The Slovene Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30635-8. Available online at http://www.meta-net.eu/whitepapers.

Robert Krovetz and W. Bruce Croft. Lexical Ambiguity and Information Retrieval. *ACM Trans. Inf. Syst.*, 10:115–141, April 1992.

Sandra Kübler and Desislava Zhekova. Singletons and Coreference Resolution Evaluation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 261–267, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee. Available online at http://aclweb.org/anthology/R11-1036.pdf.

Jonathan K Kummerfeld, Mohit Bansal, David Burkett, and Dan Klein. Mention Detection: Heuristics for the OntoNotes annotations. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 102–106, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1916.pdf.

Sobha Lalitha Devi, Pattabhi Rao, Vijay Sundar Ram R, M. C S, and A. A. Hybrid Approach for Coreference Resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 93–96, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1914.pdf.

Nguyen Thi Van Lam. Structure of English Noun Phrases. *TIL (Tun Institute of Learning)*, 2004.

Shalom Lappin and Herbert J. Leass. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535–561, December 1994. ISSN 0891-2017.

H. Leass and U. Schwall. *An Anaphora Resolution Procedure for Machine Translation*. IWBS report. IBM Deutschland, Wissenschaftliches Zentrum, Institut für Wissensbasierte Systeme, 1991.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 28–34, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1902.pdf.

Baoli Li. Learning to Model Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 129–135, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4516.pdf.

Xinxin Li, Xuan Wang, and Shuhan Qi. Coreference Resolution with Loose Transitivity Constraints. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 107–111, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1917.pdf.

Xinxin Li, Xuan Wang, and Xingwei Liao. Simple Maximum Entropy Models for Multilingual Coreference Resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 83–87, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4508.pdf.

Krista Liin, Kadri Muischnek, Kaili Müürisep, and Kadri Vider. *Eesti keel digiajastul – The Estonian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30784-3. Available online at http://www.meta-net.eu/whitepapers.

Xavier Lluís, Stefan Bott, and Lluís Màrquez. A Second-Order Joint Eisner Model for Syntactic and Semantic Dependency Parsing. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 79–84, Boulder, Colorado, June 2009. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W09-1212.pdf.

Xiaoqiang Luo. On Coreference Resolution Performance Metrics. In *HLT '05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

Xiaoqiang Luo and Imed Zitouni. Multi-Lingual Coreference Resolution with Syntactic Features. In *Proceedings of HLT/EMNLP 2005*, Vancouver, Canada, 2005.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. A Mention-Synchronous Coreference Resolution Algorithm Based On the Bell Tree. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 135–142, Barcelona, Spain, July 2004. Available online at http://acl.ldc.upenn.edu/P/P04/P04-1018.pdf.

Xiaoqiang Luo, Radu Florian, and Todd Ward. Improving Coreference Resolution by Using Conversational Metadata. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 201–204, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

Mohamed Maamouri and Ann Bies. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In Ali Farghaly and Karine Megerdoomian, editors, *International Conference on Computational Linguistics (COLING 2004), Workshop on Computational Approaches to Arabic Script-based Languages*, pages 2–9, Geneva, Switzerland, August 28th 2004. COLING. Available online at http://aclweb.org/anthology-new/W/W04/W04-1602.pdf.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Joseph Mariani, Patrick Paroubek, Gil Francopoulo, Aurélien Max, François Yvon, and Pierre Zweigenbaum. *La langue française à l' Ère du numérique – The French Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30760-7. Available online at http://www.meta-net.eu/whitepapers.

Katja Markert and Malvina Nissim. Comparing Knowledge Sources for Nominal Anaphora Resolution. *Computational Linguistics*, 31(3):367–402, 2005.

Sebastian Martschat, Jie Cai, Samuel Broscheit, Éva Mújdricza-Maydt, and Michael Strube. A multigraph model for coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 100–106, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4511.pdf.

James Mayfield, David Alexander, Bonnie Dorr, Jason Eisner, Tamer Elsayed, Tim Finin, Clay Fink, Marjorie Freedman, Nikesh Garera, Paul McNamee, Saif Mohammad, Douglas Oard, Christine Piatko, Asad Sayeed, Zareen Syed, Ralph Weischedel, Tan Xu, and David Yarowsky. Cross-Document Coreference Resolution: A Key Technology for Learning by Reading. In *Proceedings of the AAAI 2009 Spring Symposium on Learning by Reading and Learning to Read*, March 2009.

