

Computational modelling of coreference and bridging resolution

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität
Stuttgart zur Erlangung der Würde eines Doktors der Philosophie (Dr. phil.)
genehmigte Abhandlung.

Vorgelegt von
Ina Verena Rösiger
aus Göppingen

Hauptberichter Prof. Dr. Jonas Kuhn
Mitberichter Prof. Dr. Simone Teufel

Tag der mündlichen Prüfung: 28.01.2019

Institut für Maschinelle Sprachverarbeitung
der Universität Stuttgart

2019

Erklärung (Statement of Authorship)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet.

I hereby declare that this text is the result of my own work and that I have not used sources without declaration in the text. Any thoughts from others or literal quotations are clearly marked.

Ort, Datum

Unterschrift

Contents

| | |
|---|-----------|
| 1. Introduction | 5 |
| 1.1. Motivation | 5 |
| 1.2. Research questions | 7 |
| 1.3. Contributions and publications | 9 |
| 1.4. Outline of the thesis | 14 |
| | |
| I. Background | 17 |
| 2. Anaphoric reference | 19 |
| 2.1. Coreference | 22 |
| 2.2. Bridging | 30 |
| 3. Related NLP tasks | 35 |
| 3.1. Coreference resolution | 35 |
| 3.2. Bridging resolution | 41 |
| | |
| II. Data and tool creation | 47 |
| | |
| 4. Annotation and data creation | 49 |
| 4.1. Coreference annotation and existing corpora | 49 |
| 4.2. Bridging annotation and existing corpora | 53 |
| 4.3. Newly created corpus resources | 61 |
| 4.3.1. BASHI: bridging in news text | 63 |
| 4.3.2. SciCorp: coreference and bridging in scientific articles | 70 |
| 4.3.3. GRAIN: coreference and bridging in radio interviews | 79 |
| 4.3.4. Conclusion | 81 |
| 5. Coreference resolution | 83 |
| 5.1. Existing tools and related work | 84 |
| 5.2. A coreference system for German | 88 |
| 5.2.1. System and data | 88 |
| 5.2.2. Adapting the system to German | 91 |
| 5.2.3. Evaluation | 99 |
| 5.2.4. Ablation experiments | 101 |
| 5.2.5. Pre-processing pipeline: running the system on new texts | 102 |

| | | |
|-------------|---|------------|
| 5.2.6. | Application on DIRNDL | 104 |
| 5.3. | Conclusion | 106 |
| 6. | Bridging resolution | 109 |
| 6.1. | A rule-based bridging system for English | 109 |
| 6.1.1. | Reimplementation | 111 |
| 6.1.2. | Performance | 120 |
| 6.1.3. | Generalisability of the approach | 123 |
| 6.2. | CRAC 2018: first shared task on bridging resolution | 124 |
| 6.2.1. | The ARRAU corpus | 125 |
| 6.2.2. | Data preparation | 125 |
| 6.2.3. | Evaluation scenarios and metrics | 127 |
| 6.2.4. | Applying the rule-based system to ARRAU | 128 |
| 6.3. | A refined bridging definition | 131 |
| 6.3.1. | Referential bridging | 132 |
| 6.3.2. | Lexical bridging | 134 |
| 6.3.3. | Subset relations and lexical givenness | 135 |
| 6.3.4. | Near-identity | 137 |
| 6.3.5. | Priming and bridging | 137 |
| 6.4. | Shared task results | 138 |
| 6.4.1. | Rules for bridging in ARRAU | 138 |
| 6.4.2. | A learning-based method | 143 |
| 6.4.3. | Final performance | 144 |
| 6.5. | A rule-based bridging system for German | 146 |
| 6.5.1. | Adaptation to German | 147 |
| 6.5.2. | Performance | 157 |
| 6.6. | Conclusion | 160 |
| III. | Linguistic validation experiments | 163 |
| 7. | Using prosodic information to improve coreference resolution | 165 |
| 7.1. | Motivation | 165 |
| 7.2. | Background | 167 |
| 7.3. | Related work | 169 |
| 7.4. | Experimental setup | 170 |
| 7.5. | Prosodic features | 171 |
| 7.6. | Manual prosodic information | 173 |
| 7.7. | Automatically predicted prosodic information | 174 |
| 7.8. | Results and discussion | 176 |
| 7.9. | Conclusion and future work | 181 |

| | |
|---|------------|
| 8. Integrating predictions from neural-network relation classifiers into coreference and bridging resolution | 183 |
| 8.1. Relation hypotheses | 184 |
| 8.2. Experimental setup | 186 |
| 8.3. First experiment | 187 |
| 8.3.1. Semantic relation classification | 187 |
| 8.3.2. Relation analysis | 189 |
| 8.3.3. Relations for bridging resolution | 190 |
| 8.3.4. Relations for coreference resolution | 192 |
| 8.4. Second experiment | 192 |
| 8.4.1. Semantic relation classification | 193 |
| 8.4.2. Relation analysis | 194 |
| 8.4.3. Relations for coreference and bridging resolution | 194 |
| 8.5. Final performance of the bridging tool | 195 |
| 8.6. Discussion and conclusion | 196 |
| 9. Conclusion | 199 |
| 9.1. Summary of contributions | 199 |
| 9.2. Lessons learned | 202 |
| 9.3. Future work | 207 |
| Bibliography | 209 |

List of figures

| | | |
|------|--|-----|
| 1.1. | Four levels of contribution | 8 |
| 1.2. | Contributions to coreference resolution | 9 |
| 1.3. | Contributions to bridging resolution | 10 |
| 2.1. | Reference as the relation between referring expressions and referents . . . | 19 |
| 3.1. | Latent trees for coreference resolution: data structures | 38 |
| 4.1. | Contribution and workflow pipeline for coreference: data creation | 61 |
| 4.2. | Contribution and workflow pipeline for bridging: data creation | 62 |
| 5.1. | Contribution and workflow pipeline for coreference: tool creation | 84 |
| 6.1. | Contribution and workflow pipeline for bridging: tool creation | 110 |
| 6.2. | Contribution and workflow pipeline for bridging: task definition (reloaded) | 132 |
| 7.1. | Contribution and workflow pipeline for coreference: validation, part 1 . . | 166 |
| 7.2. | One exemplary pitch accent shape | 167 |
| 7.3. | The relation between phrase boundaries and intonation phrases | 168 |
| 7.4. | The relation between boundary tones and nuclear and prenuclear accents | 168 |
| 7.5. | Convolutional neural network model for prosodic event recognition | 175 |
| 7.6. | The relation between coreference and prominence: example from the DIRNDL dataset with English translation | 181 |
| 8.1. | Contribution and workflow pipeline for coreference: validation, part 2 . . | 184 |
| 8.2. | Contribution and workflow pipeline for bridging: validation | 185 |
| 8.3. | Neural net relation classifier: example of a non-related pair | 188 |
| 8.4. | Neural net relation classifier: example of a hypernym pair | 188 |
| 8.5. | Neural net relation classifier in the second experiment | 194 |
| 9.1. | A data structure based on latent trees for the joint learning of coreference and bridging | 206 |

List of tables

| | |
|--|-----|
| 4.1. Guideline comparison: overview of the main differences between OntoNotes, RefLex and NoSta-D | 50 |
| 4.2. Existing corpora annotated with coreference used in this thesis | 51 |
| 4.3. Existing corpora annotated with bridging used in this work | 58 |
| 4.4. An overview of the newly created data | 63 |
| 4.5. BASHI: corpus statistics | 69 |
| 4.6. BASHI: inter-annotator agreement on five WSJ articles | 70 |
| 4.7. SciCorp: categories and links in our classification scheme | 73 |
| 4.8. SciCorp: overall inter-annotator-agreement (in κ) | 76 |
| 4.9. SciCorp: inter-annotator-agreement for the single categories (in κ) | 77 |
| 4.10. SciCorp: corpus statistics | 78 |
| 4.11. SciCorp: distribution of information status categories, in absolute numbers | 78 |
| 4.12. SciCorp: distribution of information status categories, in percent | 79 |
| 5.1. IMS HotCoref DE: performance of the mention extraction module on TüBa-D/Z version 8 | 91 |
| 5.2. IMS HotCoref DE: performance of the mention extraction module after the respective parse adjustments, on TüBa-D/Z version 8 | 94 |
| 5.3. IMS HotCoref DE: performance of the mention extraction module on TüBa-D/Z version 10 | 94 |
| 5.4. Performance of IMS HotCoref DE on TüBa-D/Z version 10: gold vs. predicted annotations | 99 |
| 5.5. SemEval-2010 official shared task results for German | 100 |
| 5.6. SemEval-2010 post-task evaluation | 101 |
| 5.7. SemEval-2010: post-task evaluation, excluding singletons | 101 |
| 5.8. Performance of IMS HotCoref DE on TüBa-D/Z version 10: ablation experiments | 102 |
| 5.9. CoNLL-12 format overview: tab-separated columns and content | 103 |
| 5.10. Markable extraction for the DIRNDL corpus | 105 |
| 5.11. Performance of IMS HotCoref DE on DIRNDL, using predicted annotations | 106 |
| 6.1. Overview of rules in Hou et al. (2014) | 113 |
| 6.2. Contingency table for the Noun1 + preposition + Noun2 pattern | 114 |
| 6.3. Exemplary semantic connectivity scores | 114 |
| 6.4. Exemplary argument-taking ratios | 115 |
| 6.5. A bridging system for English: performance of the individual rules, their precision as well as their firing rate | 120 |

List of tables

| | | |
|-------|---|-----|
| 6.6. | Performance of the reimplementa- tion of Hou et al. (2014), with different settings | 121 |
| 6.7. | Performance of the bridging system with different coreference information, gold setting | 122 |
| 6.8. | Performance of the bridging system with different coreference information, predicted setting | 122 |
| 6.9. | Performance of the rule-based method on other corpora. | 123 |
| 6.10. | Number of bridging anaphors in the single domains of the ARRAU corpus | 125 |
| 6.11. | Bridging relations in ARRAU | 126 |
| 6.12. | The CoNLL-12-style format used in our bridging experiments | 126 |
| 6.13. | Number of bridging anaphors in the shared task after filtering out prob- lematic cases | 127 |
| 6.14. | Applying Hou et al. (2014) on the RST part of the ARRAU corpus: rule performance | 129 |
| 6.15. | Performance of the rule-based method on other corpora | 139 |
| 6.16. | Shared task performance on the domains of ARRAU | 144 |
| 6.17. | Shared task results, more detailed evaluation | 145 |
| 6.18. | Performance of the single rules on the test set of the RST dataset | 146 |
| 6.19. | Overview of German corpora annotated with bridging | 147 |
| 6.20. | Bridging resolution on DIRNDL: precision of the firing rules | 157 |
| 6.21. | Bridging resolution on DIRNDL: overall performance | 157 |
| 6.22. | Bridging resolution on DIRNDL: predicted vs. gold mentions | 158 |
| 6.23. | Bridging resolution with different coreference information in DIRNDL | 158 |
| 7.1. | ToBI types in GToBI(S) | 172 |
| 7.2. | Performance of pitch accent presence (in CoNLL score) | 178 |
| 7.3. | Performance of nuclear accent presence (in CoNLL score) | 179 |
| 7.4. | Additional features based on manual prosodic information (gold setting) | 179 |
| 8.1. | Results of the intrinsic evaluation on BLESS (without lexical overlap) | 189 |
| 8.2. | Average cosine similarities and relation classifier probabilities for corefer- ent and bridging pairs in comparison to other pairs of nouns | 190 |
| 8.3. | Correct and wrong bridging pairs found by the additional semantic rule | 190 |
| 8.4. | Effect of the cosine threshold constraint, for the relation meronymy | 191 |
| 8.5. | Results of the intrinsic evaluation on WordNet | 194 |
| 8.6. | Average relation classifier probabilities and cosine similarities for corefer- ent and bridging pairs in comparison to other pairs of nouns, experiment 2 | 195 |
| 8.7. | Final performance of the English bridging system | 195 |
| 8.8. | Final performance of the English bridging system with different corefer- ence information | 196 |
| 9.1. | Comparison of different German coreference systems | 200 |
| 9.2. | Performance of the English bridging system | 200 |

| | |
|---|-----|
| 9.3. Performance of the German bridging resolver (on DIRNDL) | 201 |
| 9.4. Performance of pitch accent and nuclear accent presence (in CoNLL score) | 202 |
| 9.5. Final performance of the English bridging system | 202 |

List of abbreviations

| | |
|--------|--|
| BCUBE | Coreference evaluation metric proposed by Bagga and Baldwin (1998) |
| BLANC | Bilateral assessment of noun-phrase coreference |
| BNC | British National Corpus |
| CEAFE | Constraining Entity-Alignment F-Measure, entity-based |
| CEAFM | Constraining Entity-Alignment F-Measure, mention-based |
| CNN | Convolutional neural network |
| CNP | Conjugated noun phrase |
| CL | Computational linguistics |
| COS | Cosine similarity |
| DL | Constituency parse tag for names |
| F1 | F1 score |
| GEN | Genetics |
| GPE | Geopolitical entity |
| IP | Intonation phrases |
| ip | Intermediate phrases |
| LEA | Link-based entity-aware metric |
| MUC | Message understanding conference score |
| N | Noun |
| NE | Named entity |
| NLP | Natural language processing |
| NN | Neural network |
| NP | Noun phrase |
| n1 | Final accent of an intermediate phrase |
| n2 | Final accent of an intonation phrase |
| LSTM | Long short-term memory network |
| ORG | Organisation |
| P | Precision |
| PDS | Demonstrative pronoun |
| PER | Person |
| pn | Preuclear (non-final) accent |
| POS | Part-of-speech |
| PP | Prepositional phrase |
| PPER | Personal pronoun |
| PPOSAT | Possessive pronoun |
| PRELS | Relative pronoun |

| | |
|------|-------------------------|
| PREP | Preposition |
| PRF | Reflexive pronoun |
| PWS | Interrogative pronoun |
| R | Recall |
| ReLU | Rectified linear units |
| ToBI | Tones and Break Indices |
| VP | Verbal phrase |
| WN | WordNet |

Acknowledgements

This thesis wouldn't exist if it weren't for the many people who have supported me and my work over the last five years.

I would like to thank my advisor Jonas Kuhn for his encouragement and advice, for letting me explore my own ideas while always providing directions in case I needed them.

I am very grateful to Simone Teufel, not only for accepting to be the second reviewer of this thesis and for her many detailed comments that helped to make this thesis better, but also for introducing me to this topic many years ago during my stay in Cambridge and for sparking my interest in pursuing a PhD.

I would like to thank Arndt Riestler for all his advice, for the very helpful feedback on so many of my publications and posters, and for being a great role model of a scientist who always goes the extra mile. I have always enjoyed when we joined forces and turned our combined knowledge into fruitful collaborations.

Writing a PhD thesis can be lonely at times, but because of the many friendly and helpful faces at IMS, I've rarely ever felt alone. I want to thank my colleagues at IMS for this lovely time, particularly the people I've had the pleasure of working with: Sabrina, Markus, Kerstin, Janis, Nils, Sarah, Max, Maximilian, Sabine, Fabienne, Kim-Anh, Thang, Johannes, Tanja, Anja, Simon, Julia and Uli Heid. A special shout-out goes to my office mates Wiltrud and Yvonne and to my Mensa group!

One of the benefits of doing a PhD is that you get to travel to so many interesting places for conferences. Some of my highlights (besides all the academic input of course) include exploring the wonderful Kyushu, staying in a ryokan, relaxing in an onsen, visiting the Alhambra or the many temples in Kyoto, taking a steamboat cruise on the Mississippi River, climbing what felt like the steepest part of the Great Wall of China, and a little detour trip from Denver to Chicago, just to name a few. To everyone who has accompanied me on these trips (you know who you are), thanks for the great time and all the memories!

Last but not least, the biggest thank you goes to my family and Micha for their love and support throughout the years.

Abstract

Anaphora resolution is an important task in natural language understanding, where the aim is to automatically extract meaning from text. Anaphora resolution subsumes the two tasks coreference and bridging resolution.

Coreference resolution deals with coreference or identity anaphora, where a context-dependent expression refers to a previously mentioned entity. This includes pronouns such as *Tim ... he*, but also definite descriptions such as *Laura ... the girl*.

Bridging resolution revolves around bridging or associative anaphora, where the context-dependent expression itself has not been introduced into the discourse, but due to an already mentioned and associated entity, the expression can be interpreted, e.g. in *a school ... the headmaster*.

The goal of this thesis is to improve coreference and bridging resolution for English and German. Our contributions comprise the four levels task definition, data creation, tool creation and linguistic validation experiments. Based on the state of the art and previous work on both tasks, our focus for coreference resolution is set on later steps in the pipeline, while for bridging resolution work on all levels was required.

Whereas the task definition for coreference is well-defined and compatible in previous research, the bridging annotations we found in available corpora contained very different phenomena and motivated use to propose a refined bridging definition. We introduce the term referential bridging to cover two types of bridging on the level of referring expressions: (i) argument slot filling, as in *the wheel (of the car)* and (ii) referential subset expressions, as in *the small pug (out of the previously mentioned group of dogs)*. In both cases, context-dependence is the main criterion for referential bridging. This is not the case for lexical or lexically induced bridging, where we have a non-anaphoric or anaphoric expression that stands in some relation with a previously introduced entity. This relation typically exists either on the word level or models a real-world relation based on the relation on the concept level (*Europe ... Spain*).

In terms of data, we create three new corpus resources annotated with bridging and coreference information to overcome the lack of data particularly evident for bridging.

We have annotated BASHI, an English corpus of Wall Street Journal articles, SciCorp as an English corpus of scientific articles and the German corpus GRAIN, which comprises radio interviews.

While many English coreference resolvers are available, not many systems exist for German. We adapt a data-driven coreference resolver designed for English to German by integrating features designed to address the specificities of German. The tool achieves state-of-the-art performance on the benchmark dataset TüBa-D/Z. For bridging resolution, there are no openly available systems. Building on a rule-based approach, we develop bridging resolvers for English and German, which both achieve state-of-the-art performance. We show that the English bridging resolver generalises well to other in-domain corpora if they are of the same type of bridging, namely referential bridging.

Finally, inspired by theoretical studies, we improve the developed tools by integrating linguistic information that is assumed to be beneficial for the tasks. First, we show that the theoretic claims on the interaction between coreference and prosody hold true in an automatic setting: we improve the performance of our coreference resolver by integrating prosodic information, which is included in the form of manual prosodic labels or by using automatic labels predicted by a CNN classifier. In a second experiment, we test the use of semantic relations predicted by a neural-net relation classifier and show that automatically predicted meronymy pairs improve our bridging resolver.

Deutsche Zusammenfassung

Anaphernresolution befasst sich mit Methoden zur automatischen Auflösung von kontextabhängigen sprachlichen Ausdrücken. Es umfasst die zwei Aufgaben Koreferenzauflösung und Bridgingauflösung. Die Auflösung kontextabhängiger Ausdrücke ist ein wichtiger Teilschritt des automatischen Textverstehens.

Koreferenzauflösung bildet kontextabhängige koreferente Anaphern, die ohne Hinzunahme bisherigen Kontexts nicht interpretierbar sind, auf bereits eingeführte Entitäten ab. Das umfasst klassischerweise Pronomen wie z.B. *Tim ... er*, aber auch andere nominale Ausdrücke wie z.B. definite Deskriptionen in *Laura ... das Mädchen*.

Bridgingauflösung beschäftigt sich mit der Abbildung kontextabhängiger Ausdrücke auf bereits eingeführte Entitäten, die im Gegensatz zur Koreferenz nicht in einer identitären Relation stehen, sondern nur assoziiert sind (*die Schule ... der Rektor*).

Das Ziel dieser Arbeit ist es, die automatische Auflösung von Koreferenz und Bridging für Englisch und Deutsch zu verbessern. Die Forschungsbeiträge dieser Arbeit umfassen dabei die vier Ebenen Problemdefinition, Erstellung von manuell annotierten Daten, Entwicklung von Werkzeugen zur automatischen Analyse sowie linguistische Validierungsexperimente.

Da der Fortschritt im Bereich Koreferenz aufgrund des großen Forschungsaufkommens deutlicher weiter ist als im Bereich Bridging und es viele große, zuverlässig mit Koreferenz annotierte Korpora gibt, liegt der Schwerpunkt im Bereich Koreferenz auf den Schritten Werkzeugerstellung und darauf basierenden linguistischen Experimenten. Im Bereich Bridging sind unsere Forschungsbeiträge auf allen vier Ebenen zu finden.

Während bisherige, verwandte Arbeiten im Bereich Koreferenz und Koreferenzauflösung vergleichbare und klare Definitionen verwenden, enthalten die annotierten Korpora im Bereich Bridging sehr unterschiedliche Phänomene, was eine genauere Betrachtung und Charakterisierung der verschiedenen Bridgingdefinitionen motivierte. Unsere Charakterisierung unterscheidet referentielles Bridging, das zwei Untertypen umfasst: (i) Bridging als Einsatz von impliziten Argumenten, wie in *das Lenkrad (des Autos)*, und (ii) referentielle Teilmengenbeziehung wie z.B. in *der Mops (aus der bereits erwähnten*

Gruppe der Hunde). Das Hauptkriterium für referentielles Bridging ist dabei stets die Kontextabhängigkeit des sprachlichen Ausdrucks. Im Gegensatz dazu beschreibt lexikalisches Bridging eine Relation auf Wort- oder Konzeptebene, bei der der sprachliche Ausdruck nicht notwendigerweise kontextabhängig sein muss (*Europa ... Spanien*).

Im Bereich der Korporaerstellung motivierte vor allem der Mangel an annotierten Daten im Bereich Bridging die Annotation von drei verschiedenen Korpora: BASHI, ein englisches Korpus aus Wall-Street-Journal-Artikeln, SciCorp, ein englisches Korpus aus wissenschaftlichen Veröffentlichungen sowie GRAIN, ein deutsches Korpus aus Radiointerviews.

Während für das Englische viele verfügbare Koreferenzauflöser existieren, gibt es für Deutsch vergleichsweise wenig Werkzeuge zur automatischen Auflösung. Basierend auf einem englischen, lernbasierten Werkzeug entwickeln wir daher ein frei verfügbares Koreferenzsystem fürs Deutsche, wobei wir besonderen Stellenwert auf die Implementierung von Features legen, die die Eigenheiten des Deutschen reflektieren. Das entwickelte Koreferenzwerkzeug erzielt die bisher besten veröffentlichten Ergebnisse auf dem Referenzkorpus TüBa-D/Z.

Für die automatische Auflösung von Bridging existieren bisher für die Sprachen Englisch und Deutsch keine frei verfügbaren Werkzeuge. Basierend auf der besten veröffentlichten Methode für englische Daten implementieren wir daher Auflösungswerkzeuge für Englisch und Deutsch, die beide den aktuellen Stand der Technik definieren.

Abschließend nutzen wir die erstellten Daten und Werkzeuge, um unsere Werkzeuge mit aus der theoretischen Literatur aufgegriffenen Ideen zur Integration von linguistischem Wissen zu verbessern und gleichzeitig die Ideen auf ihre Anwendbarkeit in einem computerlinguistischen Experiment zu überprüfen. Wir zeigen, dass der aus der theoretischen Literatur bekannte Zusammenhang von Koreferenz und Prosodie genutzt werden kann, um unser Koreferenztool zu verbessern. Auf Sprachdaten konnten wir unseren Koreferenzresolver sowohl mit manuell annotierten Pitchakzenten als auch mit Akzenten, die mit einem neuronalen Netz automatisch vorhergesagt wurden, verbessern. In einem zweiten Experiment, in dem die Integration von semantischen Relationen in die Koreferenz- und Bridgingauflösung getestet wurde, hatten automatisch vorhergesagte Meronomiepaare einen signifikant positiven Einfluss auf unseren Bridgingauflöser.

1. Introduction

1.1. Motivation

In natural language understanding, the aim is to extract meaning from text automatically. In order to interpret any sentence in a discourse, we need to know who or what entity is being talked about, as Karttunen (1969)'s early vision illustrates in the following quote.

“

Consider a device designed to read a text in some natural language, interpret it, and store the content in some manner, say, for the purpose of being able to answer questions about it. To accomplish this task, the machine will have to fulfill at least the following basic requirement. It has to be able to build a file that consists of records of all the individuals, that is, events, objects, etc., mentioned in the text, and, for each individual, record whatever is said about it.

”

Since then, a lot of work went into the question of constructing records of the entities mentioned in a text, or, in other words, into grouping references to the same discourse entity together. This includes determining where new entities get introduced in a text and where they get referred to again. In natural language, this task is non-trivial, as humans use pronouns and descriptions to establish dependencies between expressions rather than referring to an entity by always using the same surface form. This is shown in the little extract from *Alice in Wonderland*¹ in Example (1). In the modified version in Example (2), we have replaced every pronoun and paraphrase with the original surface form, which makes the text sound very unnatural.

- (1) It was the White Rabbit, trotting slowly back again, and looking anxiously about as it went, as if it had lost something; and Alice heard it muttering to itself [...]

¹Text from <https://www.gutenberg.org/files/11/11-h/11-h.htm>

1. Introduction

Alice guessed in a moment that it was looking for the fan and the pair of white kid gloves, and she very good-naturedly began hunting about for them, but they were nowhere to be seen.

- (2) It was the White Rabbit, trotting slowly back again, and looking anxiously about as the White Rabbit went, as if the White Rabbit had lost something; and Alice heard the White Rabbit muttering to the White Rabbit [...] Alice guessed in a moment that the White Rabbit was looking for the fan and the pair of white kid gloves, and Alice very good-naturedly began hunting about for the fan and the pair of white kid gloves, but the fan and the pair of white kid gloves were nowhere to be seen.

The task is considered particularly difficult because it involves the use of knowledge and reasoning, as the famous example from the Winograd Schema challenge shows.² Depending on which verb is chosen in the subordinate clause, the pronoun *they* either refers to *the city councilmen* or *the demonstrators*.

- (3) The city councilmen_a refused the demonstrators_b a permit
- because **they** feared violence.
 - because **they** advocated violence.

The subtask of natural language understanding that deals with the fundamental task of determining what entities occur in a text and where they are mentioned again is called coreference resolution and is one of the two tasks which we will investigate in this thesis. It has been proven beneficial for many applications, including question answering (Voorhees et al., 1999), text summarisation (Steinberger et al., 2007), sentiment analysis (Nicolov et al., 2008), textual entailment (Mirkin et al., 2010) and machine translation (Hardmeier and Federico, 2010), to name only a few.

Some expressions are anaphoric, i.e. they are not interpretable on their own without previous context. This includes pronouns such as *he* or definite descriptions such as *the rabbit*, which refer back to entities that have already been introduced and are covered by coreference resolution. There are, however, also context-dependent entities which do not refer to an already introduced entity, but are only related to previously introduced entities. These are called bridging anaphors. When we look at how the little snippet from Alice in Wonderland continues in Example (4), we for example find the expressions *the glass table* and *the little door*, which have not yet been introduced in the text and

²Example taken from Winograd (1972).

are only interpretable because *the great hall* has been mentioned before, and so we can infer that they are part of *the great hall*.

- (4) Everything seemed to have changed since her swim in the pool, and the great hall, with **the glass table** and **the little door**, had vanished completely.

Bridging resolution is the second task that this thesis is concerned with. It is important because it can help in tasks which use the concept of textual coherence, for example Barzilay and Lapata (2008)'s entity grid or Hearst (1994)'s text segmentation. Resolving bridging references is also of help in aspect-based sentiment analysis (Kobayashi et al., 2007), where the aspects of an object, for example the zoom of a camera, are often bridging anaphors. It might also be of use in higher-level text understanding tasks such as textual entailment (Mirkin et al., 2010), question answering (Harabagiu et al., 2001) or summarisation (Fang and Teufel, 2014).

1.2. Research questions

This thesis arose from the interest in developing and improving coreference and bridging resolution, with a focus on English and German. Improving coreference and bridging resolution can be done on different levels, some of them more on the theoretical, some of them more on the computational side. We have identified the following four levels on which contributions can benefit the two tasks: task definition, data creation, tool creation and linguistic validation experiments, as shown in Figure 1.1. In the standard setting, they represent a workflow, i.e. work on data creation requires a satisfying task definition, tool creation is only possible with at least a little bit of evaluation data, and linguistic validation experiments can only be carried out once tools and data are available. However, all these levels are also interlinked and influence each other. Hence, it might sometimes be necessary to go back one or two steps, or, after having conducted some experiments, one might also go back to the first step, task definition, and repeat another round of the pipeline with an improved understanding of the task.

Our research questions reflect the four levels of contribution. Before we can annotate a text with coreference or bridging information or work on tools that can provide automatic annotations, we need to be sure that we have developed a good understanding of the anaphoric phenomenon and that our annotations will be in line with previous definitions and guidelines, or, in cases where previous annotation efforts have shortcomings, we need to address them to avoid coming up with a new, non-compatible scheme.

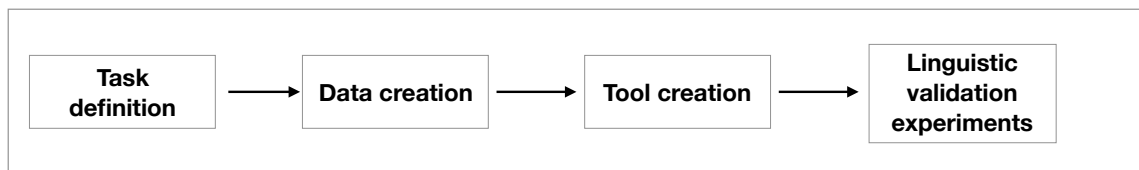


Figure 1.1.: Four levels of contribution

The first research question thus addresses the point whether the tasks are conceptually clear and whether previous work uses compatible annotation guidelines.

Once this question has been answered satisfactorily, we need to think about the corpus resources on which we base the development or improvement of resolution tools. **The research question here is whether there is enough consistently annotated data to enable the creation of automatic tools, including ones making use of statistical algorithms. If not, can we create data resources to fill the research gap?**

With consistently annotated data being available, we can move on to the next step in the pipeline, tool creation. In this step, the availability of coreference and bridging resolution tools is addressed. **Are there openly available tools aiming at providing automatic annotations on unseen text? If not, can we create tool resources to fill the research gap?**

As one of our main interests is to enrich the coreference and bridging resolution systems with linguistically informed new features, we are now at a point in the pipeline where data and tools providing automatic coreference and bridging annotations are available and where we can perform experiments based on these tools and data. On the one hand, the experiments are meant to improve the tools' performances, but they can also give insight into how theoretical claims can be integrated into an automatic setting. The final research question is thus concerned with linguistic validation experiments: **with tools and data being available, do theoretical assumptions about the tasks hold true on actual data? Can we use the theoretical notions to improve the tools?**

1.3. Contributions and publications

Parts of the research described in this thesis have been published in conference proceedings. They are marked as such with the following symbols:

 Publications on coreference resolution  Publications on bridging resolution

For coreference and bridging resolution, the contributions of this work are asymmetrical, i.e. we set our focus on different parts of the pipeline based on previous work and the state of the art. Due to the larger progress in previous work on coreference resolution, we focus on the later steps in the pipeline, mostly tool creation and validation experiments, while also contributing a couple of corpus resources on the data level. In bridging resolution, we encountered problematic issues already in the first step, task definition, and thus set our focus on all four steps in the pipeline. The contributions of this work are summarised in the following.

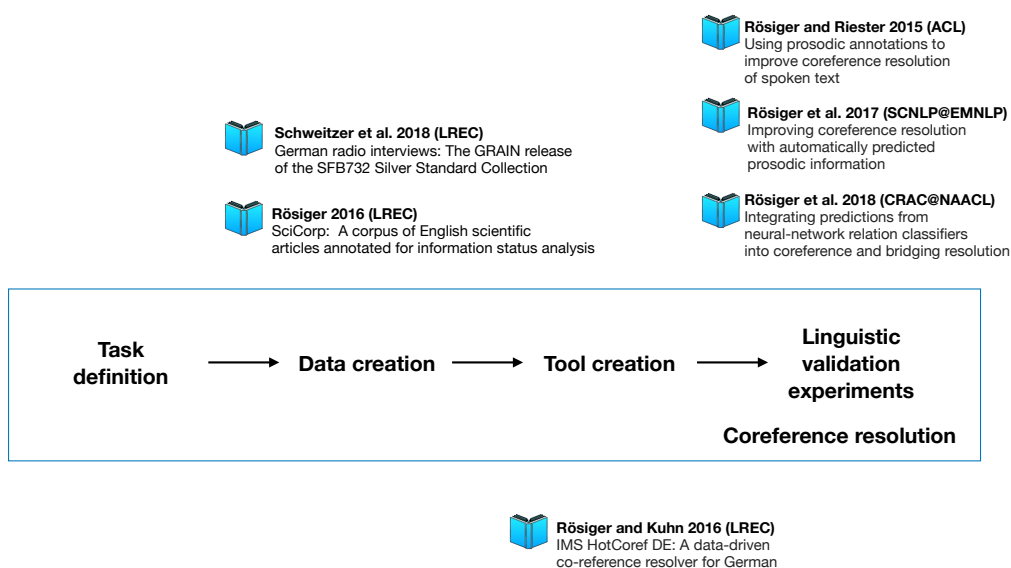


Figure 1.2.: Contributions to coreference resolution

Coreference resolution Coreference is an anaphoric phenomenon which has been studied in theoretical linguistics and semantics since the late nineteenth century (see for example Frege (1892); Russell (1905)). Work on coreference resolution started in the 1960s with some prototypical experiments and has progressed, particularly due to the use of statistical methods, to be one of the most-researched natural language processing

1. Introduction

(NLP) tasks. As a result, the linguistic understanding of the phenomenon as well as the task definition of the NLP task coreference resolution is rather clear, with a couple of exceptions that involve special cases, e.g. the handling of generic entities. In our background chapter, we will give a detailed overview of the definition of coreference and coreference resolution. In terms of data, many large corpora have been created for many languages, including OntoNotes for English (Hovy et al., 2006) and TüBa-D/Z (Naumann and Möller, 2006) for German. There is also a number of smaller corpora for specific domains, e.g. for the biomedical domain or the literary domain. Therefore, our contributions focus mostly on the third and fourth step, tool creation and linguistic validation experiments.

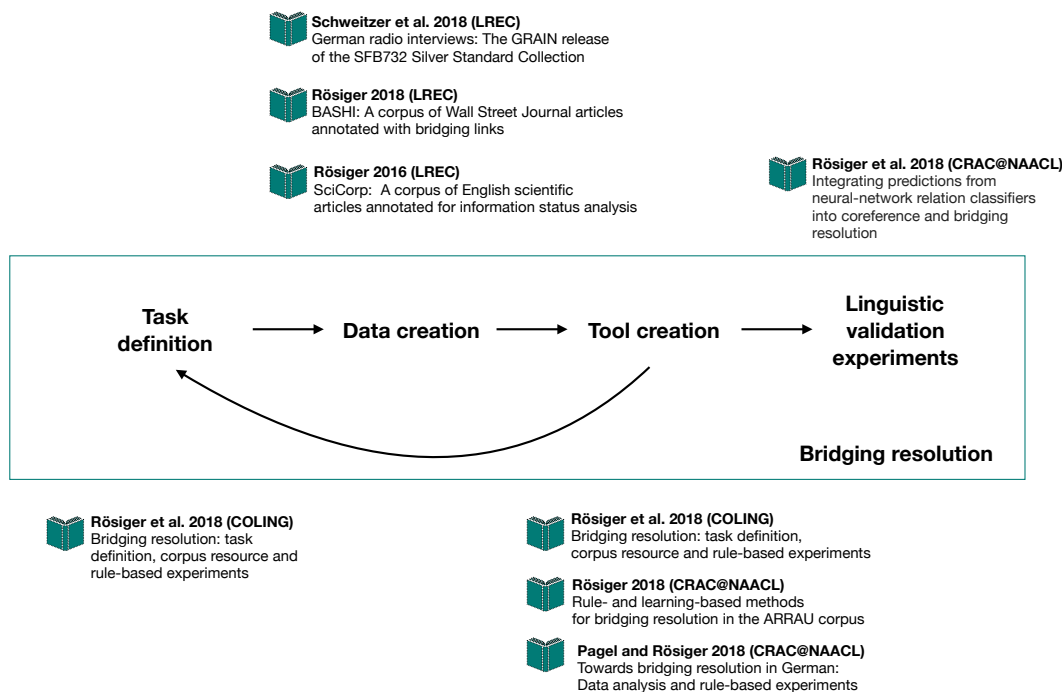


Figure 1.3.: Contributions to bridging resolution

Bridging resolution The phenomenon of bridging was first mentioned in Clark (1975). Back then, it was a term used for a couple of different phenomena, including cases of rhetorical connection and different-head-coreference. As bridging has always been a term with very different understandings and very few corpus resources, the focus in the pipeline is set on all four steps, including task definition and data creation, which enables

us to implement tools and perform validation experiments on the newly created corpus resources.

In the following, we will give an overview of the four contribution levels and list the corresponding publications.

Task definition The task of coreference resolution is generally well-studied and conceptually clear. Apart from minor differences, most previous work uses compatible guidelines.


For coreference resolution, we answer this research question by providing a summary of the anaphoric phenomenon, the task definition as well as best practices in annotating coreference in Section 2.1 and 3.1.

As just noted, bridging is a term that has been used to describe many different phenomena. Some of the critical issues have been controversial for a long time, e.g. the question of definiteness being a requirement for a bridging anaphor. We give an overview on bridging and bridging resolution in Section 2.2 and 3.2. While working on the creation of an automatic bridging resolver, we realised that there was an even more fundamental problem, where non-anaphoric pairs that stand in a particular relation, for example meronymy, are included in the bridging annotation, such as shown in Example (5).

(5) In Europe, **Spain** is the fourth largest country.

To distinguish these two different phenomena, we introduce the concepts of referential vs. lexical bridging and provide a detailed analysis of bridging types (cf. Section 6.3).

Corresponding publication:

-  Ina Rösiger, Arndt Riester and Jonas Kuhn (2018)³
Bridging resolution: task definition, corpus resources and rule-based experiments. Proceedings of COLING. Santa Fe, US 2018.

Data creation For coreference, we provide an overview of available corpora in Section 4.1, where we show that large corpora annotated with coreference information are available for English and German. For bridging, however, there are only a few, small-scale corpus resources. We give an overview of available bridging corpora in Section 4.2

³In this publication, I was responsible for the assessment of the available corpus resources, the implementation of the bridging tool as well as the evaluation of the tool's performance on the respective corpora. The refined bridging definition was the result of a joint effort with Arndt Riester.

1. Introduction

and annotate three corpus resources to overcome the lack of annotated data for bridging. This includes an English corpus of newspaper text called BASHI, an English corpus of scientific texts called SciCorp, as well as a German corpus of radio interviews called GRAIN. All newly created corpora also contain coreference annotations, so that the two anaphoric phenomena can be studied jointly in future experiments.

Corresponding publications:





-  Ina Rösiger (2018)
BASHI: A corpus of Wall Street Journal articles annotated with bridging links. Proceedings of LREC. Miyazaki, Japan 2018.
-   Ina Rösiger (2016)
SciCorp: A corpus of English scientific articles annotated for information status analysis. Proceedings of LREC. Portoroz, Slovenia 2016.
-   Katrin Schweitzer, Kerstin Eckart, Markus Gärtner, Agnieszka Falańska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwen and Jonas Kuhn (2018)⁴. German radio interviews: The GRAIN release of the SFB732 Silver Standard Collection. Proceedings of LREC. Miyazaki, Japan 2018.

Tool creation Many coreference resolvers have been developed for English (see for example Clark and Manning (2016a); Björkelund and Kuhn (2014), etc.). For German, however, there is less work. Over the last couple of years, only the rule-based CorZu (Klenner and Tuggener, 2011; Tuggener and Klenner, 2014) has been developed and improved. Our contribution to coreference resolution in this step is thus an adaptation of an English data-driven coreference resolver to German.

For bridging, there is no openly available resolution system. We thus provide a re-implementation and extension of the state-of-the-art system by Hou et al. (2014) and test the system’s generalisability on our newly developed corpora. We also develop an openly available bridging system for German and perform experiments on German data.

⁴For this resource, I have taken part in the creation of the manual information status annotations. For the paper itself, I have contributed a section describing this part of the resource.

Corresponding publications:

-  Ina Rösiger (2018)
Rule- and learning-based methods for bridging resolution in the ARRAU corpus. Proceedings of NAACL-HLT: Workshop on Computational Models of Reference, Anaphora, and Coreference. New Orleans, US 2018.
-  Janis Pagel and Ina Rösiger (2018)⁵
Towards bridging resolution in German: data analysis and rule-based experiments. Proceedings of NAACL-HLT: Workshop on Computational Models of Reference, Anaphora, and Coreference. New Orleans, US 2018.
-  Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Sadat Moosavi, Ina Rösiger, Adam Roussel, Alexandra Uma, Olga Uryupina, Juntao Yu, Heike Zinsmeister⁶. Anaphora resolution with the ARRAU corpus. Proceedings of NAACL-HLT: Workshop on Computational Models of Reference, Anaphora and Coreference. New Orleans, US 2018.
-  Ina Rösiger and Jonas Kuhn (2016)⁷
IMS HotCoref DE: A data-driven co-reference resolver for German. Proceedings of LREC. Portoroz, Slovenia 2016.

Linguistic validation experiments We address this research question by performing two experiments, which are meant to motivate further experiments using the available tools and data to assess theoretical assumptions about the tasks. The first experiment deals with the question of how prosodic information can be used to improve coreference resolution in spoken data. We show that using both manually annotated and automatically predicted prosodic information significantly increases results. In the second experiment, we test the use of automatically predicted semantic relations for coreference and bridging resolution. We show that our newly integrated features significantly improve our bridging resolver, but not our coreference resolver.





⁵I was responsible for the implementation of the bridging tool and for the experiments on DIRNDL, while Janis Pagel performed experiments on the newly created GRAIN corpus.

⁶I contributed my shared task results in the form of evaluation tables.

⁷I was responsible for the creation of the coreference tool and for writing the paper.

1. Introduction

Corresponding publications:

-  Ina Rösiger and Arndt Riester (2015)⁸
Using prosodic annotations to improve coreference resolution of spoken text. Proceedings of ACL-IJCNLP, Beijing, China 2015.
-  Ina Rösiger, Sabrina Stehwien, Arndt Riester, Ngoc Thang Vu (2017)⁹
Improving coreference resolution with automatically predicted prosodic information. 1st Workshop on Speech-Centric Natural Language Processing (SCNLP). Copenhagen, Denmark 2017.
-   Ina Rösiger, Maximilian Köper, Kim Anh Nguyen and Sabine Schulte im Walde (2018)¹⁰. Integrating predictions from neural-network relation classifiers into coreference and bridging resolution. Proceedings of NAACL-HLT: Workshop on Computational Models of Reference, Anaphora, and Coreference. New Orleans, US 2018.

1.4. Outline of the thesis

This thesis has three main parts. In the first part, we give some background on the anaphoric phenomena and the computational modelling of coreference and bridging.

Chapter 2 introduces the basic concepts of coreference, bridging and anaphoricity. We also analyse differences in annotation guidelines and divergent understandings of the phenomena.

Chapter 3 explains the NLP tasks coreference and bridging resolution and gives an overview of previous automatic approaches.

In the second part, data and tool creation, we present available corpus resources in Chapter 4, before we introduce our newly annotated corpora. To overcome the lack of available data for bridging, we annotate a newspaper corpus called BASHI. In order to be

⁸Arndt Riester and I jointly developed ideas taken from the theoretical literature to be tested in a coreference resolver. I was responsible for integrating the ideas into the resolver and evaluating different scenarios. The paper was written jointly with Arndt Riester.

⁹Sabrina Stehwien provided the automatically predicted prosodic information, which I integrated into the coreference resolver. I was also responsible for the evaluation of the newly integrated prosodic information and for the error analysis. The paper was written in a joint effort.

¹⁰Maximilian Köper and Kim-Anh Nguyen provided the automatically predicted relations for word pairs that I have extracted from the corpora used in the experiments. I was responsible for integrating the predicted information into the coreference and bridging tools and also for the evaluation of the newly integrated information. The paper was written in a joint effort.

able to test the generalisability of automatic approaches, we also create SciCorp, a corpus of a different domain, namely scientific text. For German, we annotate a corpus of radio interviews with bridging information. All corpora also contain coreference annotations.

Chapter 5 addresses the adaptation of a data-driven coreference resolver for English to German, where we focus on the integration of features designed to address specificities of German. The tool achieves state-of-the-art performance on the latest version of the benchmark dataset TüBa-D/Z version 10.

Chapter 6 is devoted to bridging resolution, where we reimplement the state-of-the-art approach for bridging resolution in English by Hou et al. (2014) and test the generalisability of the approach on our own new corpora as well as other available corpus resources. Besides the expected out-of-domain effects, we observe low performance on some of the in-domain corpora. Our analysis shows that this is the result of two very different phenomena being defined as bridging, which we call referential and lexical bridging. We think that the distinction between referential and lexical bridging is a valuable contribution towards the understanding of the phenomenon of bridging and that it can also help design computational approaches. The diverging bridging annotations became obvious when we worked on a shared task submission for the first shared task on bridging. After discussing the different properties of the two types of bridging, we compare our rule-based system against a learning-based one and design new rules to also handle lexical bridging. We also create a bridging resolution system for German, where we investigate new rules and the role of coreference information.

The third part addresses two linguistic validation experiments.

Chapter 7 explains how prosodic information can be used to improve coreference resolution. Our results show that both manually annotated and automatically predicted prosodic information improve a coreference system for German.

Chapter 8 explores the use of automatically predicted semantic relations for both coreference and bridging resolution. While our coreference resolver does not benefit from the newly added information, our bridging resolver can be improved by including automatically predicted meronymy pairs.

Part I.
Background

2. Anaphoric reference

Research Question 1: Task definition:

Are the tasks conceptionally clear? Does previous work use compatible annotation guidelines or are there very different understandings of the tasks?

This section gives an overview on anaphora and introduces the two main phenomena, coreference and bridging. Doing so, we give an answer to the question of whether the tasks are conceptionally clear (Research Question 1), to which we will come back in Section 4.

Reference Reference, in traditional semantics, is a relation between certain expressions in a text and objects of our thought (Bussmann, 1990). Hereby, the referent is the mental entity to which is referred, and the referring expressions is a noun phrase (NP) in a text which identifies some individual object. Reference thus denotes the ability of language expressions to refer to discourse entities, which may be linked to extralinguistic objects (Zikánová et al., 2015). In Figure 2.1, the relation between discourse or mental entities and referring expressions is illustrated.

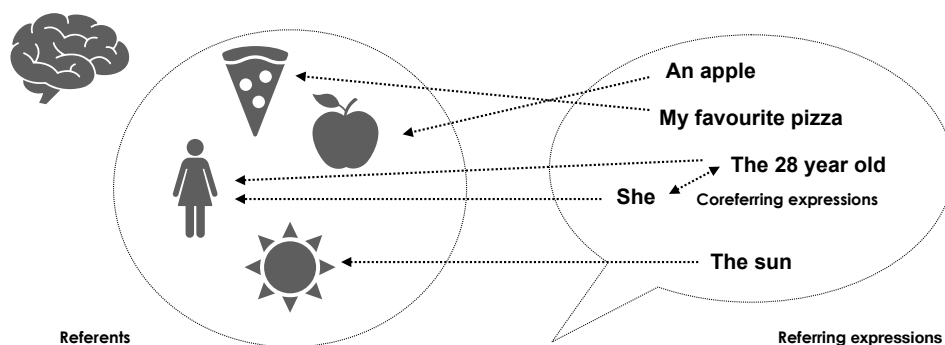


Figure 2.1.: Reference as the relation between referring expressions and referents

2. Anaphoric reference

In that sense, reference can either be specific, as in Example (1), describing a particular specimen of a class, or generic, where the reference holds between any member of the class, as shown in Example (2).

- (1) Have you seen **my cat**?
- (2) **Cats** are small, typically furry, carnivorous mammals. **Cats** are the second-most popular pet in the U.S.

A second interpretation of the term reference comprises textual links to preceding or following context. In case of textual context, we speak of anaphoric reference, while extra-textual reference is called exophora (Zikánová et al., 2015). Anaphoric reference means that an expression cannot be interpreted on its own and that it refers back to an already introduced entity or event which enables us to identify the referent that could otherwise not be established. In other words, reference to an entity that has been previously introduced into the discourse is called anaphora and the referring expression is said to be anaphoric. In Example (3), the pronoun *he*, for example, refers back to *Peter*.

- (3) Peter went into the supermarket. **He** bought a pizza.

Anaphoric reference can comprise relations of identity, i.e. coreference, where two or more expressions have the same referent, as in Example (3) or shown in Figure 2.1, where the expressions *the-28-year-old* and *she* refer to the same person, or bridging (Clark, 1975), where the relation to the expression to which it refers back is only one of association, and the referents are related, but not identical.

- (4) We went to see a movie last night. **The tickets** were rather expensive.

In Example (4), it is clear that *a movie* and *the tickets* do not refer to the same entity, but one cannot interpret the expression *the tickets* without the previously introduced expression *a movie*.

In general, the referring expression which cannot be interpreted on its own is called the anaphor (or sometimes called anchor), while the expression to which it refers back is called the antecedent. In the remainder of this thesis, anaphors are marked in boldface and their antecedents are underlined.

There is also the special case of cataphora or backward anaphora (Carden, 1982) in which the context-dependent expression, in this case called cataphor, appears before the antecedent (sometimes called postcedent).

- (5) Speaking in **his** home state of Texas, Mr Cruz urged other Republicans to quit the race and unite against Mr Trump.

Anaphors can be pronominal, as in Example (6) or nominal as in Example (7).

- (6) Peter went into the supermarket. **He** bought a pizza.
- (7) Peter bought a new book yesterday. **The novel** turned out to be very entertaining.

In fact, a lot of different pronouns, NP types, as well as adverbs can function as anaphors, as the following enumeration shows.

- Pronouns

- Personal pronouns:

- (8) Peter likes watching football matches. **He** also likes baseball.

- Demonstrative pronouns:

- (9) My friend just played the piano for us. **That** was great.

- Relative pronouns:

- (10) Have you seen the man **who** wears a striped shirt?

- Possessive pronouns:

- (11) My sister is saying that the shirt is **hers**.

- Reflexive pronouns:

- (12) Peter washed **himself**.

- Reciprocal pronouns:

2. Anaphoric reference

(13) The five people looked at **each other**.

– Indefinite pronouns:

(14) He greeted the students. **One** raised his hand in greeting.

- Definite and demonstrative NPs:

(15) Peter gave Bill a strange look. **The man** was crazy.

- Temporal, local and manner adverbs:

(16) The wedding is at 2 pm. See you **then!**

- Indefinite NPs (in bridging):

(17) Starbucks is planning their own take at the unicorn frappuccino.
One employee accidentally leaked a picture of the secret new drink.

2.1. Coreference

Coreference and anaphora Coreference and anaphora are both basic means of achieving text cohesion, as for example studied in Halliday and Hasan (1976). However, the two terms are not synonymous.