Laia Mayol. On Pronouns in Catalan and Game Theory. pages 7–11, August 2006.

Joseph McCarthy and Wendy Lehnert. Using Decision Trees for Coreference Resolution. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 1050–1055, Montreal, Canada, 1995.

Maite Melero, Toni Badia, and Asunción Moreno. *La lengua española en la era digital – The Spanish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30840-6. Available online at http://www.meta-net.eu/whitepapers.

Marcin Miłkowski. *Język polski w erze cyfrowej – The Polish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30810-9. Available online at http://www.meta-net.eu/whitepapers.

Ruslan Mitkov. 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 (ACL/-COLING)*, Montreal, Canada, 1998.

Ruslan Mitkov. Anaphora Resolution: The State Of The Art. Technical report, 1999a.

Ruslan Mitkov. Multilingual Anaphora Resolution. *Machine Translation*, 14(3): 281–299, 1999b.

Ruslan Mitkov. *Anaphora resolution*. Studies in Language and Linguistics. Longman, 2002.

Ruslan Mitkov, Shalom Lappin, and Branimir Boguraev. Introduction to the Special Issue on Computational Anaphora Resolution. *Computational Linguistics*, 27(4):473–477, 2001.

Asunción Moreno, Núria Bel, Eva Revilla, Emília Garcia, and Sisco Vallverdú. *La llengua catalana a l'era digital – The Catalan Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30677-8. Available online at http://www.meta-net.eu/whitepapers.

Tatsunori Mori, Mamoru Matsuo, and Hiroshi Nakagawa. Constraints and Defaults on Zero Pronouns in Japanese Instruction Manuals. *Proceedings of the Operational Factors in Practical, Robust Anaphora Resolution, Madrid, Spain*, pages 7–13, 1997. Available online at http://www.aclweb.org/anthology-new/W/W97/W97-1302.pdf.

Jane Morris and Graeme Hirst. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17 (1):21–48, 1991. Available online at http://www.aclweb.org/anthology-new/J/J91/J91-1002.pdf.

Hiromi Nakaiwa and Satoru Ikehara. Zero Pronoun Resolution in a Machine Translation System by Using Japanese to English Verbal Semantic

Attributes. *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 201–208, 1992.

Hiromi Nakaiwa and Satoru Ikehara. Intrasentential Resolution of Japanese Zero Pronouns in a Machine Translation System Using Semantic and Pragmatic Constraints. pages 96–105, 1995.

Karin Naumann. Manual for the Annotation of in-document Referential Relations, 2006.

Vincent Ng. Shallow Semantics for Coreference Resolution. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1689–1694, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

Vincent Ng. Unsupervised Models for Coreference Resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 640–649, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/D08-1067.

Vincent Ng and Claire Cardie. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 104–111, Philadelphia, Pennsylvania, USA, July 2002a. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/P02-1014.pdf.

Vincent Ng and Claire Cardie. Combining Sample Selection and Error-Driven Pruning for Machine Learning of Coreference Rules. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 55–62. Association for Computational Linguistics, July 2002b. http://www.aclweb.org/anthology/W02-1008.

Cicero Nogueira dos Santos and Davi Lopes Carvalho. Rule and Tree Ensembles for Unrestricted Coreference Resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 51–55, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1906.pdf.

Jan Odijk. *Het Nederlands in het Digitale Tijdperk – The Dutch Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-25977-7. Available online at http://www.meta-net.eu/whitepapers.

Tomoko Ohta, Yuka Tateisi, and Jin D. Kim. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 82–86, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

Constantin Orăsan, Dan Cristea, Ruslan Mitkov, and António Branco. Anaphora Resolution Exercise: an Overview. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

Lluís Padró and Evgeny Stanilovsky. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the 9th Conference on International Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 2005.

Martha Palmer, Olga Babko-Malaya, Ann Bies, Mona T. Diab, Mohamed Maamouri, Aous Mansouri, and Wajdi Zaghouani. A Pilot Arabic Propbank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 28-30 2008. European Language Resources Association.

Manuel Palomar and Patricio Martínez-Barco. Computational Approach to Anaphora Resolution in Spanish Dialogues. *J. Artif. Intell. Res. (JAIR)*, pages 263–287, 2001.