Anaphora, as explained in the previous section, is the phenomenon that anaphoric expressions are dependent on the previous context, and need to be linked to their antecedent in order to be interpretable.

Coreference is defined as the identity of referents signified by language expressions in discourse (Zikánová et al., 2015). As such, anaphoricity, i.e. context-dependence, is not a requirement for two expressions to be considered coreferent. Often, coreference and anaphora occur simultaneously, e.g. in Example (3). However, not all coreferring entities are anaphoric, e.g. in Example (18), where the second and third occurrence of *Google* is not dependent on the first occurrence, but of course, all expressions have the same referent (Google, the company).

(18) US lawmakers on Wednesday sent a letter to Google CEO Sundar Pichai expressing concerns regarding Huawei's ties with the Chinese government. The

lawmakers said the strategic partnership between **Google** and Huawei on instant messaging, announced in January, poses serious threats to US national security and consumers. The letter also addressed **Google’s** recent refusal to renew a research partnership, Project Maven, with the Department of Defense.

For the sake of being compatible with previous research, we nevertheless use the term coreference anaphor for all expressions which are coreferent with some expression in the previous context, e.g. also for *Google* in Example (18), and the term antecedent for coreferred expressions (in the context of coreference, of course, as there are also bridging antecedents).

In the following section, we will introduce important concepts and special cases related to coreference.

Predication The theory of reference is based on logical semantics (Frege, 1892; Strawson, 1950), where the relation between language expressions and referents was studied. One notion that was distinguished from the referential use of NPs already back then was the predicative use, sometimes also called attributive use (Donnellan, 1966).

(19) Donald Trump is the US President.

In Example (19), the expressions *Donald Trump* and *the US President* are not coreferent, as being *the US President* is a property of *Donald Trump*.

Genericity While the distinction between predicative and referential use of NPs seems to be generally accepted and is considered in most guidelines, opinions on generic entities and their ability to refer have been more diverse, and as a result, also annotated rather diversely across many corpora. It is clear that we want to distinguish between generic and non-generic entities, like in Example (20), where coreference class 1 refers to the generic class of lions and coreference class 2 refers to the specific lions at the zoo.

(20) Today I saw {a bunch of {Lions}₁}₂ at the zoo. **{They}**₁ are great animals. **{The lions in our zoo}**₂ seemed sad, though.

Reference to the type differs from reference to a concrete object, as it does not need to refer to all objects of that type, but is rather a statement about the prototypical member of the class (Zikánová et al., 2015). In Example (21), while it may be true that most cats do not like water, it might not be true for all cats. However, generic entities should

2. Anaphoric reference

still be considered in coreference, as the repetition of generic entities is important for text cohesion, and of course, generic entities can be pronominalised.

(21) **Cats** do not like water.

As a result, the handling of generic entities in the annotation of coreference is a controversial one. Some work left them out completely, while others have suggested that generic anaphoric expressions always start their own coreference chain, and should thus just be linked back to their own antecedent and not to other occurrences of the same entity, e.g. in the OntoNotes guidelines (Weischedel et al., 2011). This then accounts for pronominalised generic entities but it does not capture the repetition of generic entities throughout a text.

Besides generic entities, there is a number of other special cases, which are worth mentioning.

Abstract anaphora Coreferent anaphors can also have non-nominal antecedents, e.g. verbal phrases (VPs) or clauses, as shown in Examples (22) and (23). Because of the often abstract nature of these expressions, this phenomenon is called abstract anaphora or event reference (Asher, 1993).

(22) We found that eating a lot of sugar is detrimental to your health.
This has also been shown in previous studies.

(23) I heard him singing last night. **That** was funny.

As in most of the work on coreference resolution, in our experiments, we focus on nominal antecedents.

Aggregation or split antecedents Often, the anaphoric expression refers to a set of referents, e.g. in Example (24), where the pronoun *they* refers back to the set of the referents *Peter* and *Sam*. As they occur in a conjunction, they can be captured by a single antecedent. However, sometimes the two expressions appear separated by other syntactical elements, e.g. verbs, as in Example (25). Although conceptionally it is the same case as in Example (24), the fact that in an annotation setting there are now two links required to express that this anaphor refers to the set of two entities has caused some problems. In some previous work, for example in the PDT corpus (Hajič et al., 2018), the second case is thus treated as bridging of the type **set-subset** rather than

coreference. We think that aggregation is a special case of coreference, which should not be mixed with bridging.

- (24) Peter and Sam met a couple of years ago. **They** now like to go on holiday together.
- (25) Peter met Sam a couple of years ago. **They** now like to go on holiday together.

Anaphoric zeros Anaphoric zeros or zero anaphora (Saeboe, 1996), the textual ellipsis of a dependent element that can be determined from the context, occur in many languages (e.g. Russian, Japanese, Chinese, etc.). As they do not occur in nominal coreference in German and English, they are not a focus of this work.

Bound coreference For pronouns, some work uses the concept of bound coreference. This means that pronouns appear in quantified contexts in which they are considered to be bound.

- (26) Every female teacher raised **her** arm.

In this theory, *her* does not refer to anything but behaves like a variable bound to the quantified expressions *every teacher*. In practice, the distinction between bound pronouns and other non-bound coreference is not always made. In this thesis, we include these cases in coreference resolution, but do not distinguish between bound and non-bound coreference.

Near-identity As the question of the identity of referents is not always trivial, Recasens and Hovy (2010a) have introduced a third concept in between coreference and bridging, which they call near-identity, which has been picked up by others, e.g. in Grishina (2016). Near-identity is defined to hold between an anaphor and an antecedent whose referents are almost identical, but differ in one of four respects: name metonymy, meronymy, class or spatio-temporal functions. Example (27), taken from Recasens and Hovy (2010a), for example contains a near-identity relation between *Jews* and *the crowd*.

- (27) Last night in Tel Aviv, Jews attacked a restaurant that employs Palestinians. "We want war", **the crowd** chanted.

They believe that most of the near-identity types can be grasped on the level of grammar, semantics and concepts. However, the concept has also been criticised, e.g. by Zikánová

2. Anaphoric reference

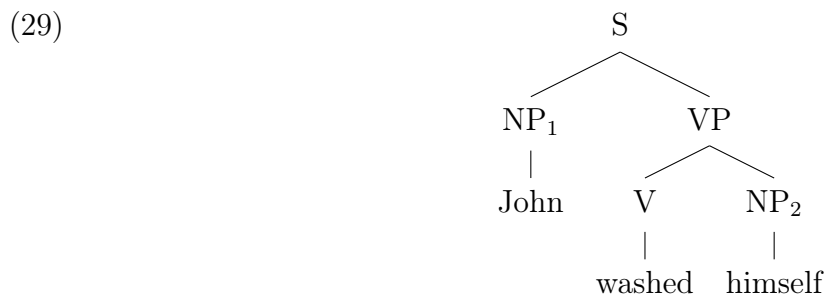
et al. (2015), who argue that coreference is a property of the discourse world, realised on the pragmatics level only. They claim that introducing an additional term is not helpful in the understanding of anaphora and mixes up separate levels of language systems and speech. We will get back to this concept in our refined bridging definition in Section 6.3.

Constraints and preferences In order for two expressions to be coreferent, they need to fulfil a number of constraints. While some of them are hard constraints, some are tendencies or preferences rather than hard constraints. The first type of constraints considers agreement constraints. Anaphor and antecedent typically agree in number, person and gender. There are, however, exceptions. Regarding the number agreement, some expressions can be referred to with a singular or plural pronoun, e.g. *the police*. In languages with grammatical gender, gender agreement does not need to hold, at least not between two non-pronominal noun phrases, such as shown in Example (28).

- (28) DE: Der Stuhl [masc.] ... **die Sitzgelegenheit** [fem.] ...
 das Plastikmonster [neut.] .
EN: the chair ... **the seating accommodation** ... **the plastic monster** .

Syntactic constraints are generally thought to be hard constraints and are based on the binding theory (Chomsky, 1981). While a full explanation of the binding theory and its underlying assumptions on syntactic structure would go beyond the scope of this thesis, we will give a short explanation of its main ideas based on simplified syntactic structures. Please refer to Chomsky (1981) for more details.

One important notion of the binding theory with respect to coreference is the concept of c-commanding nodes in the syntax tree. NP_x c-commands NP_y if and only if neither NP_x nor NP_y dominates the other; and every branching node that dominates NP_x also dominates NP_y . Or, in simpler terms, c-command summarises the relationships brother, uncle, great-uncle, great-great-uncle, etc. In Example (29), NP_1 c-commands NP_2 .



There are three main principles of the binding theory. The first one is concerned with reflexive pronouns and states that reflexives must have local antecedents (must be c-commanded). Local in this case means that they must be bound in their governing category, the clause containing the anaphor.¹ Consider Example (30), for which we have already seen a simplified syntactic structure in Example (29).

(30) John washed **himself**.

(31) * John asked Mary to wash **himself**.

In contrast, Example (31) is ungrammatical because the reflexive pronoun *himself* does not have a local, c-commanding antecedent and can thus not be bound in its governing category.

The second principle is that personal pronouns must not have local antecedents, i.e. must not be c-commanded. This means that when we replace the reflexive pronoun in Example (31) with a personal pronoun like *him*, as in Example (32), the resulting sentence is grammatical.

(32) John asked Mary to wash **him**.

The third principle is that a full NP cannot have a local (=c-commanding) antecedent. In Example (33), it is not possible that both occurrences of *John* are coreferent because the first occurrence c-commands the second.

(33) * John saw **John**.

Another constraint, or rather a preference, is called selectional restriction (Chomsky, 1988), where, in Example (34), the verb *eat* requires that its direct object denote something that can be eaten, such as *a pizza*, but not *a supermarket*.

(34) John bought a pizza from the supermarket. Peter ate **it**.

Whereas the constraints above are typically considered hard constraints, there are a number of soft factors that affects the salience of an expression, i.e. the degree of accessibility in the addressee's consciousness at the time of the speakers utterance (Prince, 1981). This has the effect that we favour some expressions as a potential antecedent over other expressions.

¹Note that this definition is overly simplistic and not entirely accurate, for details see Chomsky (1981).

2. Anaphoric reference

For example, the verb semantics in Example (35) and (36) influences the different preferences, as the implicit cause of a shouting event is considered to be its object, whereas the implicit cause of a calling event is considered to be its subject (Garvey and Caramazza, 1974). Thus, the higher degree of salience for the entity in this argument position leads to the different preferences.

(35) Peter called Adam. **He** had broken the TV.

(36) Peter shouted at Adam. **He** had broken the TV.

Another constraint is the non-accessibility of expressions under negation, e.g. in Example (37), where it is not possible to refer to *the house* with the pronoun *it*, as *the house* appears under negation and is thus not accessible for future (co)reference (Kamp, 1981).

(37) * Peter did not buy a house. **It** was big.

Recency is another factor that makes expressions more salient, i.e. also more likely to be referred to. In Example (38), we have a tendency to favour *a burger* as the antecedent, rather than *a pizza*, simply due to recency.

(38) Bill is eating a pizza. John is eating a burger. Mary wants a taste of **it**, too.

Grammatical roles are another factor which influences the salience of antecedents (Alshawi, 1987). It is generally assumed that entities introduced in subject position are more likely to be referred to by a pronoun than entities in object position, which in turn are considered more salient than other grammatical roles, such as prepositional objects or adjuncts.

(39) John went to the supermarket with Bill. **He** bought a pizza.

Plain word repetition also affects salience, as can be seen in the following example.

(40) John went to the supermarket with Bill. **He** bought a pizza. Later, **he** met with Peter. **He** had a nice time.

Parallelism is another contributing factor to salience, i.e. pronouns are more likely to refer to those entities that do not violate syntactically parallel constructions. In Example (41), *Peter* is the preferred antecedent, although *Bill* is in subject position.

(41) Bill took Peter to the supermarket. Sue took **him** to the park.

The role of prosody Prosody, or more specifically accentuation, is an important means to influence the meaning of language expressions.

- (42) If the going gets tough you don't want a criminal LAWyer – you want a CRiminal lawyer. (J. Pinkman, Breaking Bad)

In Example (42), the different accentuations lead to different interpretations: in one case, we refer to lawyers specialised in criminal law, in the other case we refer to a lawyer who is also a criminal.

In spoken language, pitch accents are often used to emphasise new information, while given (=coreferent) information is often deaccented (Terken and Hirschberg, 1994). Thus, it is important to include prosody in the analysis of coreference, as default preferences and interpretations can be overridden by prosody. Consider Example (43), taken from Lakoff (1971), where the default interpretation without prosody is that *he* refers to *John* and *him* refers to *Peter*, due to the preference for role parallelism. In Example (44), we can override the default interpretation by accenting the two pronouns.

- (43) {John}₁ called {**Peter**}₂ a republican. And then {**he**}₁ insulted {**him**}₂.
- (44) {John}₁ called {**Peter**}₂ a republican. And then {**HE**}₂ insulted {**HIM**}₁.

2.2. Bridging

Coreferential anaphoric links are not the only important type of anaphoricity that is important in order to establish textual coherence in a text. Consider Example (45), where the definiteness of *the front door* signals uniqueness, which can only be fulfilled if the reader accepts the implication that this door is part of the house mentioned in the preceding sentence.

- (45) She spends nearly four hours measuring each room in the 50-year-old house. Afterwards, she snaps photos of **the front door** and **the stairway**.

This is an example of a bridging (Clark, 1975; Asher and Lascarides, 1998) or associative anaphor (Hawkins, 1978): an expression which cannot be interpreted without previous context. To make up for this, we need to build a “bridge” in order to link the expression to previously mentioned material (Riester and Baumann, 2017). In contrast to coreference, the antecedent is not identical but associated. In other words, bridging is an anaphoric phenomenon where the interpretation of a bridging anaphor is based on the non-identical associated antecedent.

- (46) Our correspondent in Egypt is reporting that **the opposition** is holding a rally against **the constitutional referendum**.
- (47) What is the movie about? **The answer** isn't easy.

One can think about bridging anaphors as expressions with an implicit argument, e.g. *the opposition* (in Egypt) or *the answer* (to this question). The term bridging has first been introduced in Clark (1975), where a broad classification of different types was presented. In this work, three main groups are distinguished:

- Set-subset:

(48) I met two people yesterday. **The woman** told me ...

(49) I swung three times. **The first time** ...

- Indirect reference by association:

(50) I went shopping. **The walk** was nice.

(51) I looked into the room. **The size** was overwhelming.

- Indirect reference by characterisation

(52) John was murdered yesterday. **The murderer ...**

However, the definition of bridging in Clark (1975) is quite broad, also covering cases which are nowadays covered by coreference.

Bridging has also been studied in Hawkins (1978), where the term **associative anaphora** is used to refer to typically definite associative anaphors, such as *the bride* in Example (53), that can be interpreted because a previous expression has triggered the reader's associations, in this case *the wedding*.

(53) I went to a wedding yesterday. **The bride** was a friend of mine.

Furthermore, Prince (1981, 1992) introduced the term **inferrables** to refer to anaphors that can be inferred from certain other discourse entities already mentioned. In Prince (1992) she introduced the term **information status** to describe the degree of givenness of a language expression and presented a classification based on the notions **hearer-new/hearer-old** and **discourse-new/discourse-old**. We will come back to the notion of bridging as an information status category in Section 4.

Based on this work, Nissim et al. (2004) picked up the term information status, distinguishing **old** entities from **new** ones, and **mediated** in between. The category **mediated** comprises a number of different types, including generally-known entities like *the pope*, but also **mediated/bridging**. Bridging as a subcategory of information status has been applied in many works (e.g. Markert et al. (2012); Baumann and Riester (2012), among others).

As of now, bridging has been studied in many theoretical studies (Clark, 1975; Hawkins, 1978; Hobbs et al., 1993; Asher and Lascarides, 1998; Prince, 1981) as well as in corpus and computational studies (Fraurud, 1990; Poesio et al., 1997; Vieira and Teufel, 1997; Poesio and Vieira, 1998; Poesio et al., 2004; Nissim et al., 2004; Nedoluzhko et al., 2009; Lassalle and Denis, 2011; Baumann and Riester, 2012; Cahill and Riester, 2012; Markert et al., 2012; Hou et al., 2013a,b; Hou, 2016b; Zikánová et al., 2015; Grishina, 2016; Roitberg and Nedoluzhko, 2016; Riester and Baumann, 2017; Hou, 2018). One big issue is that, unlike in work on coreference, these studies do not follow an agreed upon definition of bridging. On the contrary, many different phenomena have been described as bridging. As a result, guidelines for bridging annotation differ in many respects so that they cannot be easily combined to create a larger bridging corpus resource. The latter would, however, be necessary to further research in this area, as statistical approaches

2. Anaphoric reference

to bridging resolution are limited due to the limited corpus size, as for example stated in Hou (2016b).

This section summarises the main issues of diverging bridging definitions.

Overlap with coreference One issue that came up in early work on bridging and is still present in some work is the overlap with coreference anaphora. As mentioned above, Clark (1975) proposed a very broad definition, including the anaphoric use of NPs that have an identity relation with their antecedent, e.g. in

(54) I met a man yesterday. **The man** stole all my money.

While it is nowadays non-controversial that these coreferent cases should not fall under the label of bridging, the more difficult cases of coreference where the anaphor and the antecedent do not share the same head but are in a synonymy, hyponymy or metonymy relation, are sometimes treated as bridging, e.g. in Poesio and Vieira (1998), among others. We think that independent of the surface form, identical context-dependence should be covered as a case of coreference.

(55) I met a man yesterday. **The bastard** stole all my money.

Clark (1975) and Asher and Lascarides (1998) also included rhetorical relation or connection cases, e.g. in

(56) John partied all night yesterday. He's going to get drunk **again** today.

While these are interesting cases of anaphoric use, most work nowadays limits anaphors to nominal referring expressions.

Definiteness Another important point of discussion is the question of whether definiteness should be a requirement for bridging anaphors. Many studies (Poesio and Vieira, 1998; Baumann and Riestler, 2012), among others, have excluded indefinite expressions as potential bridging candidates, stating that indefinite expressions introduce new information that can be processed without the context of the previous discourse. Löbner (1998) suggested that bridging anaphors can also be indefinite, as these indefinite expressions can occur in **whole-part** or **part-of-event** relations, with the consequence that many studies have linked them as bridging (e.g. in ISNotes, and others).

(57) I bought a bicycle. **A tire** was already flat.

(58) Standing under the old oak tree, she felt **leaves** tumbling down her shoulders.

Riester and Baumann (2017) suggested restricting the annotation of bridging to definite expressions as part of their information status annotation of referring expressions (r-level) and to treat lexical relations (in indefinite and definite expressions) on another level (called the l-level). We will get back to this question in Section 6.

Pre-defined relations Another common issue is the restriction of bridging to pre-defined relations, such as **part-of**, **set-membership**, **possession** or **event** relations, e.g. in the Switchboard corpus (Nissim et al., 2004). Other corpora do not make such limitations (e.g. ISNotes). We believe that bridging is a versatile phenomenon that cannot be satisfactorily captured with pre-defined relations.

Furthermore, some work has excluded certain relations, e.g. comparative anaphora in the ISNotes corpus (Markert et al., 2012), from the bridging category arguing that they can be found by surface markers, such as *other*, *another* etc., for instance in Example (59).

(59) About 200,000 East Germans marched in Leipzig and thousands more staged protests in **three other cities**.

Comparative anaphora have different properties than “regular bridging” cases, as they indicate co-alternativity, e.g. a relationship on equal terms, between the antecedent and the anaphor, while for typical bridging cases, the relation between the anaphor and the antecedent is a hierarchical one, with the bridging anaphor being subordinate to the antecedent.

Bridging-contained Related to bridging is a special case where the antecedent modifies the bridging anaphor, sometimes called **containing inferrable** (Prince, 1981) or **bridging-contained** (Baumann and Riester, 2012), as the antecedent is a syntactic argument within the markable, as shown in Example (60), (61) and (62).

(60) **the windows** in the room

(61) the room’s **windows**

(62) their **interest**.

As these cases of bridging are not context-dependent as such, we think that they should not be included in the regular bridging category.

3. Related NLP tasks

Research Question 1: Task definition:

Are the tasks conceptionally clear? Does previous work use compatible annotation guidelines or are there very different understandings of the tasks?

3.1. Coreference resolution

Task definition Coreference as an anaphoric phenomenon has been described in Section 2.1. The related NLP task of noun phrase coreference resolution is about determining which NPs in a text or dialogue refer to the same discourse entities. Many definitions revolve around NPs that refer to real-world entities, e.g. in Ng (2010)'s definition. However, the existence of objects in the real world is not essential, as we can also have reference to fictional or hypothetical characters or objects of our thoughts.

Prior to work on noun phrase coreference resolution, which has become one of the core topics in NLP, there was much work on pronoun resolution (Hobbs, 1979; Lappin and Leass, 1994), which aimed at finding an antecedent for every pronoun. Coreference resolution nowadays focuses on grouping references to the same discourse entity together, where language expressions referring to the same entity can be understood as an equivalence class, set or chain. This means that, in contrast to anaphora resolution, which aims at finding an antecedent for each anaphor in a text, coreference resolution is about partitioning the set of discourse entities (NPs) in a text into equivalence classes. This also includes named entities (NEs) or non-anaphoric nominal expressions.

Example (1) shows the different types of entities involved, marked with numbers to highlight the equivalence classes. Note that the classes contain truly anaphoric expressions (*the former secretary of state*) as well as non-anaphoric coreferent expressions (*Trump*).

- (1) {Democrat Hillary Clinton}₁ and {Republican Donald Trump}₂ have won the most states on the biggest day of the race for the US presidential nominations. {{The former secretary of state}₁ and {Trump, a property tycoon,}₂}₃ entered

3. Related NLP tasks

Super Tuesday as favourites to win the vast majority of states for {their}₃ respective parties. {Mr Trump}₂ won seven states while {{his}₂ closest rival, Ted Cruz,}₄ took three. Speaking in {his}₄ home state of Texas, {Mr Cruz}₄ urged other Republicans to quit the race and join {him}₄ against {Trump}₂.

Coreference resolution can be divided into two subtasks. The first subtask is to figure out what language expressions need to be partitioned into clusters, i.e. determining the span of (mostly) NPs¹. These language expressions that make up the set of discourse entities are generally called mentions or markables. The second subtask is then to group these mentions into equivalence classes referring to the same entity.

Computational approaches to coreference resolution Coreference resolution has become one of the core NLP tasks, with its own track at most NLP conferences. It dates back to the 1960s with the very first prototypical experiments and the 1970s, when work on pronoun resolution began, e.g. the syntax-based pronoun resolution algorithm for 3rd person pronouns by Hobbs (1979). The method requires a syntactic parser as well as a morphological number and gender checker. It searches syntactic trees of the current and preceding sentences and stops when it finds a matching NP. The algorithm starts with a right-to-left-search in the current sentence and checks if it finds a matching NP node as antecedent that agrees in number and gender and is not c-commanded (except for reflexive pronouns). If not, we move on to the preceding sentence and perform left-to-right-search in a breadth-first manner. If still none is found, we check left-to-right in the sentence of the pronoun, to check for cataphora.

In subsequent years, a couple of rule-based systems were developed based on linguistic information, such as the salience-based model by Lappin and Leass (1994). They employ a simple weighting scheme for recency and syntactically-based preferences, taking into account grammatical roles as well as recency. There are two types of steps, a discourse model update when new sentences are read into the model and a pronoun resolution step each time a pronoun is encountered. The discourse model step updates weights for introduced entities, by adding weights according to the factors mentioned above. Several NPs denoting the same referent are hereby treated as an equivalence class, i.e. the weights for each factor are summed. In the pronoun resolution step, we compute two extra factors that can only be computed based on the pronoun and possible antecedent

¹The question of which language expressions should be considered is non-controversial. NPs are always included and VPs or clauses are typically also allowed as antecedents. However, the latter are typically ignored in automatic approaches.

pair: role parallelism and whether the two are cataphoric. The entity that has the highest score and does not violate syntactic (*c-command*) and morphological constraints is considered the most salient and proposed as the antecedent. The progress in theoretical work on local coherence and salience, e.g. by Grosz and Sidner (1986) or Grosz et al. (1995), enabled the development of a number of additional pronoun resolution algorithms (Tetreault (1999); Brennan et al. (1987), among others).

With the creation of large, manually annotated corpora, e.g. the MUC-6 and MUC-7 corpora, rule-based approaches were soon superseded by probabilistic, data-driven models. Ge et al. (1998) presented a statistical model to pronoun resolution, Soon et al. (2001) the first machine learning approach to nominal coreference resolution. The basic idea follows that of any supervised approach: we start with gold annotated data, extract positive examples of coreferent pairs and negative examples of non-coreferent pairs and then define features for the pairs to create a classification task. As such, this represents the mention pair approach, where we pair mentions and let the classifier decide whether this individual pair is coreferent or not. Soon et al. (2001) use simple features including string match, NP type, number and semantic class agreement, to name a few. To overcome the problem that there are much more non-coreferent pairs in the training data than coreferent pairs, they restrict non-coreferent pairs to those appearing between a coreferent pair, combining the gold anaphor with all mentions that appear between the coreferent anaphor and the coreferent antecedent. As a decoding strategy, they use closest-first decoding, i.e. if for one anaphor there are several antecedents for which the classifier determined the pair to be coreferent, the closest is chosen. The mention pair model has a couple of rather obvious weaknesses, the main one being that the pairs are considered individually by the classifier, and so the transitivity which is inherent in coreference chains cannot be ensured. For example, the classifier can predict pair A-B to be coreferent as well as pair B-C, but not pair A-C. To solve the non-transitivity problem, a number of clustering or graph partitioning algorithms have been proposed, e.g. in Ng and Cardie (2002). Another weakness is the fact that the classifier only knows about the pair and not the clusters that are already formed. Hence, the entity-mention model was introduced (Luo et al., 2004), in which the NP to be resolved is compared against already formed clusters or entities. Still, each pair is considered individually, and so there is no comparison of antecedent candidates, in the sense of which antecedent candidate is the most probable antecedent. To acknowledge this, mention-ranking models have been developed, where all antecedent candidates are considered simultaneously and get ranked for an anaphor to find the most likely antecedent, e.g. in Denis and Baldrige

3. Related NLP tasks

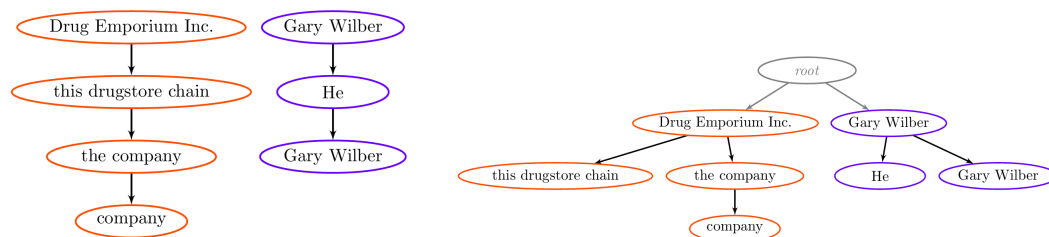


Figure 3.1.: Latent trees for coreference resolution: data structures in the pair-based approach (left) and the tree-based approach (right)

(2007). Ideally, we would like to combine the ranking models with the models that have information about the already formed clusters. Hence, the cluster-ranking approach was proposed by Rahman and Ng (2009).

Since then, numerous algorithms have been developed, enabled by the creation of many annotated corpora, in recent years particularly the OntoNotes corpus (Weischedel et al., 2011). There was also progress in the enhancement of the evaluation methodology, which resulted with the CoNLL score as the de facto official reference scorer (Pradhan et al., 2014), which will be explained in more detail in the following section. The machine learning involved in the models has progressed and more sophisticated structures have been learned, e.g. latent trees in Fernandes et al. (2012) and Björkelund and Kuhn (2014). These approaches assume a hidden structure underlying the data in the form of latent trees. The benefits of such a tree-based system are that you the system has access to more informative antecedent candidates for learning and that you have the ability to define features over the tree, e.g. involving siblings or parents, while in the standard approach you are limited to pairs. Consider Example (2)², for which Figure 3.1 shows the standard data structure and the tree-structure in latent tree approaches.

- (2) {Drug Emporium Inc.}₁ said {Gary Wilber}₂ was named CEO of {this drugstore chain}₁. {He}₂ succeeds {Philip T. Wilber, who founded {the company}₁ and remains chairman}₃. {Robert E. Lyons III, who headed {the company s}₁ Philadelphia region}₄, was appointed president and chief operating officer, succeeding {Gary Wilber}₂.

Of course, there are also other approaches which are not based on supervised methods. Haghghi and Klein (2007) presented an unsupervised approach based on a nonparametric Bayesian model. Rule-based models have also been developed, e.g. in the sieve-based

²Example taken from Björkelund and Kuhn (2014)

Stanford deterministic system (Raghunathan et al., 2010) or in CorZu, a rule-based system for German (Klenner and Tuggener, 2011).

More recently, with the advances in deep learning, neural models have also been applied to coreference resolution (Clark and Manning, 2016b). The neural-net, cluster-ranking approach works with relatively few features, mostly word embeddings and a number of additional features, including string match and distance features. Lee et al. (2017) were the first to introduce a neural end-to-end coreference resolver which is not based on mention extraction that relies on a syntactical parser, but which considers all spans in a document as potential mentions. It includes a head-finding attention mechanism. Based on this, Zhang et al. (2018) suggested a biaffine attention model to receive antecedent scores for each possible mention and jointly optimised mention detection and mention clustering. Due to the power of the word and character embeddings, the only features that are used are speaker information, document genre, span distance and span width features as 20-dimensional learned embeddings.

Evaluating coreference resolution Scoring the performance of a coreference system is an important and non-trivial task, which is why there is not one standard measure, but rather a number of evaluation measures. This section gives an overview of the standard measures as well as the CoNLL metric, which is currently the standard metric for experiments on coreference resolution.

There are a number of relevant terms in coreference evaluation. Singletons are referring expressions that could potentially corefer but occur only once in a document, in contrast to expressions which are never used to refer to anything, e.g. expletive pronouns or idioms like *on the other hand*. There is an ongoing debate on whether the determination of singletons should be a part of the evaluation, as including them affects the evaluation metrics. The term key refers to the manually annotated coreference chains (the gold standard) while response refers to the coreference chains output by a system (Vilain et al., 1995). Recall that coreferring mentions form clusters or equivalence classes. The CoNLL score is an average of the three evaluation metrics MUC, BCUBE and CEAFE. Pradhan et al. (2014) have developed an official scorer, which has also been used in previous shared tasks.³

The MUC algorithm (Vilain et al., 1995) is a link-based version of precision and recall. The recall error is computed by taking each equivalence class, counting the number of

³<https://github.com/conll/reference-coreference-scorers>

Note that earlier versions of the script contained a number of bugs, which heavily affected the performance scores.

3. Related NLP tasks

links that are missing and dividing it by the number of correct links. Reversing the roles of the key and the response leads to the precision error. MUC has a few disadvantages. As the algorithm is link-based, singletons are ignored during the computation. It also sometimes fails to distinguish system outputs of different quality and systematically favours systems that produce fewer equivalence classes.

The B³/BCUBE algorithm (Bagga and Baldwin, 1998) tries to overcome the problem of ignored singletons in MUC by looking at the presence of entities relative to other entities in the equivalence class. Thus, it is mention-based rather than being link-based like MUC.

The Constraining Entity-Alignment F-Measure, short CEAF, produces a one-to-one mapping between subsets of key equivalence classes and system output equivalence classes, with the constraint that a system equivalence class is aligned with at most one key equivalence class. In contrast to other metrics, it penalises systems that produce too many or too few equivalence classes. The metric is based on two similarity measures. One is equivalence-class based, called CEAFE, the other mention-based, further called CEAFM. For the rather complex formulae, see Luo (2005).

There is another popular evaluation metric that is not taken into account for the CoNLL score, called BLANC (Recasens and Hovy, 2010b). The motivation for BLANC was to correctly handle singletons as well as reward correct coreference chains according to their length. BLANC is short for bilateral assessment of noun-phrase coreference and is based on applying the Rand index to coreference. It is bilateral in that it takes into consideration both coreference and non-coreference links. It rewards each link to a class dependent on how large the class is, overcoming problems that both MUC and BCUBE have. Singletons are rewarded as correct full links in BCUBE and CEAF.

Very recently, Moosavi and Strube (2016) have proposed another metric called LEA, link-based entity-aware metric. They state that MUC, BCUBE and CEAFE all have their shortcomings, the agreement between the metrics is often low and they argue that using the CoNLL score as an average of three unreliable metrics does not result in a reliable score. Moosavi and Strube (2016) report a detailed analysis of the shortcomings of the previous metrics and an illustrative example of their newly proposed LEA metric.

As the LEA metric became available after our experiments had already been performed, we aim to use the new LEA metric for future experiments but report the CoNLL score for the experiments in this thesis.

3.2. Bridging resolution

Bridging as an anaphoric phenomenon has been described in Section 2.2. The corresponding NLP task of bridging resolution is about linking these anaphoric noun phrases and their antecedents, which do not refer to the same referent but are related in a way that is not explicitly stated. Reasoning is needed in the identification of the textual antecedent (Poesio and Vieira, 1998).

Bridging anaphora recognition and bridging anaphora resolution There is actually not just one NLP task, but several subtasks that revolve around the phenomenon of bridging. Full bridging resolution is the task of determining that a certain NP is a bridging anaphor and finding the correct antecedent that is necessary for the interpretation of the anaphor. As this task is complex and rather difficult, it has been broken down into two subtasks: (i) determining that a certain NP is a bridging anaphor, called bridging anaphora recognition/determination and (ii) finding an antecedent for a given bridging anaphor, called anaphora resolution. Bridging anaphora recognition is often a part of fine-grained information status classification, where bridging is one of the information status categories. Additionally, some bridging approaches also determine the relation between the bridging anaphor and antecedent.

Computational approaches to bridging resolution Bridging recognition as a subtask of fine-grained information status classification has been performed in Rahman and Ng (2012), which was based on the Switchboard corpus (Nissim et al., 2004; Calhoun et al., 2010). Switchboard contains annotated bridging anaphors, but does not contain the respective annotated antecedents. The bridging category is also limited to the pre-defined relation `mediated/part`, `mediated/situation`, `mediated/event` and `mediated/set`. On this dataset and for the four bridging subtypes, Rahman and Ng (2012) reported a rather high F1 score between 63% for the `event` subcategory, 87% for the `set` category, 83% for `part` and 80% for `situation`, using predicted coreference. As these restricted bridging types do not reflect bridging in data where there is no restriction on the relation or type of the bridging, this result has to be taken with a grain of salt. In the same year, Markert et al. (2012) presented a study on fine-grained information status classification on their corpus ISNotes based on collective classification, where they achieved an F1 score of 18.9% for the subcategory `bridging` (recall 12.1%, precision 41.7%). In Hou et al. (2013a), the model in Markert et al. (2012) was extended to better recognise bridging anaphors. For this, more linguistic features aiming to target genericity,

3. Related NLP tasks

discourse structure and lexico-semantic patterns were integrated. This improved the performance of the subcategory significantly, with an F1 score of 42.2%.

Hou (2016a) implemented an LSTM-based model for fine-grained information status prediction, also based on the corpus ISNotes, where she showed that the model based on word embeddings and a couple of simple additional features achieves comparable results to Markert et al. (2012) in terms of overall information status prediction. However, the performance for bridging was lower than in Hou et al. (2013b), in the best setting of the network the F1 score for the subcategory bridging was only 24.1%.

For German, there has been little work so far. Cahill and Riestler (2012) presented a CRF-based automatic classification of information status, which included bridging as a subclass. However, they did not state the accuracy per class, which is why we cannot derive any performance estimation for the task of bridging anaphor detection. They stated that bridging cases “are difficult to capture by automatic techniques”, which confirms similar intuitions about information status classification for English, where bridging is typically a category with rather low accuracy (Markert et al., 2012; Rahman and Ng, 2012; Hou, 2016a).

The other subtask, bridging anaphora resolution, i.e. determining an antecedent for a given bridging anaphor, has so far been the main focus of most previous work. The first work was presented by Poesio and Vieira (1998), where they investigated the use of definite descriptions. Note that cases of coreference, where the anaphor and antecedent do not share the same head, were included in the bridging category. In this study, they used a corpus of 20 Wall Street Journal articles.⁴ Based on this corpus, a number of papers revolved around resolving these bridging anaphors, mostly based on WordNet (Fellbaum, 1998), e.g. Poesio et al. (1997); Vieira and Teufel (1997); Schulte im Walde et al. (1998); Poesio et al. (2002); Markert et al. (2003). Vieira and Teufel (1997) aimed at resolving definite bridging anaphors. They tested how WordNet can help find the correct antecedent by looking for synonyms (mainly for the coreferent cases), meronyms or hyponyms. They found that only 19% of the bridging cases could be handled by WordNet.

In Schulte im Walde et al. (1998), automatic ways of deducing semantic information, which is necessary to interpret the relation between anaphor and antecedent, were explored by using cluster information. Their category *inferential* again comprised different-head coreference and bridging cases. Their main idea was to find the best

⁴The paper and the corpus were actually already published in 1997, in a manuscript from the University of Edinburgh. The paper in Computational Linguistics appeared in 1998.

antecedent by creating a high-dimensional vector space created using the BNC corpus (Clear, 1993) and computing a similarity measure, including cosine similarity, Euclidean distance and the Manhattan metric. They found that using the cosine similarity worked best, with an accuracy of 22.7%.

Poesio et al. (2002) included syntactic patterns to find meronyms in the BNC corpus, including the “NP of NP” pattern designed to find semantically connected concepts, like *the windows in the room*, or the genitive pattern “NP’s NP (*the room’s windows*). The pattern is useful to find meronym-holonym pairs, as these often occur in the above-mentioned pattern, i.e. “meronym of holonym”. For example, if we consider a bridging anaphor *the windows* and we find an occurrence of *the room* in the previous context, it is likely that *the room* is the antecedent, as they often occur as *the windows in the room*.

Markert et al. (2003) restricted their study to the 12 bridging cases classified as meronymy in Poesio and Vieira (1998) . They then used the “NP of NP” pattern (and many variations thereof) to query the web. The pair with the highest frequency was chosen as the bridging pair. They found the correct antecedent in seven out of 12 cases.

In Poesio et al. (2004), the first machine-learning based model for bridging anaphora resolution was presented: a pair-wise model fed with lexical and salience features, focusing on cases of meronymy in the GNOME corpus.

Lassalle and Denis (2011) adapted the learning-based approach to French and reported an accuracy of 23% for meronymic bridging pairs in the DEDE corpus.

Most of the work on anaphora resolution presented here is restricted to definite descriptions and included a mixture of coreference and bridging cases. However, a lot of the ideas proposed in these approaches are still very relevant, e.g. the idea of the prepositional pattern introduced in Poesio et al. (2002).

More recently, Hou et al. (2013b) presented a study on anaphora resolution that was not restricted to definites or certain semantic relations, based on the ISNotes corpus (Markert et al., 2012). They started with a pair-wise model and a rich feature set, but then stated that considering anaphor and antecedent pairs in isolation does not seem to be reasonable, as they often appear in clusters. This means that one expression is introduced, e.g. *the house* and then several aspects of it are discussed, e.g. *the floors*, *the garage*, etc. As such, antecedents are often the antecedent of several anaphors, so-called sibling anaphors. To acknowledge this, they switched to a global Markov model, in which they used the same features as in the first experiment, but added the following constraints: (i) anaphors are likely to share the same antecedent, (ii) the semantic connectivity of one antecedent to all anaphors should be modelled globally and (iii) the

3. Related NLP tasks

union of potential antecedents is considered for all anaphora instead of a fixed window size. This way, they could achieve a significant improvement over the baseline, with an accuracy of 41.32% on the ISNotes corpus.

Hou (2018) presented an experiment on bridging anaphora resolution where she created word embeddings based on extracting matches using the NP of NP pattern in the Gigaword corpus (Napoles et al., 2012) to capture the semantic connectivity between two words. She showed that using these word embeddings alone, one can achieve 30% accuracy. When integrating the PP word embeddings into the global model in Hou et al. (2013b), the state of the art on the ISNotes corpus could be improved, and reached 45.85% accuracy.

In this thesis, we will mainly consider the task of full bridging resolution, i.e. a combination of bridging anaphor detection and resolution, but will also report numbers for bridging anaphor detection.

The first work on full bridging resolution was performed in Vieira and Poesio (2000), where they classified each definite description as either direct anaphora (same head coreference), discourse-new, or a bridging description. For those definite descriptions classified as bridging, the system then identifies an antecedent. The system made use of syntactic and lexical information to classify the definite descriptions and used WordNet to resolve bridging descriptions. They did not state the performance for the category bridging, which again included cases of coreference where the anaphor and antecedent do not share the same head.

Bunescu (2003) also presented a system for full bridging resolution for definite descriptions, using lexico-syntactic patterns by searching the web. He distinguished identity and associative (=bridging) anaphors. The method was evaluated on the first 32 documents of the Brown section of the Treebank corpus, but performances for the individual classes (identity or associative) were not reported.

Hou et al. (2014) presented a rule-based system that consists of eight hand-crafted rules. Some of the rules are rather specific, for example aiming to find buildings and their parts, while other rules make use of the “NP of NP” pattern to determine the semantic connectivity of two words. The system will serve as a baseline in one of our experiments and will be explained in more detail in Section 6.

In our own previous work (Rösiger and Teufel, 2014), we aimed at resolving bridging anaphors in scientific text by training a coreference resolver on bridging references, together with some additional WordNet features.

Sasano and Kurohashi (2009) presented a probabilistic model to resolve bridging anaphors in Japanese. Their model considers bridging anaphora as a kind of zero anaphora and applies techniques used to resolve zero anaphora, based on automatically acquired lexical knowledge.

For full bridging resolution in German, Hahn et al. (1996) and Markert et al. (1996) have presented a resolver for bridging anaphors, back then called textual ellipsis or functional anaphora, in which they resolved bridging anaphors in German technical texts using centering theory (Grosz et al., 1995) and a knowledge base. The corpus and the knowledge base, as well as the overall system, are, however, not available.

Evaluating bridging resolution We adopt the evaluation metrics applied in previous research (Hou, 2016b), where the evaluation of bridging resolution is computed using the widely known measures precision and recall (and the harmonic mean between them, F1). The precision of a rule or a system is computed by dividing the correctly predicted bridging pairs by the number of all predicted bridging pairs. The recall is computed by dividing the correctly predicted bridging pairs by the number of all gold bridging pairs. The bridging anaphor is considered a mention, while the antecedent is considered an entity, which is taken into account by including gold coreference chains during the evaluation. If the predicted antecedent is one of the mentions in the coreference chain of the gold antecedent, the bridging pair is considered correct.

This rather basic evaluation has a few shortcomings: Firstly, the evaluation is rather strict, as overlapping markables (where, for example, the predicted anaphor contains an adverb which the gold anaphor does not contain) are considered wrong. This is particularly relevant for experiments with automatically predicted markables, as they might sometimes differ from the gold markables which are annotated.

Furthermore, bridging anaphors with more than one link, e.g. comparative anaphora in the sense of Example (3), are only correct if all antecedents have been found by the system. Partial correctness is not taken into account, i.e. when the pair *the US* and *other countries* is suggested by the system in Example (3), it is considered wrong.

(3) Canada, the US and **other countries**

The same holds for discontinuous antecedents, as in Example (4), where the anaphor *those in Europe wanting to invest in IT technology* was annotated and a part of the NP, *or Asia*, was left out. It is probably controversial whether allowing parts of NPs to be

3. *Related NLP tasks*

markables is a good annotation strategy, but as this is present in some of the corpora, it would be nice to have some way of dealing with it in the evaluation.

(4) those in Europe or Asia wanting to invest in IT technology.

Another special case are anaphors without antecedents, so-called empty antecedents, which are also contained in some of the bridging corpora.

We adopt the rather simple evaluation metrics precision and recall in order to be comparable with previous research, but it should be noted that a more complex evaluation metric, as available for coreference resolution, would be preferable to ensure a fairer evaluation.

Part II.

Data and tool creation

4. Annotation and data creation

Research Question 2: Data creation

Is there enough consistently annotated data to enable the creation of automatic tools, including ones making use of statistical algorithms? If not, can we create data resources to fill the research gap?

This section gives an overview on previous work on coreference and bridging annotation, particularly on the compatibility of annotation guidelines, and summarises available corpus resources, with a focus on the corpora used in the remainder of the thesis. Thus, it gives an answer to Research Question 1 on how well defined the phenomenon and the tasks are, which was already partially answered in Section 2.1 and 2.2, as well as to Research Question 2 on the availability of data.

4.1. Coreference annotation and existing corpora

As explained in Section 2.1, the phenomenon of coreference is generally well-understood and clearly defined, with the exception of a few special cases. These differences can, of course, be of importance when the aim of the work is to study one of these special phenomena. To give an impression of what differences remain, Table 4.1 compares three exemplary coreference guidelines, the OntoNotes guidelines (Weischedel et al., 2011), the RefLex guidelines (Baumann and Riester, 2012; Riester and Baumann, 2017) as well as the NoSta-D guidelines developed for non-standard text (Dipper et al., 2013).

The first important question is always how the markables, the expressions that we want to annotate, are defined. Most work suggests annotating the maximum span of NPs as well as embedded NPs as additional markables. As can be seen in Table 4.1, RefLex includes prepositions in the markables, which means that prepositional phrases (PPs) are annotated rather than NPs, in cases where an NP is embedded in a PP. This is due to the fact that the guideline schema was developed for German, where there are merged forms of a preposition and a determiner, for example in *am Bahnhof* (*at the station*). Another common difference is that other types of pronouns are included or

4. Annotation and data creation

excluded, as well as the handling of certain pronouns. Relative pronouns, for example, are sometimes annotated as a markable, whereas in the RefLex scheme, they are part of the relative clause and not annotated as a separate markable because they trivially corefer with the referent of the head noun (or the whole span, respectively). Other types of difference stem from special constructions such as aggregation, which is for example not annotated in OntoNotes, or the handling of generic entities. In OntoNotes, generic pronouns can be linked to their antecedent, but they always only make up a coreference chain of two, consisting of the anaphor-antecedent pair. Additionally, some guidelines (e.g. NoStaD) distinguish non-anaphoric and anaphoric coreference¹, whereas most guidelines do not make such a distinction.

Since we are interested in general coreference resolution and not in a certain special case, e.g. generic anaphors or abstract anaphors, we accept the minor differences contained in most corpora.

| | OntoNotes | RefLex | NoSta-D |
|-------------------------------------|--------------------------|-----------------------------------|------------------------|
| Prepositions | excluded from markable | included in markable | excluded from markable |
| Relative pronouns | annotated separately | part of complex relative markable | annotated separately |
| Antecedent of abstract anaphor | verbal head | entire clause or VP | not annotated |
| Aggregation | no | yes | yes |
| Apposition | separate link | included in markable | included in markable |
| Generic expressions | annotated | annotated | not annotated |
| Generic anaphors | only pronouns are linked | linked | not linked |
| Non-anaphoric/anaphoric coreference | not distinguished | not distinguished | distinguished |

Table 4.1.: Guideline comparison: overview of the main differences between OntoNotes, RefLex and NoSta-D

As a result of the well-understood phenomenon, many high-quality corpus resources have been developed. Nowadays, automatic tools are typically trained on the benchmark dataset OntoNotes, which spans multiple genres (mostly newswire, broadcast news, broadcast conversation, web text, among others) across three languages – English, Chinese and Arabic (Weischedel et al., 2011). Before OntoNotes, the (much smaller) benchmark

¹As explained in the introduction, non-anaphoric coreference occurs for example when certain named entities, such as *Google*, occur several times throughout a document.

datasets used were the MUC (Hirschman and Chinchor, 1998) and ACE (Doddington et al., 2004) corpora. OntoNotes differs from these two corpora with respect to corpus size and the inclusion of a few more genres. Benchmark datasets have of course also been created for other languages, e.g. the Prague Dependency Treebank (Hajič et al., 2018) for Czech, the ANCORA newspaper corpora of Spanish and Catalan (Martí et al., 2007) or TüBa-D/Z (Naumann and Möller, 2006) as a newspaper corpus for German, to name only a few.

Despite the fact that OntoNotes contains multiple genres, it is unsuited as a data basis for other domains with very different properties. Hence, annotated corpora have also been created for many other domains. One example for such a domain is the biomedical domain, for which Gasperin and Briscoe (2008) have shown that the text differs considerably from other text genres such as news or dialogue and that the complex nature of the texts is for example reflected in the heavy use of abstract entities, such as results or variables. As a result, many corpora have been annotated (Castaño et al. (2002), Cohen et al. (2010), Gasperin et al. (2007), Batista-Navarro and Ananiadou (2011), among others) for this domain. It has been shown that coreference resolution for the biomedical domain benefits a lot from in-domain training data (Rösiger and Teufel, 2014). Another example for a domain where corpora have been developed is scientific text, e.g. in Schäfer et al. (2012), where a large corpus of computational linguistics papers has been annotated.

Due to the large amount of published coreference corpora, we refrain from including a full literature review. For a more detailed analysis of the most important available corpus resources, see Poesio et al. (2016).

Corpora used in this thesis This section presents the three corpora containing coreference annotation that we will use in our experiments. Table 4.2 shows in which sections the existing corpora are used.

| Corpus | Language | Annotations | Used in | Section |
|-----------|----------|-----------------------|--|----------------------------|
| OntoNotes | EN | Coreference | Validation experiments | Section 8 |
| TüBa-D/Z | DE | Coreference | Tool development | Section 5.2 |
| DIRNDL | DE | Coreference, bridging | Tool development Validation experiments | Section 5.2.6 Section 7 |

Table 4.2.: Existing corpora annotated with coreference used in this thesis

4. Annotation and data creation

OntoNotes The OntoNotes corpus (Weischedel et al., 2011, 2013) has been the benchmark dataset for English, Arabic and Chinese since the shared tasks on coreference resolution in 2011 and 2012 (Pradhan et al., 2011, 2012). We will use the English portion, which contains 1.6M words, in our linguistic validation experiments where we include automatically predicted semantic relation information to improve coreference resolution.

TüBa-D/Z The reference corpus for coreference resolution experiments on German data is TüBa-D/Z² (Naumann and Möller, 2006). The TüBa-D/Z treebank is a syntactically and referentially annotated German newspaper corpus of 1.8M tokens based on data taken from the daily issues of ‘die tageszeitung’ (taz). We will use the TüBa-D/Z data in the tool development section, where we adapt an existing coreference tool to German. The NoSta-D guidelines, as shown in Table 4.1, are based on the TüBa-D/Z guidelines³.

DIRNDL The DIRNDL corpus (Eckart et al., 2012; Björkelund et al., 2014) is a German corpus of spoken radio news. Coreference anaphors have been annotated as a subcategory of referential information status according to the RefLex scheme (Riester et al., 2010; Baumann and Riester, 2012) (also contained in Table 4.1) and the coreference anaphors have also been linked to their antecedents. DIRNDL contains nominal, verbal and clausal antecedents. We adopt the official training, test and development split. As DIRNDL is a corpus of spoken text, we will also be using it for the validation of theoretical claims on the effect of prosody on coreference.