Bolette Sandford Pedersen, Jürgen Wedekind, Steen Bøhm-Andersen, Peter Juel Henrichsen, Sanne Hoffensetz-Andresen, Sabine Kirchmeier-Andersen, Jens Otto Kjærum, Louise Bie Larsen, Bente Maegaard, Sanni Nimb, Jens-Erik Rasmussen, Peter Revsbech, and Hanne Erdman Thomsen. *Det danske sprog i den digitale tidsalder – The Danish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30626-6. Available online at http://www.meta-net.eu/whitepapers.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A Universal Part-of-Speech Tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

M. Poesio, F. Bruneseaux, and L. Romary. The MATE Meta-Scheme for Coreference in Dialogues in Multiple Languages. In *Proceedings of Towards Standards and Tools for Discourse Tagging*. Association for Computational Linguistics, 21 June 1999. Available online at http://www.aclweb.org/anthology-new/W/W99/W99-0309.pdf.

Massimo Poesio, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira. Acquiring Lexical Knowledge For Anaphora Resolution. In *Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC)*, pages 1220–1224, 2002.

Massimo Poesio, Olga Uryupina, and Yannick Versley. Creating a Coreference Resolution System for Italian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

Simone Paolo Ponzetto and Michael Strube. Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2006. Available online at http://acl.ldc.upenn.edu/N/N06/N06-1025.pdf.

Andrei Popescu-Belis and Isabelle Robba. Cooperation between Pronoun and Reference Resolution for Unrestricted Texts. *Proceedings of the ACL'97/EACL'97*, pages 94–99, 1997. Available online at http://www.aclweb.org/anthology-new/W/W97/W97-1314.pdf.

Marta Recasens Potau. Towards Coreference Resolution for Catalan and Spanish, 2008.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 1–27, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1901.pdf.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4501.pdf.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: A Unified Relational Semantic Representation. In *Proceedings of the International Conference on Semantic Computing*, ICSC '07, pages 517–526, Washington, DC, USA, 2007. IEEE Computer Society.

Long Qiu, Min yen Kan, and Tat seng Chua. A Public Reference Implementation of the RAP Anaphora Resolution Algorithm. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004*, pages 291–294, 2004.

Altaf Rahman and Vincent Ng. Supervised Models for Coreference Resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 968–977, Singapore, August 2009. Association for Computational Linguistics. http://www.aclweb.org/anthology/D/D09/D09-1101.pdf.

Altaf Rahman and Vincent Ng. Narrowing the Modeling Gap: a Cluster-Ranking Approach to Coreference Resolution. *J. Artif. Int. Res.*, 40(1):469–521, January 2011.

Lance A. Ramshaw and Mitchell P. Marcus. Text Chunking Using Transformation-Based Learning. In *Proceedings of the Association for Computational Linguistics 3rd Workshop on Very Large Corpora*, pages 82–94, Cambridge, MA, 1995.

William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Marta Recasens and Eduard Hovy. A Deeper Look into Features for Coreference Resolution. In *Proceedings of Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*, 2009.

Marta Recasens and Eduard Hovy. BLANC: Implementing the Rand Index for Coreference Evaluation. *Natural Language Engineering*, 2011.

Marta Recasens and M. Antònia Martí. AnCora-CO: Coreferentially Annotated Corpora for Spanish and Catalan. *Language Resources Evaluation*, 44(4):315–345, 2010.

Marta Recasens, Toni Martí, Mariona Taulé, Lluís Màrquez, and Emili Sapena. SemEval-2010 task 1: Coreference Resolution in Multiple Languages. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pages 70–75, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W09-2411.pdf.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden, July 2010. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/S10-1001.pdf.

Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series.* Springer, 2012. Available online at http://www.meta-net.eu/whitepapers.

Eric Reuland. *Anaphora and Language Design*. Linguistic Inquiry Monographs; 62 [i.e.63]. MIT Press, Cambridge, Mass. [u.a.], 2011.

Nick Rizzolo and Dan Roth. Learning Based Java for Rapid Development of NLP Systems. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Language Resources and Evaluation (LREC 2010)*. European Language Resources Association, 2010.

Kepa Joseba Rodríguez, Francesca Delogu, Yannick Versley, Egon W. Stemle, and Massimo Poesio. Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 2010. European Language Resources Association (ELRA).

Eiríkur Rögnvaldsson, Kristín M. Jóhannsdóttir, Sigrún Helgadóttir, and Steinþór Steingrímsson. *Íslensk tunga á stafrænni öld – The Icelandic Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30173-5. Available online at http://www.meta-net.eu/whitepapers.