Conclusion In the area of coreference, much theoretical work on coreference and coherence has built the foundation for a good understanding of the phenomenon. Hence, annotation guidelines typically differ only in minor aspects, such as the handling of genericity or abstract anaphors. Although, of course, an agreed upon handling of all special cases would be desirable, it is not of greatest importance for studies that do not focus on these special cases. Since we are concerned with general coreference, we accept the minor differences that are present in the corpus resources. Huge corpora have been developed for many domains and languages, including the benchmark datasets for English, OntoNotes with 1.6M tokens, and for German, TüBa-D/Z with 1.8M tokens. These enable

²<http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html>

³For a description of the coreference annotation scheme, please refer to the stylebook for anaphoric annotation, which can be found at <http://www.sfs.uni-tuebingen.de/resources/tuebadz-coreference-manual-2007.pdf>

the creation of automatic resolvers, including statistical models that require larger data in order to work properly.

Hence, at least for coreference, we can answer the research questions positively: yes, the task is conceptionally clear, almost all previous work uses compatible annotation guidelines, and as a result, large corpus resources have been created, which have already enabled a lot of progress in the field of coreference resolution. Therefore, there is not much need for the development of new data, and the annotation and creation of new corpora is thus not a focus in our workflow pipeline for coreference. However, for the sake of completeness, all the resources which we will create to overcome the data problem for bridging will also contain coreference annotations. This way, the relation between the two anaphoric phenomena can later be analysed and computational approaches can learn coreference and bridging jointly.

4.2. Bridging annotation and existing corpora

As explained in Section 2.2, the term bridging stands for many different phenomena, and many aspects thereof have been controversial for a long time. It is therefore not a surprise that annotation guidelines and the created corpus resources also vary quite a bit, in terms of pre-defined relations, the definiteness requirement for bridging anaphors, whether the antecedents can be nominal or also verbal or clausal, and whether there is an overlap with coreferent anaphors. We have discussed many of the controversial issues in Section 2.2. This section aims at giving an overview of the corpus resources and their properties. Although our main focus is set on English and German, as there are far fewer corpora for bridging than for coreference we will include other languages as well as information on the inter-annotator-agreement, where available.

Poesio/Vieira corpus The first real dataset that was introduced was the one in Poesio and Vieira (1998), which consists of 33 Wall Street Journal articles annotated according to their classification scheme of definite descriptions.

Anaphors: definite NPs.

Relations: **identity** (overlap with coreference), **compound noun** and **meronymy**.

Antecedent: entity (nominal) or event (verbal, clausal).

GNOME The overlap with coreference is not present in the GNOME corpus (Poesio, 2004), which comprises about 500 English sentences in museum object descriptions and

4. Annotation and data creation

drug leaflets.

Anaphors: all NPs.

Relations: `set membership`, `subset` and `generalised possession`, including `meronymy` and `ownership` relations.

Antecedent: entity (nominal).

PAROLE The bridging subcorpus of the PAROLE corpus (Gardent et al., 2003) is a 65k words corpus of French newspaper texts.

Anaphors: definite NPs.

Relations: `set membership`, `thematic`, `definitional` (including `meronymy`, `attribute`, `associate`), `co-participants` and `non-lexical circumstantial`.

Antecedent: strictly nominal or verbal, not clausal.

DEDE corpus Gardent and Manuélian (2005) presented a French newspaper corpus of roughly 5000 definite descriptions, with bridging as one of the categories in their classification scheme.

Anaphors: definite NPs.

Relations: `meronymy`, `modifier-modified` relation and `predicate-argument`.

Antecedent: entity (nominal).

Caselli/Prodanoff Caselli and Prodanoff (2006) presented a corpus study of definite descriptions in Italian news text (17 articles, 10k words). They presented high inter-annotator-agreement for bridging anaphora recognition (κ 0.58-0.71) and antecedent selection (κ 0.78).⁴

Anaphors: definite NPs.

Relations: not restricted.

Antecedent: entity (nominal).

Switchboard The Switchboard corpus (Nissim et al., 2004) comprises information status annotations, which refer to the degree of givenness of an NP. Bridging was contained in the category `mediated`, namely in the subcategories `part-of`, `set`, `situation` or `event`. Other subcategories of `mediated` do not contain cases of bridging. The information status scheme was annotated in a subpart of the Switchboard corpus. The

⁴ κ is a statistical measure for assessing the reliability of agreement between a fixed number of annotators. It measures the degree of agreement over what would be expected by chance by taking in the distribution of the categories (Fleiss, 1971).

annotation consisted only of labelling NPs with their information status and did not include linking bridging anaphors to their antecedents. Hence, the corpus only contains bridging anaphors, and no bridging pairs. There were some additional constraints for the bridging annotation, for example, the restriction that anaphors of the **part-whole** type could only be annotated if they appeared in WordNet (Fellbaum, 1998), or the restriction to FrameNet (Baker et al., 1998) frames for the type **mediated/situation**.

Anaphors: all NPs.

Relations: **part-of** (WordNet), **set**, **situation** (FrameNet) and **event**.

Antecedent: not annotated.

CESS-ECE Recasens et al. (2007) presented guidelines to add different coreference subtype annotations to the Spanish CESS-ECE corpus, with bridging as one subtype of coreference. How much of this corpus was actually annotated remains unclear.

Anaphors: all NPs.

Relations: bridging as a subtype of coreference, not further restricted.

Antecedent: nominal or verbal, not clausal.

SemDok In a subset of the German corpus SemDok (Bärenfänger et al., 2008), definite descriptions were annotated in three scientific articles and one newspaper text. However, the exact number of bridging anaphors in this corpus is unknown and the corpus is currently not available.

Anaphors: all NPs.

Relations: **possession**, **meronymy**, **holonym**, **hasMember**, **setMember** and **undefined**.

Antecedent: entity (nominal), event (verbal, clausal).

ARRAU The ARRAU corpus, first released in Poesio and Artstein (2008), contains English texts from three domains: newspaper, spoken narratives and dialogue. In the newest version (Uryupina et al., 2018), the corpus contains 5,512 bridging pairs. Most annotated bridging pairs are of the category **subset** or **element-of**.

Anaphors: all NPs.

Relations: **set membership**, **subset**, **possession**, **other** and **unrestricted**.

Antecedent: entity (nominal), event (verbal, clausal).

COREA corpus Hendrickx et al. (2008) presented guidelines and a corpus for Dutch, which mainly focused on coreference, but also included bridging as a subtype. Bridging was restricted to **superset-subset** or **group-member** relations. Bridging turned out to

4. Annotation and data creation

be the subtype with the lowest inter-annotator-agreement (33% MUC F1 score).

Anaphors: all NPs.

Relations: bridging as a subcategory of coreference, with the annotated relations `group-member` and `subset`.

Antecedent: entity (nominal).

Prague dependency treebank (PDT) Bridging has been annotated in a subset of the Czech PDT corpus (annotation guidelines described in Nedoluzhko et al. (2009), corpus last released in Hajič et al. (2018)). They state that they did not perform an unrestricted annotation of bridging because they feared it would be too inconsistent, as Czech lacks a definite article. Therefore, they specified a couple of relations to be annotated, including `meronymy`, `subset`, `function` and `contrast`, among others.

Anaphors: all NPs.

Relations: `meronymy`, `subset`, `function`, `contrast`, `explicit anaphoricity` (demonstrative article without coreference), `rest` (with some additional, quite specific subcategories: `relatives`, `event-argument` and a few others).

Antecedent: entity (nominal), event (verbal, clausal).

Italian Live Memories Corpus The Italian Live Memories Corpus (Rodríguez et al., 2010) is an Italian corpus of annotated Wikipedia articles and blog texts. It is relatively large, with 142k tokens of Wikipedia articles and 50k tokens of blog texts, but restricts bridging to only three pre-defined relations.

Anaphors: all NPs.

Relations: `part-of`, `set-member`, and `attribute`.

Antecedent: entity (nominal).

Copenhagen Dependency Treebank A subpart of the Copenhagen Dependency Treebank (Korzen and Buch-Kromann, 2011), a multi-language corpus, has also been annotated with anaphoric information, including bridging. The exact size of the subcorpus is not known. Anaphors: all NPs.

Relations: 16 quite detailed relations under the two categories `semantic role` and `lexical semantics and generativity`.

Antecedent: entity (nominal).

DIRNDL The DIRNDL corpus (Eckart et al., 2012; Björkelund et al., 2014), as mentioned above in the coreference section, is a German corpus of spoken radio news.

Bridging has been annotated as a subcategory of referential information status (Riester et al., 2010; Baumann and Riester, 2012). In this scheme, indefinite expressions introduce new information and are thus excluded from the bridging category. As a result, all bridging anaphors in DIRNDL are definite.

Anaphors: definite NPs.

Relations: unrestricted.

Antecedent: entity (nominal) and event (verbal, clausal).

ISNotes The ISNotes corpus (Markert et al., 2012), a corpus of newspaper text (50 Wall Street Journal articles). It contains bridging as a subclass of information status annotation, with 633 annotated bridging pairs. It contains definite and indefinite bridging anaphors, but no comparative anaphors, as these cases were considered a different information status category. For the bridging category, the kappa values are over 0.6 for all three possible annotator pairings.

Anaphors: all NPs.

Relations: not restricted, with the exception of comparative anaphora, which are not included in the bridging category.

Antecedent: entity (nominal).

Coref pro corpus Grishina (2016) recently described a parallel corpus of German, English and Russian texts with 432 German bridging pairs that have been transferred to their English and Russian counterparts, resulting in 188 transferred English bridging pairs. The corpus contains narrative and news text as well as medicine instruction leaflets. In contrast to the other corpora, she applies a three-way classification: anaphors can be **coreferent**, **bridging** or of the category **near-identity**. In terms of the bridging definition, they base their work on the assumption that the speaker intends the listener to be able to compute the shortest possible bridge from the previous knowledge to the antecedent which is therefore unique (determinate) in the discourse. Hence they only annotate definite descriptions as bridging anaphors. On a subset of the German part of the corpus, the paper reports rather high inter-annotator agreement for bridging anaphora recognition (F1 score of 64%) and antecedent selection (F1 score of 79%).

Anaphors: definite NPs.

Relations: **meronymy**, **set-membership**, **entity-attribute/function** (*Kosovo-the government*), **event-attribute** (*the attack- the security officers*), **location-attribute**

4. Annotation and data creation

(*Germany- in the south*), and **other** to capture other types of bridging.

Antecedent: entity (nominal).

RuGenBridge Roitberg and Nedoluzhko (2016) presented a Russian corpus annotated with genitive bridging. They define genitive bridging as “the case where two elements (an anchor/antecedent and a bridging element/anaphor) can form a genitive construction, where the anchor is marked with the genitive case in Russian”. In other words, they only mark bridging cases which can be paraphrased as a genitive construction, i.e. *the room* and *the ceiling* could be a bridging pair as it is possible to utter *the ceiling of the room*. They argue that this limitation helps overcome the vagueness of many previous annotation efforts, which is often reflected in the low inter-annotator-agreement. As the paper mainly presented the annotation scheme, the annotation and the corpus development is still underway.

Anaphors: all NPs.

Relations: only genitive bridging cases.

Antecedent: entity (nominal).

GUM The GUM corpus (Zeldes, 2017), an English multi-domain corpus of (currently) 85,350 tokens annotated with bridging links and coarse-grained information status, has recently been released. As the corpus is expanded by students as part of a curriculum at Georgetown University, it continues to grow.

Anaphors: all NPs.

Relations: not restricted to certain relations.

Antecedent: entity (nominal) and event (verbal, clausal).

| Corpus | Language | Annotations | Used in | Section |
|---------|----------|-----------------------|--------------------------------|--------------------------|
| ISNotes | EN | Bridging | Tool development Validation | Section 6.1 Section 8 |
| ARRAU | EN | Coreference, bridging | Tool development | Section 6.2 |
| GUM | EN | Coreference, bridging | Tool development | Section 6.1 |
| DIRNDL | DE | Coreference, bridging | Tool development Validation | Section 6.5 Section 7 |

Table 4.3.: Existing corpora annotated with bridging used in this work

Corpora used in this thesis We make use of a few of the corpora presented in this section. To develop a freely available bridging tool, we use the corpus ISNotes, as it con-

tains reliable and unrestricted bridging annotations. To check how well the approaches presented in Hou (2016b) generalise to other in-domain corpora, we also use the ARRAU corpus. This assessment of generalisability will also include some GUM annotations. For German, the DIRNDL corpus was the only available corpus containing bridging annotations at the time of this research. We will thus use the corpus to develop a bridging resolver for German. Some of the corpora will also be used in the validation experiments. Table 4.3 shows in which sections existing corpora are used. The other corpora that were presented above are not included as they either contain data in another language than our two languages of interest, English and German, have major restrictions such as an overlap with coreference in their bridging definition or are not openly available.

Conclusion The phenomenon of bridging has been studied in many theoretical and computational works, as highlighted in the previous sections. Different phenomena have been described as bridging, and, as a result, the corpora have very different properties. One of the most apparent differences is the **limitation to a certain set of pre-defined relations** (e.g. Poesio and Vieira (1998); Poesio (2004); Nedoluzhko et al. (2009), among many others). The reason for the limitation is often argued to be the improved annotation quality, e.g. in Poesio (2004), as the annotation of bridging without any relation restrictions tends to result in low inter-annotator-agreement. Reducing bridging to e.g. only cases of meronymy makes the task clearer for human annotators, but does in our opinion not reflect the complexity inherent in bridging relations. We see bridging as a versatile phenomenon on the pragmatic level, where anaphoricity is signaled by the speaker or writer. Simplifying the task to finding anaphoric cases of pre-defined relations can, of course, be a subtask, which however leaves the difficult cases of bridging, where the relation cannot be described with relations such as meronymy, subset-member or attribute-of, unresolved. With the improved understanding of the phenomenon of coreference, the **overlap between coreference and bridging**, i.e. considering non-identical head coreference as bridging, seems to be a thing of the past, although the terminological confusion remains, e.g. in Feuerbach et al. (2015), where the term “bridging” mentioned in the title of the work actually refers to non-identical-head coreferent mentions. Other limitations, like the **definiteness requirement for bridging anaphors**, is still very present in current work on bridging, for example in Grishina (2016). **The restriction to NP antecedents** is also common in previous work (Poesio, 2004; Gardent and Manuélian, 2005; Grishina, 2016). This excludes a smallish percentage of bridging anaphors with a verbal or clausal antecedent. In the

4. Annotation and data creation

corpus ISNotes, where they are included, they make up about 10% of all bridging cases. One could argue that computational work typically focuses on NP antecedents, even in coreference resolution, where there is much more progress, and that event reference thus is a special case which is not of great overall importance for the current state of bridging resolution. On the other hand, it is, of course, a part of the complex phenomenon that is bridging, and cannot be studied when there is not enough data that includes this in the annotation. In addition to the different interpretations of the task, the size of annotated data resources is the biggest issue for researchers aiming to apply statistical algorithms to the data. As a comparison, OntoNotes, the benchmark dataset for coreference resolution, contains 35,000 coreference pairs, taking into account the transitivity of the coreferent pairs. ISNotes, the corpus on which most recent work has been reported (Hou et al., 2014; Hou, 2016b, 2018), comprises only 633 bridging pairs. Of course, coreference anaphors are also more frequent than bridging anaphors and the relation is transitive, i.e. we can pair every mention in a certain chain with another member of the chain to create more data for learning, but still, the difference in corpus size is major, to say the least.

As a consequence, and to answer the last part of Research Question 2, **much work is needed on the creation of reliable and unrestricted bridging data**. We think that even small, reliably annotated resources can help check how generalisable previous approaches are, and can help make the approaches less tuned to the very small available datasets. Hence, during the course of the last four years, we have developed three resources for our main languages of interest, English and German, which we think will be beneficial for our own work as well as for future work in this area:

BASHI: a corpus of English Wall Street Journal (WSJ) articles, where bridging is defined as unrestricted as possible, in order to be compatible with ISNotes, which also contains WSJ articles. We define a number of subcategories (definite anaphors, indefinite anaphors, comparative anaphors) so that it can also be used with a couple of other corpora which do not include indefinite anaphors, and so that people have the choice to focus on one type of bridging. The corpus is presented in Section 4.3.1. The work on BASHI was published in Rösiger (2018a).

SciCorp: an English corpus of a different domain, scientific text, which contains genetics articles as well as computational linguistics articles. We want to use this corpus to assess how well approaches developed for news text transfer to other domains. As

this corpus was developed quite early in the PhD progress, it is however also limited to definite anaphors. Nowadays, with our growing understanding of the task, we would strongly argue not to make this restriction. However, it is not limited in terms of the annotated relations, and can still be used in computational approaches that focus on the subset of definite anaphors. The corpus is presented in Section 4.3.2. The work concerning SciCorp was published in Rösiger (2016).

GRAIN: a corpus of German radio interviews, annotated for referential information status including coreference and bridging, and a number of other annotation layers including syntax. The corpus contains twenty 10-minute interviews. As the information status was annotated according to the RefLex scheme (Baumann and Riester, 2012; Riester and Baumann, 2017), all anaphors are again definite. As mentioned above, ideally, we would like to include also indefinite bridging. In contrast to BASHI and SciCorp, the corpus was created in a joint effort of many researchers at IMS, and we were involved in the training and guidance of the information status annotators. The corpus is presented in Section 4.3.3. The work on GRAIN has been published in Schweitzer et al. (2018).

4.3. Newly created corpus resources

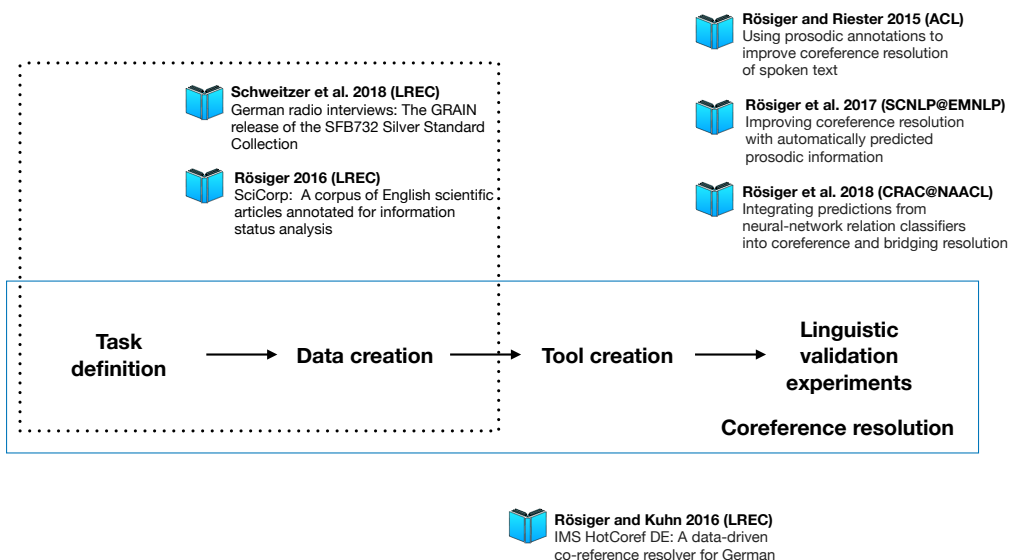


Figure 4.1.: Contribution and workflow pipeline for coreference: data creation

4. Annotation and data creation

For coreference resolution, there are many high-quality corpus resources. Hence, as can be seen in Figure 4.1, most publications revolve around later steps in the pipeline, including tool creation and validation, but we have included coreference in all three newly created corpora, in order to have a complete picture of anaphoric relations. These joint bridging and coreference annotations could also be exploited in future work.

As explained in the last section, not many reliably annotated bridging corpora are available. As a basis for the tool development and validation step in the pipeline, we create three corpus resources annotated with coreference and bridging links: BASHI, a corpus of news text to check whether current methods designed for news text generalise well to other corpora, and SciCorp, a corpus for scientific text, to see how well these methods work on out-of-domain text. We also created a bridging resource for German, GRAIN. Figure 4.2 presents the contributions for bridging in the first two steps. In contrast to coreference, task definition and data creation are very important steps in our work on bridging. This section gives an overview of the newly created corpora, provides details on the respective annotations and guideline decisions and compares the corpora with previously created data. Table 4.4 presents the newly created corpora and the sections in which they are used.

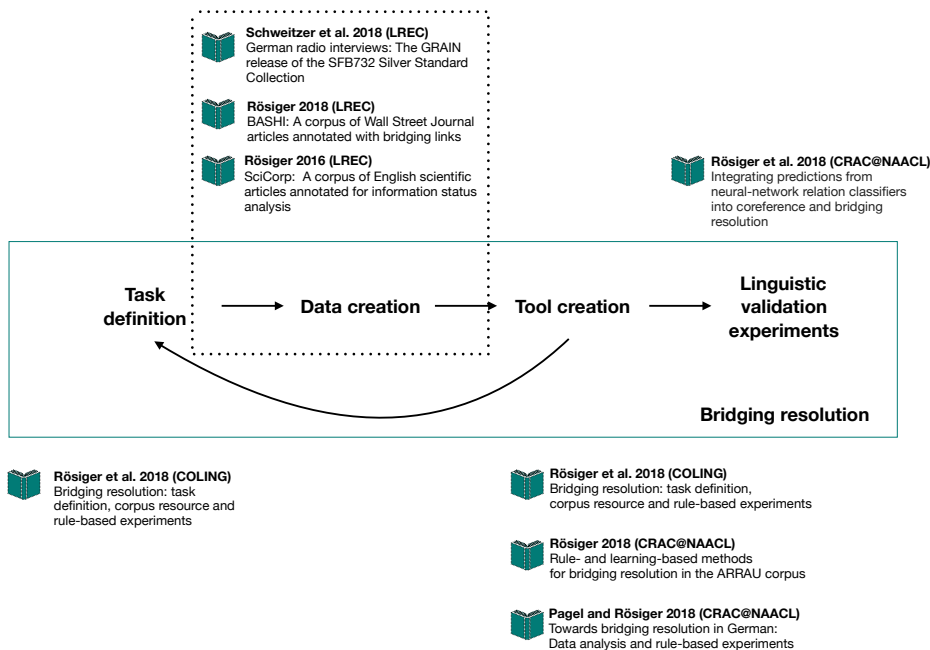


Figure 4.2.: Contribution and workflow pipeline for bridging: data creation

| Corpus | Language | Annotations | Used in |
|---------|----------|-----------------------|--------------------------------|
| BASHI | EN | Bridging | Tool development Section 6.1.3 |
| SciCorp | EN | Coreference, bridging | Tool development Section 6.1.3 |
| GRAIN | DE | Coreference, bridging | Tool development Section 6.5 |

Table 4.4.: An overview of the newly created data

4.3.1. BASHI: bridging in news text

This section presents the annotation guidelines and annotation process of the BASHI corpus, as well as the resulting resource.

We annotated 50 articles from the WSJ that are already part of OntoNotes, meaning they already come with coreference annotations. The articles were selected blindly, but we excluded articles that were already annotated as part of the ISNotes corpus (Markert et al., 2012) and those articles that give an overview of what happened in a certain time frame, thus containing several separate discourses in one document. The corpus is named BASHI, bridging anaphors hand-annotated inventory⁵. It is a relatively small corpus, but because of its categorised bridging links it can be combined with many other corpus resources (e.g. ISNotes), in order to create a larger corpus resource. Our annotation guidelines are on the one hand broad enough to cover many cases, following these principles:

- Bridging anaphors have to be truly anaphoric, i.e. not interpretable without an antecedent;
- Bridging relations are not restricted to certain pre-defined relations;
- Bridging anaphora can be definite or indefinite, but we use two different labels to distinguish them;
- Bridging antecedents can be nominal entities or events (VPs or clauses).

On the other hand, we propose a clear separation from other tasks:

- No overlap with coreference resolution:
context-dependent anaphors that refer to the same entity as their antecedent are considered given information (independent of their surface realisation), and are thus covered by coreference resolution;

⁵Bashi can mean “bridge” in Japanese. The corpus was presented at LREC 2018 in Miyazaki, Japan.

4. Annotation and data creation

- Hence, bridging anaphors are context-dependent expressions that do not refer to the same entity as their antecedent, but to a related entity;
- We focus on referring expressions, excluding rhetorical or connection cases (Asher and Lascarides, 1998): anaphors are nominal; antecedents can be nominal, verbal or clausal.

The annotation guidelines are tailored to Germanic languages like English and German as they focus on the distinction between definiteness and indefiniteness. The idea of a broad, but clear definition of bridging without an overlap with the concept of coreference can of course also be applied to other languages.

Annotation scheme

Markables Markables (and thus candidates for bridging anaphors) are all NPs that have been gold annotated in the OntoNotes corpus (Weischedel et al., 2011). Pre-marked NPs in OntoNotes include:

- definite and demonstrative nominal phrases: *the president*,
- proper names: *Mr Bush*,
- quantifier phrases: *all the products*,
- pronouns: personal, possessive, demonstrative, reflexive.

If the annotator thought that an NP has not been pre-marked, he or she added a markable to the set of markables (this was rarely the case).

The annotators were told to mark the longest span of the NP that refers to an entity, including determiners and adjectives, dependent PPs and relative clauses.

- (1) There have been concerns that the Big Board's basket could attract investors with a short term perspective who would rapidly turn over the product, thus increasing volatility.

Non-markables The pre-marked NPs do not include

- nominal premodification: the *US* president,
- interrogative or relative pronouns.

Bridging anaphors

In our annotation, bridging anaphors are discourse-new, anaphoric expressions which are dependent on the previous context, and for which the text presents an antecedent NP which does not stand in the relation of identity, but in some other form of relation to the associative phrase. The antecedent may be an associate in a typical relation such as **part-of**, **part-of-event** or any kind of associate as long as there is a clear relation between the two phrases.

- (2) My sister celebrated her birthday last weekend.
I offered to help her make **the cake**.

Often, the anaphor is lacking an argument (the antecedent) which enables the interpretation of the expression. This is also reflected in the bridging definition of Roitberg and Nedoluzhko (2016), called genitive bridging, where they restrict bridging cases to those that can form a genitive construction with the antecedent. While genitive constructions might be a bit too restrictive and the use of genitive constructions is very language-dependent, we agree that bridging pairs can often be seen as head-argument constructions.

- (3) **the cake** (at her birthday)

Definite Use Most bridging anaphors are definite NPs. Note that bare singulars can sometimes also count as definite, in cases where the insertion of the definite article is more plausible than the insertion of an indefinite article. Bare plurals usually count as indefinites.

- (4) I went into the room. **The windows** were broken.
- (5) We performed the experiments using
Evaluation is done by means of 10-fold cross validation.

Indefinite Use Some bridging anaphors are indefinite expressions. In this case, we label the NP as indefinite and link it to the preferred antecedent. Indefinite cases of bridging are typically either **part-of** or **part-of-event** relations. We annotate them as bridging in cases where we feel that the interpretation strongly benefits from an argument in the form of the antecedent.

- (6) I bought a bicycle. **A tire** was already flat.

4. Annotation and data creation

- (7) Afghanistan ... **Millions of refugees** would rush home.

Comparative anaphors Comparative anaphors were excluded from the bridging category and treated as a separate category in the ISNotes corpus. We include them in the bridging cases, but label them as *comparative* and link the comparative markable to the antecedent.

- (8) About 200,000 East Germans marched in Leipzig and thousands more staged protests in **three other cities**.

- (9) President Bush, the Canadian prime minister and **14 other members of the Committee**.

Antecedents

As a general principle, one antecedent has to be chosen. In special cases, e.g. comparative cases where two antecedents are needed, the annotator may create two or more links.

- (10) President Bush, the Canadian prime minister and **14 other members of the Committee**.

We include nominal and abstract antecedents, where the anaphors link back to a VP or a clause.

- (11) What is the meaning of life? **The answer** cannot be expressed in one sentence.

The antecedent should be the best fitting semantically related expression. In the case of several possible antecedents, the closest should be chosen.

Bridging should not be used as a substitution category for aggregated coreference, where we need two coreference links to for example state that *all sides* involve *the media* and *the congressman* (in a context where these two expressions do not appear in a coordination).

Link types

As there are different types of links covered under the term bridging in previous annotation efforts, we distinguish a number of bridging types, for purely pragmatic reasons. The phenomena can then be studied separately, if needed, or certain anaphor types can

be excluded when merging data from different source corpora. Cases of the category **bridging-contained**, as described in Baumann and Riester (2012), are not annotated as bridging because it is not an anaphoric phenomenon and as such a special case where the antecedent modifies the bridging anaphor.

(12) *the windows in the room*

(13) *the mother's room or her room*

The annotated bridging link categories are the following: (i) definite bridging links, (ii) indefinite bridging links and (iii) comparative bridging links. Cataphoric bridging links are not allowed.

Annotation procedure

The annotation was done using the annotation tool Slate (Kaplan et al., 2012) using our own annotation guidelines.⁶ The markables, i.e. the gold annotated NPs in OntoNotes, are presented in green. Coreferent entities shown in red are already marked and can thus not be marked as bridging anaphors. Exceptions are the first mention in a coreference chain, which can, of course, be of the category bridging. We refrained from annotating attributes in order not to complicate the annotation process. The annotation involved two annotators (both graduate students in computational linguistics, who have previously been involved in information status annotation) for five WSJ articles, to establish the inter-annotator agreement. The rest of the corpus was annotated by a single annotator.

Difficult annotation decisions

Some cases of bridging are very clear, particularly for definite anaphors that occur in a well-defined relation with their antecedent, e.g. **whole-part** (*the house - the window*). In this case, it is obvious that the definite anaphor requires the antecedent for its interpretation.

Generic use vs. bridging Other cases are less clear, and they are often a question of generic use vs. bridging. Consider the following example that is taken from the Wall

⁶Annotation guidelines:

<http://www.ims.uni-stuttgart.de/institut/mitarbeiter/roesigia/guidelines-bridging-en.pdf>

4. Annotation and data creation

Street Journal and is thus concerned with the US (which is often not explicitly stated, but obvious given the WSJ's location).

(14) **The police** would be waiting.

The question whether *the police* is a generic reference to the concept police or whether a bridging link should be placed between *the police* and the previously mentioned *the US* is not obvious. When does such an entity need an antecedent or when does it simply add (optional) information? In cases of obvious generic use, we do not link the two entities. If we are not referring to the generic class *police*, but more specifically about the police in, say, *Baltimore*, we link the two entities. As a general rule, if the entity is interpretable on its own, we do not link it, e.g. in

(15) When you annotate a text, **bridging anaphors** are the most difficult issue (*not linked in this case*).

Still, this distinction remains a little vague.

Unused vs. bridging Another difficult choice is the distinction between the information status category **unused** (sometimes called **mediated-general**) and bridging, i.e. in a case like

(16) Iran ... **foreign secretary Mottaki**

where some people might consider this a bridging case, as the *foreign secretary Mottaki* is probably not interpretable alone for a typical WSJ reader without the mentioning of *Iran* first. However, others might argue that his discourse referent might already be identified by his name.

Furthermore, while we typically assume entities like *the moon* to be unique, known entities, and thus of the category **unused/mediated-general**, there might be contexts where there are several moons, and one might want to link *the moon* to the entity *the earth* via a bridging relation.

Determining a single antecedent In some contexts, the writer/speaker introduces a topic into the discourse and then talks about aspects referring to this topic. In cases where there are several noun phrases representing this topic, it is not always obvious which NP should be chosen as the antecedent.

- (17) No age group is more sensitive than younger voters, like Ms. Ehman. A year ago this fall, voters under 30 favored George Bush by 56 to 39% over Michael Dukakis, [...]. Voters in **the same age group** backed Democrat Florio 55% to 20% over Republican Courter.

It is relatively obvious that *the same age group* is a bridging anaphor, but whether *younger voters, like Ms. Ehman*, *Ms. Ehman* or *voters under 30* should be chosen as the antecedent remains unclear (and does not really make a big difference in terms of the interpretation of the anaphor).

Resulting corpus

As can be seen in Table 4.5, the corpus consists of 459 bridging links, 114 of which contain an indefinite anaphor, 275 a definite anaphor and 70 are comparative anaphors. Out of these 70 comparative anaphors, 12 have more than one link to an antecedent. The corpus contains 57,709 tokens.

| | |
|----------------|-----|
| Bridging links | 459 |
| Definite | 275 |
| Indefinite | 114 |
| Comparative | 70 |

Table 4.5.: BASHI: corpus statistics

Inter-annotator agreement

Five WSJ articles have been annotated by a second annotator, in order to assess the inter-annotator-agreement. Table 4.8 shows the agreement for the respective categories. We only report the observed agreement, as the expected agreement for linking markables is considered extremely low (as one can potentially link every NP with all preceding NPs) and can thus be neglected.

It can be seen that the agreement is high for comparative anaphora: as these almost always occur with surface markers such as *other*, *another*, *etc.*, they can be easily spotted. The agreement for the chosen antecedent is also higher, as they are typically local antecedents in a rather narrow window. As expected, the agreement for anaphor detection as well as for full bridging resolution is higher for definites than for indefinites. This confirms our hypothesis that for definites, it is easier to decide whether they are

4. Annotation and data creation

| Bridging anaphor type | Anaphor | | | Anaphor+antecedent | | |
|-----------------------|---------|-------|-----------|--------------------|-------|-----------|
| | same | diff. | agreement | same | diff. | agreement |
| Definite | 34 | 13 | 73.9% | 30 | 17 | 63.8% |
| Indefinite | 15 | 11 | 57.7% | 11 | 15 | 42.3% |
| Comparative | 12 | 2 | 85.2% | 10 | 4 | 71.4% |
| Total | 31 | 25 | 70.9% | 51 | 36 | 59.3% |

Table 4.6.: BASHI: inter-annotator agreement on five WSJ articles

anaphoric or not. Overall, for anaphor detection, we achieve an agreement of 70.9% and 59.3% agreement for the overall links. As the overall agreement on the bridging links is rather low (also for other corpora), one could think about evaluating the task of bridging resolution differently than with the typical precision/recall metrics, particularly for contexts such as Example (17).

Format and download

The corpus is made available in the form of a download link⁷. The download contains the annotations in an offset-based XML format as well as CoNLL-12 style columns. For the single anaphor type categories (`definite`, `indefinite`, `comparative`) we have created separate columns, as well as one joint column which contains all the bridging links. For copyright reasons (the OntoNotes data has to be obtained separately via the LDC), the download includes instructions on how to merge the annotations with the actual corpus data and the annotations in the OntoNotes release (words, part-of-speech, coreference, etc.).

4.3.2. SciCorp: coreference and bridging in scientific articles

In this section, we present SciCorp, a scientific corpus of two different disciplines, namely computational linguistics and genetics.⁸ Apart from automatic pre-processing layers, the corpus features three types of manual annotation: coreference clusters, bridging entities and their antecedents, and information status labels. In this thesis, we will focus on the

⁷<http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/bashi.html>

⁸We addressed the research question of resolving coreferent and bridging references in scientific literature in Rösiger and Teufel (2014). In this paper, an earlier version of the corpus was used as a basis for the experiments, but the corpus was not made publicly available as it was only annotated by one person. Over the course of this dissertation, the annotation guidelines and the annotation setting have been improved and new inter-annotator-agreement evaluations are provided.

coreference and bridging information. For more information on the information status labelling, see Rösiger (2016).

We chose scientific text as a domain, as it differs quite heavily from news text, and we are interested in testing the generalisability of our bridging approaches. Scientific text differs from news text mostly with respect to the heavy use of abstract entities such as results or variables, while easy-to-resolve named entities are less frequently used (Gasperin and Briscoe, 2008). The more complex nature of the texts is also reflected in the high proportion of definite descriptions (Watson et al., 2003). These typically require domain knowledge to be resolved. It has been shown in Rösiger and Teufel (2014) that in-domain training data helps improve coreference resolution in scientific text.

This section presents details of the annotation process and describes the new corpus that was annotated by three independent annotators and that can be downloaded from our website.⁹

Corpus creation

The computational linguistics (CL) papers were taken from the ACL anthology, the genetics (GEN) papers from PubMed. Papers were selected blindly, not focusing on one topic, any specific year or the first language of the authors. The CL papers cover various topics ranging from dialogue systems to machine translation; the GEN papers deal mostly with the topic of short interfering RNAs, but focus on different aspects of it. The corpus contains a number of short papers as well as some long papers (see Table 4.10 for details). The manual annotations were performed on plain text versions of the papers.¹⁰ After the annotation, we enriched the corpus with a number of automatic annotations.

Manual annotations

We manually annotated the corpus using the annotation tool Slate (Kaplan et al., 2012). Slate does not feature pre-defined mentions, so the identification of markables was part of the annotation task. The tool shows the whole text with a slide bar at the side and the annotator is asked to mark the markables with different colours depending on the information status category. Coreference and bridging links are also highlighted in

⁹www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/scicorp.html

¹⁰The papers were provided in the framework of the FUSE project (Foresight and Understanding from Scientific Exposition)(McKeown et al., 2016). The CL papers were converted from Latex source by Simone Teufel, the GEN papers by Dain Kaplan and Diarmuid Ó Séaghdha as well as other members of the FUSE project.

4. Annotation and data creation

different colours. Three annotators, all graduate students of computational linguistics, independently annotated the documents according to the following annotation scheme. Detailed annotation guidelines were provided.¹¹ The annotators were given two papers (one from genetics, one from computational linguistics) to familiarise themselves with the task before starting the annotation work on the texts included in this corpus.

The remainder of this section describes the annotation scheme in detail. This fine-grained scheme is based on other schemes (Riester et al., 2010; Poesio and Vieira, 1998), but has been adapted to this special domain.

Markables

To limit the number of markables, back then we decided to restrict the annotation to definite NPs and allowed only nominal phrases as antecedents for both coreference and bridging anaphors. Therefore, no event reference is covered in the corpus. These are two serious limitations, and in hindsight, we would recommend not making these limitations. It has been proven difficult for the annotators to determine what markables are definite (particularly for special cases such as bare singulars, bare plurals, modifiers, etc.) and in the end, the initial purpose of the restriction, namely being able to annotate more data in less time, did not hold true. However, the corpus annotation has been performed this way, and we will now report the annotation scheme as it was designed back then. Despite these two limitations, we still think it will be beneficial to see how well approaches designed on newspaper text work on out-of-domain corpora, even if only for the (very large) subset of definite anaphors.

We consider the following types of NPs as definite:

- Definite descriptions or similar: NPs starting with the definite determiner *the*, a demonstrative determiner such as *this*, a possessive pronoun like *my* or a universal quantifier such as *all*. Examples: *the most efficient siRNAs*, *the siRNAs*, *all algorithms*.
- Named entities such as *Noam Chomsky*, *siRNAs* but also variables like *x* and *y*.
- Personal pronouns (*we*, *it*, *they*), possessive pronouns (*our*, *their*, *its*) and demonstrative pronouns like *this* or *these*.

¹¹www.ims.uni-stuttgart.de/institut/mitarbeiter/roesigia/annotationguidelines.pdf

Non-markables

Non-markables are the following:

- We do not mark relative pronouns and expletive or pleonastic *it*. *It* in cases like *since it was discovered that ...* is not considered a markable.
- Indefinite NPs, including indefinite descriptions with the indefinite determiner *a*, such as *an important part*. It also comprises existential quantifier phrases like *some siRNAs*, *most siRNAs* or *15 siRNAs*. Bare plurals such as *proteins* are also considered indefinite and are thus not included in the annotation.
- Bare singulars and the existential *there* are also not annotated.

Overview: Annotated categories and links

We label information status and create reference links for a subset of the information status categories. Table 4.7 shows the categories in the annotation scheme and how they interact with the coreference and bridging links: we create coreference links for all entities of the category given and bridging links for all bridging entities.

| | Category | Example |
|-------------------|----------------------------|---|
| Coreference links | Given | We present <u>the following experiment</u> . It ... |
| Bridging links | Bridging | Xe-Ar was found to be in a <u>layered structure</u> with Ar on the surface . |
| | Bridging (self-containing) | The structure of the protein ... |
| | Description | The fact that the accuracy improves ... |
| Categories | Unused | Noam Chomsky introduced the notion of ... |
| without links | Deictic | This experiment deals with ... (<i>non-anaphoric use</i>) |
| | Predicative | Pepsin, the enzyme , ... |
| | Idiom | On the one hand ... on the other hand |

Table 4.7.: SciCorp: categories and links in our classification scheme

Information status categories

Table 4.7 overviews the information status categories. As mentioned above, we will not go into detail here, but focus on the given and bridging category and their coreference and bridging links. As mentioned above, for a description of the information status categories please refer to Rösiger (2016).

Given We consider a definite noun phrase **given** when the entity refers back to a discourse entity that has already been introduced in the discourse and is thereby known to the reader. This includes lexically new material, pronouns and repetitions or short forms of entities that have been referred to before. **Given** entities include synonyms and are not limited to entities that have the same head.

Bridging For bridging anaphors, the text presents an antecedent NP which does not stand in the relation of identity, but in some other form of relation to the associative phrase. The antecedent may be an associate in a typical relation such as **part-of**, **is-a**, or any kind of associate as long as there is a clear relation between the two phrases. We do not limit bridging references to any predefined relations.

Bridging (self-containing) In some constructions, e.g. genitives or PP modifiers, we identify a bridging relation between the head noun phrase and the modifier. We consider them **bridging self-containing** and do not create a link.

(18) *The structure of the protein*

(19) *the thoracic circuit stage in HK mutants*

(20) *the giant fiber escape pathway of Drosophila*

Attributes We additionally annotate two attributes that are only applied to entities in a coreference chain (mostly given entities, but also to the first-mention entities).

- +/- **Generic**: Generic expressions include reference to a kind, i.e. a general quantification, whereas a specific reading has a fixed referent. This means that we know which exact referent is selected of the set of entities that fulfil the description.

(21) Generic:

In 2006, they shared the Nobel Prize in Physiology or Medicine for their

work on RNA interference in the nematode worm C. elegans. **C. elegans** is unsegmented, vermiform, and bilaterally symmetrical.

(22) Specific:

We present the following experiment. **It** deals with ...

- +/- **Part of compound:**

It is controversial whether non-heads of compounds should be markables (when they are definite since we only mark definite NPs). On the one hand, one might want to include them in the set of coreferential entities, but on the other hand, they do not allow for anaphoric reference, cf. Example (23).

(23) The siRNA activity. *It ...

We decided to include them in the list of mentions when they can be coreferenced to other mentions, but to mark them with the attribute **part-of-compound** so that they can be filtered out if required. Adjectives and common nouns are never marked. This means that in Example (24), we have got two markables, *the siRNA experiments* and *siRNA*.

(24) *The siRNA experiments*

Coreference annotation

All anaphors must be definite noun phrases. Definite noun phrases include bare singulars if the insertion of a definite determiner is possible and more plausible than the insertion of an indefinite determiner. Again, bare plurals are excluded as they are considered indefinite.

(25) The efficiency of RNAi is
RNAi efficiency can also be influenced by ...

The antecedent can be any type of nominal phrase (indefinite, definite, named entity, etc.). Abstract anaphora are not included in the corpus, i.e. verbal phrases or clauses are excluded as antecedents of a coreferent anaphor. The links follow the chain principle, so we always choose the closest occurrence of the entity as the antecedent.

Bridging annotation

As for coreference anaphors, bridging anaphors must also be a definite noun phrase as described before. The antecedent can be any type of nominal phrase. The links do not have to follow the chain principle, the annotators are told to choose the best fitting antecedent, not the last occurrence in the text. Bridging antecedents can also have two antecedents (and two links), if this fits best. In our scheme, bridging links are only annotated when there is a clear relation between the two entities. As we do not pre-define possible bridging relations, this definition is a little vague, but it is necessary to keep the task as general as possible.

Agreement study

After the annotators familiarised themselves with the annotation task and annotated two papers that are not part of the final corpus, we analysed the inter-annotator-agreement on two papers (one GEN, one CL) that are part of the corpus and computed Fleiss' κ (Fleiss, 1971). As can be seen in Table 4.8, for information status we achieve a κ between 0.68 (GEN) and 0.73 (CL), which is considered moderate agreement (Landis and Koch, 1977).¹² It is not surprising that the number for CL is a little higher given the fact that the annotators are students of computational linguistics.

| Agreement | GEN | CL |
|-----------|------|------|
| Actual | 0.79 | 0.82 |
| By chance | 0.34 | 0.34 |
| κ | 0.68 | 0.73 |

Table 4.8.: SciCorp: overall inter-annotator-agreement (in κ)

Table 4.9 shows the inter-annotator agreement for the single categories.¹³ It can be seen that **given**, **deictic** and **idiom** entities are easier to reliably annotate while **bridging**, **description**, **unused** and **predicative** entities are more difficult.

For the coreference links the agreement was 0.81 and for bridging links it was 0.62. The agreement for the attribute **generic** was 0.51 and for **part-of-compound** 0.85.

¹²Calculation based on markables. When there was disagreement about the markables, we resolved these cases via discussion between the three annotators. Parameters of the kappa computation: $k=8$, $N=3$, $n=552$ for GEN and $n=482$ for CL.

¹³Calculation for category x based on those mentions where one of the annotators classified it as category x.

| Category | GEN | CL |
|------------------------|------|------|
| κ given | 0.72 | 0.77 |
| κ bridging | 0.62 | 0.63 |
| κ bridging (sc) | 0.68 | 0.74 |
| κ description | 0.67 | 0.69 |
| κ unused | 0.65 | 0.67 |
| κ deictic | 0.73 | 0.76 |
| κ predicative | 0.53 | 0.57 |
| κ idiom | 0.85 | 0.83 |

Table 4.9.: SciCorp: inter-annotator-agreement for the single categories (in κ)

Annotation challenges

This section presents a few observations concerning some of the difficulties that came up during the annotation. We include this here because we think it might be helpful for further, similar annotation experiments.

One major obstacle was that not all the texts were written by native speakers. For example, sometimes the authors clearly had problems with definite articles. If the annotators are asked to mark only definite NPs and the authors leave out the definiteness marker, this is problematic. We resolved these cases by adding a rule to the guidelines that in cases where it was very clear that the author made a mistake, the entity should be marked. However, a few cases remained where it was less clear, and we did not mark these cases. Paying more attention to paper selection in the first place would have helped here. With this aspect in mind, while we originally intended to limit the annotation to definite NPs due to time constraints, in hindsight we think that it turned out to be more difficult and as a result also slower to identify definite markables than to just annotate every NP, disregarding their definiteness. We nowadays think that indefinites can be bridging anaphors and should, in any case, be included.

The annotation of the attribute `generic` turned out to be difficult for the annotators, with an agreement of only 0.51. As the decision whether an entity is generic or not is not trivial (and probably needs much more detailed guidelines), the annotation of `+/-generic` should be the focus of an annotation task, not a by-product. Nevertheless, we include this attribute in the distribution of the data. For `part-of compound`, this problem did not exist: deciding whether something is part of a compound or not is trivial enough to be annotated at the same time.

4. Annotation and data creation

For the GEN texts it would have been nice to include experts as it was difficult to understand what refers to what in a few cases.

Resulting corpus

| CL | | | GEN | | |
|----------|--------|-----------|----------|--------|-----------|
| (doc id) | words | sentences | (doc id) | words | sentences |
| 9704004 | 6,104 | 217 | 346034 | 3,040 | 116 |
| 9505025 | 5,085 | 222 | 135797 | 2,437 | 74 |
| 9812005 | 1,368 | 59 | 340083 | 4,030 | 154 |
| 9708001 | 4,416 | 160 | 149283 | 5,404 | 228 |
| 9607011 | 2,804 | 104 | 152674 | 5,711 | 223 |
| 9606028 | 1,981 | 68 | 148263 | 7,286 | 253 |
| 9606011 | 3,276 | 138 | 153544 | 8,103 | 275 |
| Total | 25,034 | 968 | Total | 36,011 | 1,320 |

Table 4.10.: SciCorp: corpus statistics

| | Total | CL | GEN |
|----------------------------|-------|-------|-------|
| Markables (incl. Unmarked) | 9,407 | 3,879 | 5,528 |
| Markables (excl. Unmarked) | 8,708 | 3,564 | 5,144 |
| Given | 4,730 | 1,851 | 2,879 |
| Bridging | 1,366 | 561 | 805 |
| Bridging(sc) | 321 | 113 | 208 |
| Description | 1,034 | 507 | 527 |
| Unused | 1026 | 424 | 602 |
| Deictic | 70 | 45 | 25 |
| Predicative | 147 | 58 | 89 |
| Idiom | 14 | 5 | 9 |
| (Unmarked | 699 | 315 | 384) |
| Links | 6,201 | 2,436 | 3,765 |
| Coreference | 4,712 | 1,837 | 2,875 |
| Bridging | 1,489 | 599 | 890 |

Table 4.11.: SciCorp: distribution of information status categories, in absolute numbers

Our annotated corpus contains 14 full-text scientific papers, 7 papers for each of the two disciplines. As shown in Table 4.10 and 4.11, the annotated computational linguistics papers contain 968 sentences, 25,034 words and 3,564 annotated definite descriptions

while the annotated genetics papers contain 1,320 sentences, 36,011 words and about 5,144 definite descriptions; the genetics subcorpus is thus a little bigger than the CL one.

The gold annotation was created by taking the majority vote of the three annotators. Disagreements with respect to the annotation or the markables were resolved via discussion between the annotators.

Table 4.11 and Table 4.12 show the distribution of categories in absolute numbers and in percent.

| Category | CL | GEN |
|--------------|------|------|
| Given | 51.9 | 56.0 |
| Bridging | 15.7 | 15.6 |
| Bridging(sc) | 3.2 | 4.0 |
| Description | 14.2 | 10.2 |
| Unused | 11.9 | 11.7 |
| Deictic | 1.3 | 0.5 |
| Predicative | 1.6 | 1.7 |
| Idiom | 0.1 | 0.2 |

Table 4.12.: SciCorp: distribution of information status categories, in percent

Automatic annotations and format

For the pre-processing of the texts, we used the Stanford Core NLP pipeline¹⁴ to automatically do tokenisation, part-of-speech (POS) tagging, constituency parsing and named entity recognition.

Our distribution of the data contains the source PDF and plain text versions of the papers, the annotated categories and links in an offset-based format as well as the coreference annotations in the tabular CoNLL-12 format.

4.3.3. GRAIN: coreference and bridging in radio interviews

GRAIN is a corpus of German radio interviews and is annotated on multiple linguistic layers.¹⁵ We will not go as much into detail as for the other two corpora, for two reasons: (i) the creation of GRAIN was a joint effort of a number of IMS collaborators, where I was involved in the training of the information status annotators and overall guidance of

¹⁴nlp.stanford.edu/software/corenlp.html

¹⁵Persistent identifier: <http://hdl.handle.net/11022/1007-0000-0007-C632-1>

4. Annotation and data creation

the annotation process, and (ii) the annotation followed the RefLex scheme (Baumann and Riester, 2012), based on the newest guidelines in Riester and Baumann (2017). However, we will present the main properties of the corpus and also introduce the main ideas of the RefLex scheme.