Monique Rolbert. *Résolution de formes pronominales dans l'interface d'intérrogation d'une base de données [Resolution of Pronominal Forms in a Database Interrogation Interface]*. PhD thesis, Faculté des Sciences de Luminy, France, 1989.

Mike Rosner and Jan Joachimsen. *Il-Lingwa Maltija Fl-Era Diġitali – The Maltese Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30680-8. Available online at http://www.meta-net.eu/whitepapers.

Dan Roth. Memory Based Learning in NLP. Technical Report UIUCDCS-R-99-2125, University of Illinois, March 1999.

Karin C. Ryding. *A Reference Grammar of Modern Standard Arabic*. Cambridge UP, 2005.

Emili Sapena, Lluís Padró, and Jordi Turmo. RelaxCor: A Global Relaxation Labeling Approach to Coreference Resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 88–91, Uppsala, Sweden, July 2010. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/S10-1017.pdf.

Emili Sapena, Lluís Padró, and Jordi Turmo. RelaxCor Participation in CoNLL Shared Task on Coreference Resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 35–39, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1903.pdf.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. Improving Coreference Resolution Using Bridging Reference Resolution and Automatically Acquired Synonyms. In *Discourse Anaphora and Anaphor Resolution Colloquium*, pages 125–136, 2007.

Helmut Schmid. Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the European Chapter of the Association for Computational Linguistics, Workshop of the Special Interest Group for Linguistic Data and Corpus-based Approaches to Natural Language Processing (EACL 1995 SIGDAT-Workshop)*, pages 47–50, 1995.

Helmut Schmid and Florian Laws. Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 777–784, Manchester, UK, August 2008. Coling 2008 Organizing Committee. Available online at http://www.aclweb.org/anthology/C08-1098.pdf.

Ur Shlonsky. The Form of Semitic Noun Phrases. *Lingua*, 2003.

Heming Shou and Hai Zhao. System paper for CoNLL-2012 shared task: Hybrid Rule-based Algorithm for Coreference Resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 118–121, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4514.pdf.

K. Simov, G. Popova, and P. Osenova. HPSG-based Syntactic Treebank of Bulgarian (BulTreeBank). Lincom-Europa, Munich, Germany, 2002.

Inguna Skadiņa, Andrejs Veisbergs, Andrejs Vasiljevs, Tatjana Gornostaja, Iveta Keiša, and Alda Rudzīte. *Latviešu valoda digitālajā laikmetā – The Latvian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30875-8. Available online at http://www.meta-net.eu/whitepapers.

Yang Song, Houfeng Wang, and Jing Jiang. Link Type Based Pre-Cluster Pair Model for Coreference Resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 131–135, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1922.pdf.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544, 2001.

Antonella Sorace and Francesca Filiaci. Anaphora Resolution in Near-Native Speakers of Italian. *Second Language Research*, 22(3):339–368, July 2006.

Marcus Stamborg, Dennis Medved, Peter Exner, and Pierre Nugues. Using Syntactic Dependencies to Solve Coreferences. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 64–70, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4505.pdf.

Craig Stanfill. Memory-Based Reasoning Applied to English Pronunciation. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 577–581, 1987.

Craig Stanfill and David L. Waltz. Towards Memory-Based Reasoning. *Communications of the ACM*, 29(12):1213–1228, 1986.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Jeek. Two Uses of Anaphora Resolution in Summarization. *Inf. Process. Manage.*, 43(6): 1663–1680, November 2007.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 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 Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664, Suntec, Singapore, August 2009. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/P/P09/P09-1074.pdf.

Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. Coreference Resolution with Reconcile. In *Proceedings of the Association for Computational Linguistics (ACL 2010) Conference Short Papers*, ACLShort '10, pages 156–161, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Veselin Stoyanov, Uday Babbar, Pracheer Gupta, and Claire Cardie. Reconciling OntoNotes: Unrestricted Coreference Resolution in OntoNotes with Reconcile. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 122–126, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1920.pdf.

Michael Strube, Stefan Rapp, and Christoph Müller. The Influence of Minimum Edit Distance on Reference Resolution. In *Proceedings of the ACL-02 Conference*

on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, pages 312–319, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

Roland Stuckardt. Resolving Anaphoric References on Deficient Syntactic Descriptions. In *Proceedings of the Association for Computational Linguistics and the European Chapter of the Association for Computational Linguistics (ACL-EACL'97) Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 30–37, Madrid, Spain, 1997. Available online at www.aclweb.org/anthology-new/W/W97/W97-1305.pdf.