The corpus consists of German radio interviews of about 10 minutes each. A subpart of the corpus has been annotated manually, but the biggest part contains a number of automatic annotations in parallel, which serve as a silver standard. Twenty of the interviews have been selected for the gold standard (manually annotated) part of the corpus. Three additional interviews have been used to introduce the annotators to the annotation task and for training. The 20 gold interviews have been manually annotated with syntactic information (part-of-speech, parses for a subset of the corpus) and referential information status according to the RefLex scheme, which has also been the guideline schema for the DIRNDL corpus.

The RefLex scheme RefLex distinguishes information status at two different dimensions, namely a referential and a lexical dimension. The referential level analyses the information status of referring expressions (i.e. noun phrases) according to a fine-grained version of the given/new-distinction, whereas the lexical level analyses the information status at the word level, where content words are analysed as to whether the lemma or a related word has occurred before. In the case of GRAIN, only referential information status was annotated, i.e. every NP in the text has been categorised as to whether it is *given/coreferential*, *bridging*, *deictic*, *discourse-new*, *idiomatic*, etc. In contrast to the information status annotation in SciCorp, indefinites are also marked (as *discourse-new*). Bridging anaphors are thus a subclass of referential information status and are labelled as *r-bridging*. Coreferent expressions are labelled as *r-given* (except the first mention in a coreference chain, which can, of course, be of a different category). On the referential level, indefinite expressions are considered to be *discourse-new* and are thus treated as expressions of the information status category *r-new*. Therefore, the bridging and coreferent anaphors in our data are always definite. However, there are no further restrictions in terms of pre-defined relations between the bridging anaphor and antecedent, or in terms of entity and event antecedents. Antecedents can be nominal, verbal or clausal. Besides the labels for referring expressions, the annotations also contain coreference chains and bridging links.

Inter-annotator-agreement Each of the interviews was annotated independently by two annotators, applying the Slate tool (Kaplan et al., 2012). Adjudication was either done by a third person, or in a discussion round of the project group. The inter-annotator-agreement has been studied in two recent student theses (Pagel, 2018; Draudt, 2018). They report that for markables with the same span, the inter-annotator-agreement is substantial, with a Cohen’s κ of 0.75. Five different annotators were involved in the annotation (all students of computational linguistics) and the pair-wise agreement for different annotator pairs (Cohen’s κ) for information status ranges between 0.64 and 0.82. For the bridging category, Pagel (2018) reports a κ of 0.2 up to acceptable κ values of 0.6, dependent on the annotator pair. For more details on the inter-annotator-agreement, please refer to Pagel (2018) and Draudt (2018).

4.3.4. Conclusion

We have presented three resources for bridging. The first resource is called BASHI, an English corpus of Wall Street Journal articles. It can be used together with the ISNotes corpus, on which most current experiments have been conducted, as both corpora are of the same domain and contain comparable guidelines. The BASHI corpus contains 459 bridging links. In terms of the inter-annotator-agreement, the agreement for anaphor detection is 71% and 59% for full bridging resolution, with higher numbers for the subset of definite bridging anaphors and comparative anaphors and lower numbers for indefinite anaphors. Coreference annotations are already contained in the corpus as part of the OntoNotes annotations. We will use this corpus to design our bridging resolver and test the generalisability of previous experiments performed on the ISNotes corpus.

The second resource is called SciCorp, an English corpus of scientific articles from two disciplines, computational linguistics and genetics. As this corpus is of a different domain than ISNotes and BASHI, we will use it to assess how our bridging resolver works on other domains than news text. It contains 1366 bridging pairs. The inter-annotator agreement for bridging resolution is in a similar range than for BASHI, with 62% for genetics and 63% for computational linguistics. We have additionally annotated coreference as well as some other information status classes.

The third resource is called GRAIN, a German corpus of radio interviews. The annotations follow the same guidelines as the ones used for the DIRNDL corpus, the only available German corpus at the time of the experiments, and contain 274 bridging pairs.

4. Annotation and data creation

The inter-annotator-agreement has been studied in two student theses, which report an agreement of up to 60% for bridging resolution. Coreference has also been annotated.

Overall, we have created three medium-sized corpora aimed at providing data for bridging resolution in English and German. Bridging has been annotated reliably in these corpora, with inter-annotator-agreement values around 60%. The corpora will serve as a basis for the experiments in the remainder of the thesis. As the annotations in GRAIN were only recently completed, experiments using GRAIN could not be included in this thesis. However, bridging in GRAIN has been the study of a recent student thesis (Pagel, 2018), and our joint results on bridging in German data have been published in Pagel and Rösiger (2018).

In the next part of the thesis, tool creation, the focus is set on developing anaphora resolution tools based on the available and newly created data.

5. Coreference resolution

Research Question 3: Tool creation

Are there openly available tools aiming at providing automatic annotations on unseen text? If not, can we create tool resources to fill the research gap?

Coreference resolution has been extensively addressed in NLP research, e.g. in the CoNLL shared task 2011 and 2012 (Pradhan et al., 2011, 2012) and in the SemEval shared task 2010 (Recasens et al., 2010). Nowadays, most NLP conferences feature coreference resolution as an own track as well as workshops focusing on coreference resolution. The recent CORBON workshops in 2016 and 2017 and the CRAC workshop in 2018 were designed to address special cases of coreference that go beyond “simple” entity coreference, such as for example abstract anaphora/event reference. Coreference resolution, at least for English, has reached a state where the performance for standard entity coreference has reached a satisfactory level, and performance on the standard benchmark datasets keeps getting improved year by year. Furthermore, work on the handling of special cases such as event reference or zero anaphora is in progress.

Most coreference research focuses on English, resulting in a number of high performing, openly available English coreference systems, e.g. Clark and Manning (2016b), Durrett and Klein (2013) or Björkelund and Kuhn (2014).

For German, however, there has been less work. Since the SemEval shared task 2010, only a few systems have been improved or developed, such as the rule-based CorZu system (Klenner and Tuggener, 2011; Tuggener and Klenner, 2014) or Krug et al. (2015)’s system, which is tailored to the domain of historical novels and focuses on the resolution of characters.

For coreference, our contribution to the tool creation step is thus to adapt an existing learning-based coreference resolver for English to German. Figure 5.1 highlights this contribution in our pipeline. The newly developed coreference tool for German will serve as the basis for further validation experiments in the next step, e.g. on the role of prosody on coreference.

5. Coreference resolution

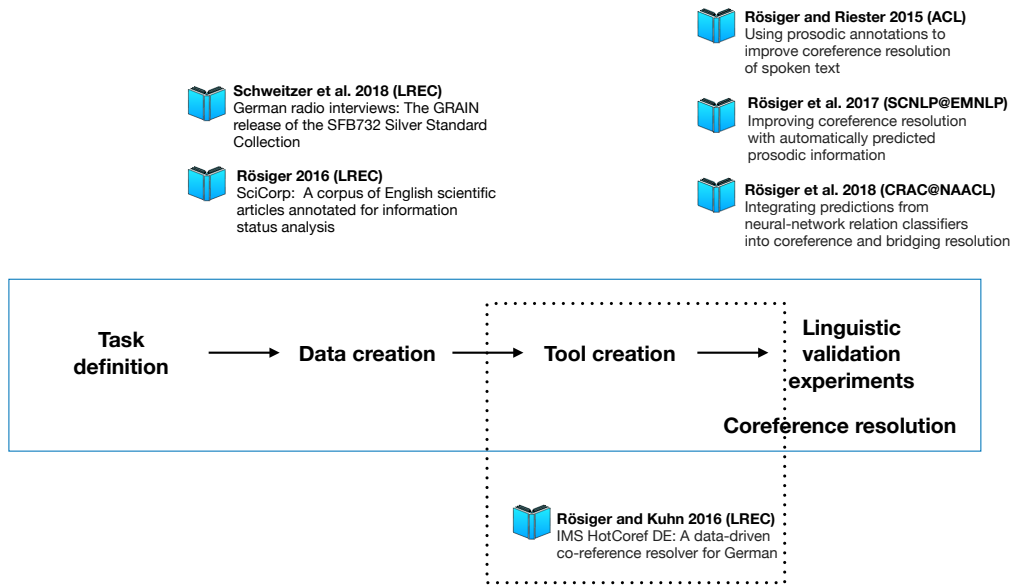


Figure 5.1.: Contribution and workflow pipeline for coreference: tool creation

5.1. Existing tools and related work

In the SemEval shared task 2010 on coreference resolution in multiple languages, a number of systems participated in the German track: BART (Broscheit et al., 2010a,b), SUCRE (Kobdani and Schütze, 2010), TANL-1 (Attardi et al., 2010) and UBIU (Zhekova and Kübler, 2010). Four different settings were evaluated in the shared task, using external resources (open) or only the resources provided (closed), combined with gold vs. regular preprocessing. In our own SemEval post-task evaluation, we will compare the performance of the three best-performing systems, BART, SUCRE and TANL-1, in Section 5.2.3.

Since then, only a few systems have been developed or improved. Ziering (2011) improved the scores of SUCRE by integrating linguistic features. This resulted in an improvement of the average of MUC and BCUBE of about 5 percentage points. It is, however, difficult to compare these numbers as the official scorer scripts have changed and as neither the system output nor the system itself are available.

Klenner and Tuggener (2011) implemented CorZu, a rule-based incremental entity-mention co-reference system which has received the best results on TüBa-D/Z, the benchmark dataset for German, since SemEval. The system was improved in Tuggener

and Klenner (2014). Krug et al. (2015) compared their own rule/pass-based system tailored to the domain of historical novels with CorZu in this specific domain, restricting coreference resolution to the resolution of persons, and found that their own system outperformed the rule-based CorZu. As this system does not aim at resolving general coreference, we will not include it in our overview of general German coreference systems.

Mikhaylova (2014) adapted the IMSCoref system (Björkelund and Farkas, 2012), a predecessor of the IMS HotCoref (Björkelund and Kuhn, 2014), to German as part of a Master thesis. To the best of our knowledge, this system was not made publicly available.

The following section introduces the available systems that have been proposed for coreference resolution in German text in more detail.

CorZu Klenner and Tuggener (2011) presented an entity-mention model for German and English with restrictive antecedent accessibility. The motivation for this approach was the flaws of the mention-pair approach, such as the restriction to local decisions, i.e. only pairs of mentions are classified prior to the construction of the coreference clusters, without being able to enforce global constraints. A postprocessing clustering step has been proven to help remove inconsistencies by ensuring the transitivity of the pairs, but the problem of unbalanced data remains. Therefore, they implemented an incremental entity-mention model where the candidate pairs are evaluated on the basis of the already formed coreference sets. The main idea is that one virtual prototype of the cluster bears all morphological and semantic information of the members of the cluster and is used to compare it with another mention.

The system uses only automatic preprocessing, including a syntactic parser, and extracts markables from the chunks based on part-of-speech tags delivered by the preprocessing pipeline. The extracted markables are then resolved per type, in the following way:

- reflexive pronouns are resolved to the subject governed by the same verb;
- relative pronouns are resolved to the nearest preceding NP;
- personal and possessive pronouns are resolved to morphologically compatible candidates (NE, nouns and pronouns) within a window of three sentences;
- named entities either match completely or the antecedent must be more than one token and all tokens of the anaphor must be contained in the antecedent (*Hillary Clinton ... Clinton*);

5. Coreference resolution

- demonstrative pronouns are mapped to nominal NPs by matching their heads;
- definite NPs are resolved to other NPs if they match completely, without the determiner;
- to find non-matching anaphors, they perform hyponymy and synonymy search in GermaNet (Hamp and Feldweg, 1997).

As can be seen from the rules, the model makes heavy use of the binding theory (Chomsky, 1981) and the c-commanding constraints explained in Section 2.1. In Example (1), *sie* and *Clinton* cannot be coreferent, as the pronoun is c-commanded by *Clinton*.

(1) Clinton traf sie.

Hence, the pair does not need to be considered at all. All mentions in the already formed Clinton cluster are transitively exclusive and can be disregarded as antecedents.

Based on TüBa-D/Z as the gold standard dataset, they calculate the salience of a dependency label as the number of coreferent mentions in the gold standard that bear that label, divided by the total number of coreferent mentions. As a result, they get a hierarchy of salient dependency categories according to which the antecedent candidates are ranked, where subjects are for example more salient than objects, which are in turn more salient than other categories.

We will include Corzu in the evaluation and compare the performance of CorZu against our newly developed model.

BART Broscheit et al. (2010a,b) presented an adaptation of their system BART to German. They base their system on the simple pair-wise approach by Soon et al. (2001), which we explained in Section 3.1, using TüBa-D/Z for training and testing. First, they extract all nominal projections if their grammatical function is not included among the following ones: appositions, items in copula constructions, noun phrases governed by *als* and the Vorfeld-*es*. They state that all cases of non-referring *es* can be easily identified by their grammatical function label, making use of hand annotated information.

As features, they use common features taken from the literature, including distance, part-of-speech tags, grammatical functions and head matching, as well as semantic class distinctions. The semantic class labels are based on GermaNet. They also include a couple of additional features, including information on quoted speech, the distance in the parse tree, partial match and GermaNet relatedness.

SUCRE SUCRE (Kobdani and Schütze, 2010) is a coreference system that is able to separately carry out noun, pronoun and full coreference resolution. It is based on a relational database model and a regular feature definition language. The main algorithm is based on Soon et al. (2001), where positive and negative training instances are extracted from the gold data, and then classified as to whether they are coreferent or not. After the classification, they apply best-first decoding, i.e. the antecedent candidate with the highest likelihood is chosen to be the antecedent. There are four classifiers integrated into SUCRE: decision tree, Naive Bayes, support vector machines and maximum entropy. They state that the best results were achieved using decision trees.

SUCRE also participated in the SemEval-2010 shared task, in the gold and regular closed annotation tracks of six languages. SUCRE’s feature set for German was improved in a Master thesis (Ziering, 2011).

UBIU UBIU (Zhekova and Kübler, 2010) is a language-independent system for detecting full coreference resolution of named entities, pronouns, and full noun phrases. It applies a statistical model, making use of memory based learning. It is language independent in the sense that it only requires syntactic dependency parses and some effort to adapt the feature extractor to the language.

UBIU was also one of the participating systems in the SemEval-2010 shared task, where they submitted systems for all languages.

TANL-1 TANL-1 (Attardi et al., 2010) is another system that participated in the SemEval-2010 shared task. The system makes use of dependency parses and similarity clustering. In the first phase of the system, a binary classifier based on maximum entropy is used to classify pairs of mentions. In the second phase, the mentions detected in the first phase are clustered according to the output of the classifier, using a greedy clustering algorithm. Hereby, each mention is compared to all previous mentions. If the pair-wise classifier suggests a probability greater than a given threshold, it is assigned to that entity. They also apply best-first decoding.

Wallin and Nugues 2017 Wallin and Nugues (2017) present a coreference system for Swedish and German based on distant supervision that does not use manually annotated data. For training, they apply the Stanford CoreNLP pipeline including coreference to parallel corpora in English-Swedish and English-German. To transfer the coreference annotations from the English text, they automatically align words and afterwards carry

5. Coreference resolution

out the mention transfer. Based on these transferred mentions, they then apply the mention-based approach of Soon et al. (2001) using a number of different classifiers: C4.5, random forest, and logistic regression. For German, they evaluate on a subpart of TüBa-D/Z, where they obtain a CoNLL score of 13.16 using the transferred mentions and 36.98 using gold mentions. These results are of course a bit lower than the state-of-the-art results on TüBa-D/Z for rule-based and supervised methods (although we cannot directly compare against this method, as it does not use the whole TüBa-D/Z corpus), as errors in the alignment stage and the predicted coreference resolution for English are propagated to the actual coreference resolution part.

5.2. A coreference system for German

This section presents a data-driven coreference resolution system for German that has been adapted from IMS HotCoref, a coreference resolver for English. It describes the difficulties when resolving coreference in German text, the adaptation process and the features designed to address linguistic challenges brought forth by German. We report performance on the reference dataset TüBa-D/Z and include a post-task SemEval 2010 evaluation, showing that the resolver achieves state-of-the-art performance. We also include ablation experiments that indicate that integrating linguistic features increases results. Furthermore, this section describes the steps and the format necessary to use the pre-trained resolver on new texts. The tool is freely available for download. Parts of this research have been published in Rösiger and Kuhn (2016).

5.2.1. System and data

IMS HotCoref

As a basis for the adaptation, we chose the English IMS HOTCoref system (Björkelund and Kuhn, 2014). The IMS HotCoref system models coreference within a document as a directed latent rooted tree.¹ The benefits of such a latent tree-based approach have already been illustrated in Section 2.1, the most important one being that one can learn more meaningful antecedents, and can profit from non-local features, which are not restricted to only the current pair of mentions. The problem with using non-local features is that it requires an approximate search algorithm to keep the problem tractable. The focus in the original paper was set on the machine learning side, particularly on search

¹The name HotCoref stands for higher order tree coreference.

strategies. They investigate different perceptron techniques and suggest to use a modified version of LaSo (Learning as Search Optimization, Daumé and Marcu (2009)), where updates are delayed until each document is processed. As we base our adaptation on the already implemented features, we will give an overview of the available feature types.

Local features The local features are the same as in the predecessor, Björkelund and Farkas (2012), and include different types of (mostly linguistic) information. Features are for example based on the surface forms of the anaphor and the antecedent, the part-of-speech tags of (parts of) the mentions or the previous and following word, syntactic features where subcategorisation frames and paths in the syntax tree are analysed, distance-based features, semantic class information as well as a number of other features.

Non-local features As non-local features, they introduce features such as the size of the already formed clusters, the shape of a cluster in terms of mention type or local syntactic context, e.g. paths in the syntax tree.

TüBa-D/Z

The reference corpus for coreference resolution experiments in German is TüBa-D/Z² (Naumann and Möller, 2006), a gold annotated newspaper corpus of 1.8 M tokens with articles from the daily issues of “die tageszeitung” (taz). To evaluate our system, we use version 10 (v10) as the newest dataset available, as well as version 8 (v8), as this was used in the SemEval shared task. We adopt the official test, development and training set splits for the shared task data. For version 10, there was no standard split available, so we split the data ourselves.³

TüBa-D/Z gold annotated version The gold annotations for both v8 and v10 were obtained via download from the TüBa-D/Z download page. TüBa-D/Z v10 comes in a number of different formats, including PENN for c-structure trees (with fine-grained syntactical annotations where topological fields such as “Vorfeld” are marked) and a CoNLL-2011 file.

In order to use TüBa-D/Z with the coreference resolver IMS HotCoref, we took the following steps:

²<http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html>

³We take the first 727 docs as test, the next 727 docs (728-1455) as dev and the remaining 2190 documents as training data. This equals a 20-20-60 test-dev-train ratio.

5. Coreference resolution

- Named entities (NEs): as the coreference resolver cannot process embedded named entities, we removed nested entries. In Example (2), we would for example remove the information that New York is a location.

(2) (Die (New York)_{LOC} Times) _{ORG}

- Lemmata: we removed all characters which are not part of the actual lemma, e.g. the # in *hinter#ziehen* or the %aux in *müssen%aux*
- Parses: for the syntactical features to work, we simplified the rich vocabulary of the annotated gold parses, i.e. we removed the subcategories after :-, :- and =. This means that we changed the labels *NX:OA* to *NX* and in a second step, *NX*, *PX* and *SIMPX* was changed to *NP*, *PP* and *S*.
- Format: we adjusted the format so that it matches the conventions for the CoNLL-12 format.

TüBa-D/Z predicted version The predicted version for v8 (i.e. using only automatic preprocessing) was obtained from the SemEval shared task, to be compatible with the other shared task systems, and had to be converted into the CoNLL-12 format. We parsed the text with the Berkeley parser (Durrett and Klein, 2013), post-processed the parses by inserting NPs into flat PPs (also embedded ones), inserted NPs into single-word NPs which were not marked as NPs before, and adjusted PNs where they overlap with NPs. We also inserted NPs into conjugated NPs (CNP) in order to be able to extract them as markables. The parsing adaptations are described in more detail below. We also included part-of-speech and morphological tagging using the Mate tool (Bohnet and Nivre, 2012) and named entities as recognised by the Stanford named entity system (Faruqui and Padó, 2010). As we are using external tools, our system can only be evaluated in the open track of the shared task.

For the predicted version of TüBa-D/Z v10, we processed the data using the same steps as described above. The steps involved are also explained in the section on how to run the tool on your own text, in Section 5.2.5.

5.2.2. Adapting the system to German

Mention extraction

The goal of the mention extraction module is to achieve high recall and to provide the coreference resolver with a high number of correctly determined mentions. This is crucial for the performance of the final system.

Mention extraction for TüBa-D/Z v8 (the SemEval data) The following experiments were performed on the predicted version of TüBa-D/Z v8 (the SemEval data). First, we computed the recall for a number of different markable types. As can be seen in Table 5.1, the recall is quite low when extracting NPs only (28 percent). Adding other types, e.g. personal pronouns (PPER), increases the recall to 36 percent and finally to 41 percent by extracting possessive pronouns (PPOSAT), relative pronouns (PRELS) and interrogative pronouns (PWS). As names are sometimes annotated as DL in the constituency parse, adding DL as a category increases recall to 46.6 percent. However, the final recall is still low, which is why further adjustments are necessary.

| Tag | Description | Recall |
|----------|------------------------|--------|
| NT-NP | noun phrases | 28.2 |
| T-PPER | personal pronouns | 36.5 |
| T-PPOSAT | possessive pronouns | 39.1 |
| T-PWS | interrogative pronouns | 41.1 |
| T-PRELS | relative pronouns | 41.5 |
| NT-DL | names | 46.6 |

Table 5.1.: IMS HotCoref DE: performance of the mention extraction module: markable types and their recall in percent, for TüBa-D/Z 8

Post-processing of the parse bits Parse bits are parts of a constituency parse that span a certain number of words. There are a number of reasons why the extracted parse bits from the constituency parser (Durrett and Klein, 2013) and the annotated coreferent NPs do not match. The first problem is that the annotated PPs are flat, i.e. they do not contain embedded NPs. Hence, we need to insert NPs into flat PPs. In Example (3), markable 22 (*seinem umfanglichen dichterischen Schaffen*) does not have a matching NP in the original parse bit.

5. Coreference resolution

| | Token | Before | | | After |
|-----|---------------|--------|-----------|-------------|-----------|
| | | POS | Parse bit | Coreference | Parse bit |
| (3) | Aus | APPR | (S(PP* | - | (S(PP* |
| | seinem | PPOSAT | * | (22 | (NP* |
| | umfänglichen | ADJA | * | - | * |
| | dichterischen | ADJA | * | - | * |
| | Schaffen | NN | *) | 22) | *) |

Of course, embedded PPs also require the insertion of NPs, as illustrated in Example (4).

| | Token | Before | | | After |
|-----|--------------|---------|-----------|-------------|-----------|
| | | POS | Parse bit | Coreference | Parse bit |
| (4) | wegen | APPR | (VP(PP* | - | (VP(PP* |
| | seiner | PPOSAT | | (16 | (NP* |
| | Gegnerschaft | NN | * | | |
| | zur | APPRART | (PP* | - | (PP* |
| | Diktatur | NN | | (18 | (NP* |
| | Primo | NE | * | | (NP* |
| | de | NE | * | - | |
| | Riveras | NE | *) | 16) 18) | *) |

One issue with the parser output is that single-word proper nouns or common nouns do not have an NP label in the parses, so we need to insert an NP label, as shown in Example (5). We cannot just extract all proper or common nouns as markables as they are typically part of larger NPs, where the single word alone is not considered a markable.

| | Token | Before | | | After |
|-----|-------|--------|-----------|-------------|-----------|
| | | POS | Parse bit | Coreference | Parse bit |
| (5) | Gott | NN | (S* | (497) | (S(NP*) |
| | guckt | VVFIN | * | - | * |
| | uns | PPER | * | - | * |
| | nicht | PTKNEG | * | | * |
| | zu | PTKVZ |) | - | *) |

Conjugated NPs (CNP) do not have embedded NPs, which is why we additionally need to insert NPs into CNPs, shown in Example (6).

| | Token | Before | | | After |
|-----|----------------|--------|-----------|-------------|--------------------|
| | | POS | Parse bit | Coreference | Parse bit |
| (6) | Übersetzungen | NN | (CNP* | (492 (42) | (CNP(NP*) |
| | und | KON | * | - | * |
| | Inszenierungen | NN | *) | (43) 492) | (NP*))) |

Independently of the presence of PPs, some NPs are not annotated by the parser. We have implemented a script that inserts NPs if it finds a determiner that is followed by NN or NE (and maximally 10 arbitrary tokens in between). One example is given in Example (7).

| | Token | Before | | | After |
|-----|---------------|--------|-----------|-------------|--------------|
| | | POS | Parse bit | Coreference | Parse bit |
| (7) | der | ART | * | (2 | (NP |
| | deutsche | ADJA | * | | * |
| | Bundeskanzler | NN | * | 2) | (*) |

The parsing adjustments have a large effect on the recall, as can be seen in Table 5.2. The final recall when using predicted annotations is about 78%. The remaining 22% are not extracted mainly due to parsing errors. With gold annotations, the recall is about 99%.

After all these adjustments, there are still gold markables for which we do not have a matching NP in the parse tree. Adding these in automatically (where the tree allows it) leads to an increase in markable detection from 78 to 91%. As this involves gold information, we do not use this information in our experiments. In some situations, the tree does not allow a multi-word NP, e.g. if the node is the start of a markable but the parse has a phrase end. These account for the remaining difference between 91 and 100 percent recall.

5. Coreference resolution

| | Recall |
|-----------------------|--------|
| Basic markables | 46.6 |
| NPs in PPs | 66.2 |
| NPs in PPs (embedded) | 68.0 |
| Single word NPs | 74.6 |
| Adjusting CNPs | 75.6 |
| Inserting NPs | 78.0 |

Table 5.2.: IMS HotCoref DE: performance of the mention extraction module after the respective parse adjustments, recall in percent on TüBa-D/Z version 8.