Honglin Sun and Daniel Jurafsky. Shallow Semantc Parsing of Chinese. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 249–256, 2004. http://acl.ldc.upenn.edu/hlt-naacl2004/main/pdf/125_Paper.pdf.

Marko Tadić, Dunja Brozović-Rončević, and Amir Kapetanović. *Hrvatski Jezik u Digitalnom Dobu – The Croatian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30881-9. Available online at http://www.meta-net.eu/whitepapers.

Hristo Tanev and Ruslan Mitkov. Shallow Language Processing Architecture for Bulgarian. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

Erkan Tin and Varol Akman. Situated Processing of Pronominal Anaphora. *Proceedings of the KONVENS'94 Conference, Vienna, Austria*, pages 369–378, 1994.

Erik F. Tjong Kim Sang. Transforming a Chunker to a Parser. In Walter Daelemans, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel, editors, *Computational Linguistics in the Netherlands 2000*, pages 177–188. Rodopi, Amsterdam, 2001.

Diana Trandabăț, Elena Irimia, Verginica Barbu Mititelu, Dan Cristea, and Dan Tufiș. *Limba română în era digitală – The Romanian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30702-7. Available online at http://www.meta-net.eu/whitepapers.

Olga Uryupina. Linguistically Motivated Sample Selection for Coreference Resolution. In *Proceedings of Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2004)*, Sao Miguel, Portugal, 2004.

Olga Uryupina. Corry: A System for Coreference Resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 100–103, Uppsala,

Sweden, July 2010. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/S10-1020.pdf.

Olga Uryupina, Sriparna Saha, Asif Ekbal, and Massimo Poesio. Multi-Metric Optimization for Coreference: The UniTN / IITP / Essex Submission to the 2011 CONLL Shared Task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 61–65, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1908.pdf.

Olga Uryupina, Alessandro Moschitti, and Massimo Poesio. BART goes multilingual: The UniTN / Essex submission to the CoNLL-2012 Shared Task. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 122–128, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4515.pdf.

Daiva Vaišniene and Jolanta Zabarskaite. *Lietuvių kalba skaitmeniniame amžiuje – The Lithuanian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30757-7. Available online at http://www.meta-net.eu/whitepapers.

G. Van Noord, I. Schuurman, and V. Vandeghinste. Syntactic annotation of large corpora in STEVIN. In *Proceedings of the Fourth International Language Resources and Evaluation (LREC 2006)*, pages 1811–1814, Genoa, Italy, 2006.

Jorn Veenstra and Sabine Buchholz. Fast NP Chunking Using Memory-Based Learning Techniques. In *Proceedings of BENELEARN'98*, pages 71–78, Wageningen, the Netherlands, 1998.

Yannick Versley, Alessandro Moschitti, Massimo Poesio, and Xiaofeng Yang. Coreference Systems Based on Kernels Methods. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 961–968, Morristown, NJ, USA, 2008a. Association for Computational Linguistics.

Yannick Versley, Simone Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. BART: A Modular Toolkit for Coreference Resolution. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008b. European Language Resources Association (ELRA).

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Message Understanding Conference*, Columbia, MD, 1995. Available online at http://aclweb.org/anthology-new/M/M95/M95-1005.pdf.

Duško Vitas, Ljubomir Popović, Cvetana Krstev, Ivan Obradović, Gordana Pavlović-Lažetić, and Mladen Stanojević. Српски језик у дигиталном добу – *The Serbian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30754-6. Available online at http://www.meta-net.eu/whitepapers.

Mária Šimková, Radovan Garabík, Katarína Gajdošová, Michal Laclavík, Slavomír Ondrejovič, Jozef Juhár, Ján Genči, Karol Furdík, Helena Ivoríková, and Jozef Ivanecký. *Slovenský jazyk v digitálnom veku – The Slovak Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30369-2. Available online at http://www.meta-net.eu/whitepapers.

Bonnie L. Webber. *A Formal Approach to Discourse Anaphora*. PhD thesis, Harvard University, 1978.

Holger Wunsch. *Rule-based and Memory-based Pronoun Resolution for German: A Comparison and Assessment of Data Sources*. PhD thesis, Universität Tübingen, 2010.

Holger Wunsch, Sandra Kübler, and Rachael Cantrell. Instance Sampling Methods for Pronoun Resolution. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgaria, 2009.

Chenhai Xi and Rebecca Hwa. A Backoff Model for Bootstrapping Resources for Non-English Languages. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 851–858, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/H/H05/H05-1107.pdf.

Hao Xiong and Qun Liu. ICT: System Description for CoNLL-2012. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 71–75, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4506.pdf.

Hao Xiong, Linfeng Song, Fandong Meng, Yang Liu, Qun Liu, and Yajuan Lv. ETS: An Error Tolerable System for Coreference Resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 76–80, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1911.pdf.

Ruifeng Xu, Jun Xu, Jie Liu, Chengxiang Liu, Chengtian Zou, Lin Gui, Yanzhen Zheng, and Peng Qu. Incorporating Rule-based and Statistic-based Techniques for Coreference Resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 107–112, Jeju Island, Korea, July 2012. Association for

Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4512.pdf.

Naiwen Xue, Fei Xia, Fu-dong Chiou, and Marta Palmer. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207–238, 2005.

Nianwen Xue and Martha Palmer. Adding Semantic Roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143–172, 2009.

Xiaofeng Yang and Jian Su. Coreference Resolution Using Semantic Relatedness Information from Automatically Discovered Patterns. In *Association for Computational Linguistics*, 2007. Available online at http://aclweb.org/anthology-new/P/P07/P07-1067.pdf.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. Coreference Resolution Using Competition Learning Approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2003. Available online at http://acl.ldc.upenn.edu/P/P03/P03-1023.pdf.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. Kernel-based Pronoun Resolution with Structured Syntactic Knowledge. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 41–48, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Yaqin Yang, Nianwen Xue, and Peter Anick. A Machine Learning-Based Coreference Detection System for OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 117–121, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1919.pdf.

David Yarowsky and Grace Ngai. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

Bo Yuan, Qingcai Chen, Yang Xiang, Xiaolong Wang, Liping Ge, Zengjian Liu, Meng Liao, and Xianbo Si. A Mixed Deterministic Model for Coreference Resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 76–82, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4507.pdf.

Wajdi Zaghouani, Mona Diab, Aous Mansouri, Sameer Pradhan, and Martha Palmer. The Revised Arabic PropBank. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 222–226, Uppsala, Sweden, July 2010. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W10-1836.pdf.

Xiaotian Zhang, Chunyang Wu, and Hai Zhao. Chinese Coreference Resolution via Ordered Filtering. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 95–99, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4510.pdf.

Yue Zhang and Stephen Clark. Syntactic Processing Using the Generalized Perceptron and Beam Search. *Computational Linguistics*, 37(1):105–151, 2011.

Shanheng Zhao and Hwee Tou Ng. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach. In *Proceedings of EMNLP-CoNLL 2007*, Prague, Czech Republic, 2007.

Desislava Zhekova. Instance Sampling for Multilingual Coreference Resolution. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 150–155, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee. Available online at http://aclweb.org/anthology/R11-2024.pdf.

Desislava Zhekova and Sandra Kübler. UBIU: A Language-Independent System for Coreference Resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99, Uppsala, Sweden, July 2010. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/S10-1019.pdf.

Desislava Zhekova and Sandra Kübler. UBIU: A Robust System for Resolving Unrestricted Coreference. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 112–116, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1918.pdf.

Desislava Zhekova, Sandra Kübler, Joshua Bonner, Marwa Ragheb, and Yu-Yin Hsu. UBIU for Multilingual Coreference Resolution in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 88–94, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W12-4509.pdf.

Huiwei Zhou, Yao Li, Degen Huang, Yan Zhang, Chunlong Wu, and Yuansheng Yang. Combining Syntactic and Semantic Features by SVM for Unrestricted Coreference Resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 66–70,

Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Available online at http://www.aclweb.org/anthology/W11-1909.pdf.

INDEX

COLOPHON

This thesis was typeset with LaTeX using André Miede's `classicthesis` package available for LaTeX via CTAN under `http://www.ctan.org/tex-archive/macros/latex/contrib/classicthesis`.

ERKLÄRUNG

Ich versichere hiermit, dass ich meine Dissertation

"Towards Multilingual Coreference Resolution "

selbstständig und ohne fremde Hilfe angefertigt und sie nicht vorher in einem anderen Prüfungsverfahren eingereicht habe und dass ich alle von anderen Autoren wörtlich übernommenen Stellen wie auch die sich an die Gedankengänge anderer Autoren eng anlehnenden Ausführungen meiner Arbeit besonders gekennzeichnet und die Quellen zitiert habe.

Bremen, den 24. Mai 2013

_____

Desislava Zhekova