Mention extraction for TüBa-D/Z v10 We also analysed whether we could use the same markables with the newer version, TüBa-D/Z v10. As some changes have been made in the newer version, we ended up with a different set of markables, which is presented in Table 5.3. Interestingly, despite the slightly different markables, the markable extraction module has the same performance on Tüba-D/Z v8 and v10: 78% using the predicted version and 99% using gold annotations.

| Tags | Description | Recall |
|-------------------------|------------------------|--------|
| NPs (after adjustments) | noun phrases | 43.0 |
| PPER | personal pronouns | 59.3 |
| PPOSAT | possessive pronouns | 68.1 |
| PRELS | relative pronouns | 74.0 |
| PDS | demonstrative pronouns | 74.9 |
| PRF | reflexive pronouns | 76.1 |
| PN | proper noun phrases | 76.1 |
| NE (ORG,LOC,PER,GPE) | named entities | 78.4 |

Table 5.3.: IMS HotCoref DE: performance of the mention extraction module after the respective parse adjustments, recall in percent on TüBa-D/Z version 10.

As the final pre-trained model is based on TüBa-D/Z v10, the final default markables for German were set to be NPs with the label NP or PN in the parse bit, personal pronouns (PPER), possessive pronouns (PPOSAT), relative pronouns (PRELS), demonstrative pronouns (PDS), reflexive pronouns (PRF) and named entities with the label LOC, PER, GPE and ORG.

Number and gender information

In the English version of IMS HotCoref, number and gender information comes in the form of a lookup from lists created by Bergsma and Lin (2006). We decided to include gender and number prediction in the pre-processing and rely on the predicted information. We have included gender and number lookup lists for personal and possessive pronouns in case the morphological analyser does not predict a label.

Head rules

The system includes a module that tries to identify the syntactic head of certain syntactic phrases. We have adapted the rules to German. The main rule for German noun phrases is to take the left-most common or proper noun (or named entity), if present, and if this fails, to look for the left-most pronoun. If this also fails, there is a number of backup strategies to find the most proper solution.

Features for German

IMS HotCoref offers a wide range of language-independent features (single and pair-based). We added a number of new features or changes that are explained in the following. After the implementation of the new features, we ran a number of feature selection experiments to come up with a final set of features that performed best. The feature selection process is described after the new features have been introduced.

Lemma-based rather than word form-based Whereas word form-based features are effective for English, due to the rich inflexion, they are less suitable for German. This is why we chose lemmata as a basis for all the features. The following example illustrates the difference, where a feature that captures the exact repetition of the word form suffices in English but where lemmata are needed for German.

- (8) DE: Sie nahm das Buch des Vaters [gen.] und hoffte, **der Vater** [nom.] würde es nicht bemerken.
 EN: She took the book of the father and hoped **the father** wouldn't notice.

F1: Gender agreement Number agreement is one of the standard features used to find suitable antecedents for pronouns. For German, we additionally need gender agreement. Contrary to English, non-animate entities are often not neuter, but feminine or masculine. On the one hand, this makes the resolution more difficult as it introduces

5. Coreference resolution

ambiguity, see Example (9). On the other hand, as shown in Example (10), it might also make the resolution of inanimate objects easier. Note that this feature is mainly relevant for pronominal reference as nominal anaphor-antecedent pairs do not need to have the same gender, see Example (11).

(9) DE: Emma schaute hoch zur Sonne. **Sie** [fem.] schien heute sehr stark.
EN: Emma looked up at the sun. **It** was shining quite brightly.

(10) DE: Das neue Auto [neut.] stand in der Garage [fem].
Es [neut.] sah ziemlich sauber aus.
EN: The new car? was parked in the garage?.
It was rather clean.

(11) DE: Der Stuhl [masc.] ... **die Sitzgelegenheit** [fem.] ...
das Plastikmonster [neut.] .
EN: the chair ... **the seating accommodation** ... **the plastic monster** .

F2: Compound head match Whereas English compounds are multi words where a simple (sub-)string match feature suffices to find similar compounds, German compounds are single words. Therefore, matching a compound and its head, as shown in Example (12), is more complicated.

(12) DE: Menschenrechtskomiteevorsitzender ... **der Vorsitzende**
EN: human rights committee chairman ... **the chairman**

We have implemented two versions to treat these compound cases, a lazy one and a more sophisticated approach. The lazy version is a boolean feature that returns true if the lemma of the head of the anaphor span ends with the five same letters as the head of the antecedent span, not including derivatives ending with *ung*, *nis*, *tum*, *schaft*, *heit* or *keit* to avoid a match for cases like *Regierung* (*government*) and *Formulierung* (*phrasing*).

The more sophisticated version uses the compound splitting tool COMPOST (Cap, 2014). The tool splits compounds into their morphemes using morphological rules and corpus frequencies. Split lists for TüBa-D/Z as produced by COMPOST have been integrated into the resolver. Split lists for new texts can be integrated via a parameter. In this case, the boolean feature is true if the two markables are compounds that have the same head or if one markable is the head of the other markable that is a compound.

F3: GermaNet lookup A GermaNet interface is implemented based on the Java API⁴ to include world knowledge and to allow the lookup of similar words. We have added three features that search for synonyms, hypernyms and hyponyms. They return true if the antecedent candidate is a synonym (hypernym or hyponym, respectively) of the anaphor.

F4: Distributional information Another source of semantic knowledge comes from distributional models, where the similarity in a vector space can be used to find similar concepts. This type of information is particularly important in cases where string match does not suffice, as in Example (13), and GermaNet does not contain both head words.

- (13) DE: Malaria wird von Stechmücken übertragen. **Die Krankheit** ...
 EN: Malaria is transmitted by mosquitoes. **The disease** ...

We thus implemented a boolean feature that is true if two mentions have a similarity score of a defined threshold (cosine similarity of 0.8 in our experiments, can be adjusted), and false otherwise. To compute the similarity score, we use a module in the coreference resolver that extracts syntactic heads for every noun phrase that the constituency parses has predicted, in order to create our list of noun-noun pairs and their similarity values. To get the similarity values, we built a vector space from the SdeWaC corpus (Faaß and Eckart, 2013), part-of-speech tagged and lemmatised using TreeTagger (Schmid, 1994). From the corpus, we extracted lemmatised sentences and trained a CBOW model (Mikolov et al., 2013). This model builds distributed word vectors by learning to predict the current word based on a context. We use lemma-POS pairs as both target and context elements, 300 dimensions, negative sampling set to 15, and no hierarchical softmax. We used the DISSECT toolkit (Dinu et al., 2013) to compute the cosine similarity scores between all nouns of the corpus.⁵

This idea is further explored in more detail on English data in our validation experiments in Section 8.

F5/F6: Animacy and name information Three knowledge sources have been integrated that are taken from Klenner and Tuggener (2011): a list of words which refer to people, e.g. *Politiker* (*politician*) or *Mutti* (*Mummy*), a list of names which refer to females, e.g. *Laura*, *Anne*, and a list of names which refer to males, e.g. *Michael*, *Thomas*, etc. We use this information in two features:

⁴<https://github.com/Germanet-sfs/GermaNetApi/>

⁵The cosine similarity values based on the CBOW model were provided by Max Kisselew.

5. Coreference resolution

The first feature, called person match, is true if the anaphor is a masculine or feminine pronoun and the antecedent is on the people list. It is also true if the antecedent and the anaphor are both on the people list.

The second feature, called gender match names, is true if the antecedent is a female name and the anaphor a singular female pronoun or if the antecedent is a male name and the anaphor a singular male pronoun, respectively.

Other newly implemented features There are a couple more features that are not included in the final set of features, but might still be helpful for other settings or training data. We give a short explanation for each of the features. For more details, please refer to the source code.

- **NumberMatch:**
a boolean feature that returns true if two expressions match in number.
- **GenderMatch:**
a boolean feature that returns true if two expressions match in gender.
- **HeadLemmaExactStringMatch:**
a boolean feature that returns true if the head lemmata of two expressions match.
- **SubStringMatch:**
a boolean feature that returns true if the two noun phrases match in either an adjective or a common noun.
- **Anaphor is Definite:**
a boolean feature that is true if the anaphor contains a definite marker.
- **Anaphor is Demonstrative:**
a boolean feature that is true if the anaphor contains a demonstrative marker.
- **PronounTreat:**
a boolean feature that adapts string match for pronouns, reflecting the fact that the same pronouns tend to refer to the same entity.

Feature selection In IMS HotCoref, three things make the feature selection process complicated: (i) features can have a negative effect on the overall performance, (ii) features can be combined with other features and contribute more as a combined feature

as the two separate features and (iii) features can have negative interactions with other features.

Therefore, we have implemented a feature selection script that adds features incrementally. If the feature improves the overall performance, it gets added as a candidate to the list of features, if not it gets excluded. When adding the next feature, we check whether the combination of the current and the previous feature improves the performance. If so, the previous feature is added as a feature and the current feature is added as a feature candidate. If the performance decreases, we check whether the current feature alone improves performance. If so, the previous feature candidate is removed and the feature is added as a feature candidate. In the end, we also combine features with other features to check whether the combination of two features gives an additional increase in performance.

5.2.3. Evaluation

Performance on TüBa-D/Z-v10

On the newest testset available (TüBa-D/Z, version 10), our resolver currently achieves a CoNLL score of 65.76. Table 5.4 compares the performance of our system using gold annotations with our system trained on predicted annotations (Section 5.2.5 lists the tools involved). Since TüBa-D/Z v10 is a rather new dataset, other systems have not reported their performance on this data. In this thesis, the best result on a dataset is always marked in bold face.

| IMS HotCoref DE using ... | MUC | BCUBE | CEAFM | CEAFE | BLANC | CoNLL |
|---------------------------|-------|-------|-------|-------|-------|--------------|
| gold annotations | 69.64 | 62.85 | 66.63 | 64.79 | 57.18 | 65.76 |
| predicted annotations | 52.57 | 45.13 | 52.44 | 48.22 | 41.23 | 48.54 |

Table 5.4.: Performance of IMS HotCoref DE on TüBa-D/Z version 10:
gold vs. predicted annotations

Performance on TüBa-D/Z v8: SemEval post-task evaluation

In Table 5.5, the official results given on the SemEval-2010 shared task website⁶ are presented. Note that these results have to be taken with a grain of salt, as an older scorer script was used for the evaluation which was later corrected due to a number of

⁶<http://stel.ub.edu/semEval2010-coref/>

5. Coreference resolution

bugs. As mentioned above, four different settings were evaluated in the shared task, using external resources (open) or only the provided resources (closed), combined with gold vs. regular preprocessing.

| System | CEAFE | MUC | BCUBE | BLANC | CoNLL |
|---------------------------|-------|------|-------|-------|-------|
| Closed, gold setting | | | | | |
| SUCRE | 72.9 | 58.4 | 81.1 | 66.4 | 70.8 |
| TANL-1 | 77.7 | 25.9 | 85.9 | 57.4 | 55.5 |
| UBIU | 68.2 | 21.9 | 75.7 | 64.5 | 55.3 |
| Closed, predicted setting | | | | | |
| SUCRE | 59.9 | 40.9 | 64.3 | 53.6 | 54.7 |
| TANL-1 | 49.5 | 15.4 | 50.7 | 44.7 | 38.5 |
| UBIU | 44.8 | 10.4 | 46.6 | 48.0 | 33.9 |
| Open, gold setting | | | | | |
| BART | 66.9 | 51.1 | 73.4 | 62.8 | 63.8 |
| Open, predicted setting | | | | | |
| BART | 61.3 | 45.5 | 65.7 | 57.3 | 57.5 |

Table 5.5.: SemEval-2010 official shared task results for German: F1 values taken from the website.

However, the system outputs are available on the shared task webpage, which is why we can use the newest, bug-free version of the official CoNLL scorer (Pradhan et al., 2014) and re-evaluate the system results as well as compare our own performance against those of the shared task systems. In a post-task SemEval 2010 evaluation our system achieves a CoNLL score of 48.61 in the *open, regular* track and a CoNLL score of 63.61 in the *open, gold* track. Table 5.6 compares our scores with the three best-performing systems in the shared task, BART, SUCRE and TANL-1 as well as with the newer system CorZu.⁷ The CoNLL scores for all systems participating in the shared task have been computed using the official CoNLL scorer v8.01 and the system outputs provided on the SemEval webpage. The scores differ from those published on the SemEval website due to the newer, improved scorer script and because we did not include singletons in the evaluation, as we think they should not be part of the actual coreference evaluation. More detailed scores can be found in Table 5.7.

⁷Performance of CorZu: Don Tuggener, personal communication.

| System | CoNLL gold ⁸ | CoNLL regular |
|------------------------|----------------------------|------------------|
| IMS HotCoref DE (open) | 63.61* | 48.61* |
| CorZu (open) | 58.11 | 45.82 |
| BART (open) | 45.04 | 39.07 |
| SUCRE (closed) | 51.55 | 36.32 |
| TANL-1 (closed) | 20.39 | 14.17 |

Table 5.6.: SemEval Shared Task 2010 post-task evaluation for track *regular* and *gold* (on TüBa 8), excluding singletons

The difference in CoNLL score between CorZu and our system is statistically significant. We compute significance using the Wilcoxon signed rank test (Siegel and Castellan, 1988) at **the 0.05 or *the 0.01 level. The compared pairs are the documents in TüBa-D/Z.

| | MUC | BCUBE | CEAFE | CEAFM | BLANC | CoNLL |
|--------------------------|-------|-------|-------|-------|-------|--------------|
| IMS (open, gold) | 67.43 | 60.90 | 62.50 | 64.12 | 55.49 | 63.61 |
| IMS (open, regular) | 52.11 | 45.55 | 48.61 | 48.17 | 38.47 | 48.61 |
| CorZu (open, gold) | 61.63 | 55.18 | 58.35 | 58.35 | - | 58.11 |
| CorZu (open, regular) | - | - | - | - | - | 45.82 |
| BART (open, gold) | 50.56 | 40.74 | 46.35 | 43.82 | 31.78 | 45.88 |
| BART (open,regular) | 42.46 | 34.64 | 42.01 | 39.52 | 26.64 | 39.70 |
| SUCRE (closed, gold) | 58.42 | 47.25 | 50.26 | 48.99 | 38.86 | 51.98 |
| SUCRE (closed, regular) | 37.64 | 32.32 | 39.00 | 37.31 | 21.7 | 36.32 |
| TANL-1 (closed, gold) | 25.87 | 16.56 | 23.72 | 18.73 | 14.21 | 22.05 |
| TANL-1 (closed, regular) | 15.36 | 9.84 | 17.32 | 13.36 | 7.37 | 14.17 |

Table 5.7.: SemEval-2010: post-task evaluation, excluding singletons

5.2.4. Ablation experiments

For the features presented above, we perform ablation experiments using the gold annotations of TüBa-D/Z v10. Statistical significance is computed for all comparisons against the best performing version, using the Wilcoxon signed ranked test again.

Table 5.8 shows the results when leaving out one of the previously described features at a time. Computing all the features on a word form rather than lemma basis results in the biggest decrease in performance (about 2 CoNLL points), followed by leaving out gender agreement, GermaNet and the animacy features. Two features, compound head

5. Coreference resolution

match and distributional information, only had a minor influence on the performance. We include them here because they have proven to be effective in other settings, e.g. when using regular annotations.

| IMS HotCoref DE | CoNLL |
|-----------------------------------|---------|
| Best performing version | 65.76 |
| - lemma-based | 63.80* |
| - F1: gender agreement | 65.03* |
| - F2: compound head match | 65.72 |
| - F3: GermaNet | 65.32** |
| - F4: Distributional information | 65.76 |
| - F5: Animacy: gender match names | 65.59** |
| - F6: Animacy: person match | 65.58** |

Table 5.8.: Performance of IMS HotCoref DE on TüBa-D/Z version 10: ablation experiments

5.2.5. Pre-processing pipeline: running the system on new texts

One of the main problems for people who want to apply a coreference resolver on new text is the pre-processing of the texts. Most systems, like ours, require a few annotation layers such as part-of-speech or constituency parses. In order to achieve the best results, one should use the same tools with which the training data has been processed, so that the annotations are compatible. Together with the specific CoNLL-12 format, this has lead to people having to spend a lot of time setting up their own pipeline or giving up during pre-processing and not using the tool at all.

To simplify the application of IMS HotCoref DE on new texts, we have set up a pipeline that takes plain text as input, performs all the pre-processing steps with the same tools that we have used, creates the right format and runs the coreference resolver as a final step, with default settings and the model pre-trained on the predicted version of TüBa-D/Z v10.⁹

In this section, we describe the required annotations as well as the final format that IMSHotCoref DE takes as input.

Required annotations The system requires preprocessed text with the following annotations in CoNLL-12 format: POS tags, lemmata, constituency parse bits, number

⁹The pre-processing pipeline can be found here:

<https://github.com/InaRoesiger/conversion2conll12>

and gender information and (optionally) named entities. The mention extraction module, the part in the resolver that chooses the markables which we want to resolve in a later step, is based on the constituency parse bits and POS tags. It can be specified which POS tags and which non-terminal categories should be extracted. Per default, noun phrases, named entities and personal, possessive, demonstrative, reflexive and relative pronouns, as well as a set of named entity labels, are extracted. Note that most parsers for German do not annotate NPs inside PPs, i.e. they are flat, so these need to be inserted before running the tool.

Pre-trained models There are two pre-trained models available: one trained on the gold annotations (this one is preferable if you can find a way to create similar annotations to the TüBa gold annotations for your own texts.). We have also uploaded a model trained on predicted annotations: We used the Berkeley parser (Petrov et al., 2006) (out of the box, standard models trained on Tiger) to create the parses, the Stanford NER system for German (Faruqui and Padó, 2010) to find named entities and `mate`¹⁰ (Bohnet and Nivre, 2012) to lemmatise, tag part-of-speech and produce the morphological information.¹¹

Format The tool takes input in CoNLL-12 format. The CoNLL-12 format is a standardised, tab-separated format in a one-word-per-line setup. Table 5.9 shows the information contained in the respective columns.

| Column | Content |
|--------|---------------------------------------|
| 1 | docname |
| 2 | part number |
| 3 | word number in sentence |
| 4 | word form |
| 5 | POS tag |
| 6 | parse bit |
| 7 | lemma |
| 8 | number information: pl or sg |
| 9 | gender information: fem, masc or neut |
| 10 | named entity (optional) |
| 11 | coref information |

Table 5.9.: CoNLL-12 format overview: tab-separated columns and content

¹⁰www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools.html

¹¹Two example documents for the annotations are provided on the webpage.

Annotating coreference in new texts This section explains how to use the pre-trained models to annotate coreference in new documents if you do not want to use the standard pipeline or want to play around with a few of the parameters. A detailed manual on how to train a model is contained in the webpage documentation.

- Download the tool, the model and the manual from the webpage;
- Pre-process your texts so that you have all the necessary annotation layers;
 - make sure that the parse bits have NPs annotated inside of PPs;
 - the parse bits should be comparable to those in the example document: either the gold ones or the ones created by the Berkeley parser;
- Get your texts into the right format: see example document;
- Specify the markables you want to extract;
- Specify the additional information: you can include distributional information, compound splits, etc. for your own texts. Details on the single formats are contained in the manual;
- Specify the features (you can play around with this or just use the default features);
- Training and testing commands can be found in the manual;
- If you have plain text and want to use the tool with default settings, simply apply the pipeline script.

5.2.6. Application on DIRNDL

So far, the experiments have been conducted with the TüBa-D/Z corpus, as it is the benchmark dataset for coreference resolution in German. It is also by far the largest corpus, a fact of which data-driven systems like ours benefit. However, there are also other corpora, such as the DIRNDL corpus, which could be of interest for studies on coreference resolution as DIRNDL is of a different text type, spoken radio news, and was for example also manually labelled with prosodic information. To study the interaction between coreference and prosody, as we plan to do in Section 7, we need a system that is applicable to DIRNDL.

The system presented above, pre-trained on TüBa-D/Z, yields a CoNLL score of 37.04 on the DIRNDL test set with predicted annotations. One issue here is that the predicted

annotations of TüBa-D/Z and DIRNDL are not completely compatible. Hence, the learned features are not as effective as they are on TüBa-D/Z. This comparatively low score also confirms the assumption that the performance of a system trained on written text drops when applied to spoken text. The drop in performance can also be explained by the slightly different domains (newspaper text and radio news).

However, the DIRNDL corpus is big enough to train a model on the concatenation of the training and development set, which is why we decided to train a model based on DIRNDL. We first check whether we should use different markable types for DIRNDL.

Mention extraction for DIRNDL

As DIRNDL was annotated according to the RefLex guidelines (Baumann and Riester, 2012), it has different mentions than TüBa-D/Z, for example no possessive pronouns and no relative pronouns. The most important difference is that PPs are annotated instead of NPs. This is to include cases where the determiner and the preposition are merged into one word, such as in

(14) am Bahnhof = an dem Bahnhof (*at the station*)

To deal with this, we insert NPs into PPs, as described in Section 5.2.2.

As can be seen in Table 5.10, the recall with the best performing markables is about 85.6%, which is slightly higher than the 78% achieved for TüBa-D/Z.

| Tag | Description | Recall |
|-----------|---------------------------------|--------|
| NT-NP | nominal phrases | 35.2 |
| +T-PPER | personal pronouns | 40.8 |
| +T-PPOSAT | attributive possessive pronouns | 40.8 |
| +T-PWS | interrogative pronouns | 40.8 |
| +T-PDS | demonstrative pronouns | 42.7 |
| +T-NE | named entities | 49.9 |
| +T-PRF | reflexive pronouns | 55.5 |
| +NT-PP | PPs | 75.4 |
| +T-PROAV | pronominal adverbs | 78.7 |
| +NT-CNP | conjunctive NPs | 79.8 |
| +T-ADV | adverbs | 80.3 |
| +NT-PN | proper NPs | 82.0 |

Table 5.10.: Markable extraction for the DIRNDL corpus

5. Coreference resolution

In DIRNDL, abstract anaphors can have a VP or clausal antecedent, such as in Example (15), taken from the DIRNDL corpus. These cannot be captured by a system based on nominal antecedents.

- (15) DE: Der niedrigen Geburtenrate sei durch mehr Krippenplätze nicht beizukommen, meinte der Kardinal. **Dies** belege die Situation in Ostdeutschland, wo das Betreuungsangebot besonders hoch sei, die Geburtenrate aber besonders niedrig.
- EN: You cannot overcome low birth rates with more places in day nurseries, said the cardinal. **This** is proven by the situation in East Germany ...

Another issue is that some NPs that have been annotated with coreference do not have a PP or NP label. This is due to errors in the automatic pre-processing and has to be accepted as part of the automatic setting.

Feature engineering We repeated our process of feature selection, as explained above, for DIRNDL. The result is a list of features that slightly deviates from the list of features used for TüBa-D/Z.

Performance on DIRNDL

As can be seen in Table 5.11, the system trained on DIRNDL achieves a CoNLL score of 46.11, which is comparable to the score reported on the predicted version of Tüba-D/Z v10 (48.61). As we will show in Section 7, it can be further improved by including prosodic features.

| MUC | BCUBE | CEAFM | CEAFE | BLANC | CoNLL |
|-------|-------|-------|-------|-------|--------------|
| 44.93 | 45.13 | 50.94 | 48.27 | 35.14 | 46.11 |

Table 5.11.: Performance of IMS HotCoref DE on DIRNDL, using predicted annotations

5.3. Conclusion

As mentioned in the beginning of this section, there are many well-performing and openly available coreference resolvers for English. For German, there is the rule-based CorZu, as well as a number of mostly learning-based systems from the SemEval shared task

2010, whose performance on the benchmark dataset TüBa-D/Z is worse than that of CorZu. Most of these systems, for example SUCRE, are also not publically available. Therefore, we have adapted the learning-based system IMS HotCoref, which at the time of the experiments achieved state-of-the-art results for English on the benchmark dataset OntoNotes, to German by integrating linguistic features designed to address specificities of German, such as for example gender agreement. In ablation experiments we have shown that computing all features based on the lemma rather than the word forms had the biggest influence on the performance on the system. The adapted system achieves state-of-the-art results on TüBa-D/Z. We have also shown that the system also works well when trained on other data, e.g. on the DIRNDL corpus, which is of a different domain than TüBa-D/Z (radio news instead of newspaper). We have described the steps involved in using the system on unseen text and presented some of the parameters with which the system can be optimised.

IMS HotCoref DE is used in one of our linguistic validation experiments, where we integrate prosodic information into coreference resolution. In the next chapter, we will continue with the creation of bridging resolution tools for English and German.

6. Bridging resolution

Research Question 3: Tool creation

Are there openly available tools aiming at providing automatic annotations on unseen text? If not, can we create tool resources to fill the research gap?

An overview of work in the area of bridging resolution has already been presented in Section 3.2. Of all the previous approaches for bridging anaphora detection, bridging anaphora resolution or full bridging resolution, no system has been made publicly available. The latter would be necessary, however, to assess the generalisability of the approaches, or in other words to check how well the suggested approaches work on other data or domains than the ones on which they were designed, without much reimplementing work. Open source systems can also easily be extended, instead of implementing entirely new systems.

In this chapter, we describe the reimplementing of the state-of-the-art system for full bridging resolution (Hou et al., 2014), which will serve as a basis to assess the tool's performance on other corpora and domains, including our newly created newspaper corpus BASHI and our scientific corpus SciCorp as well as a shared task submission for the first shared task on bridging at CRAC 2018. The tool is openly available.¹ Besides reimplementing this tool for English, we will also describe an adaptation to German. We are thus making a first step towards filling the research gap of non-existing openly available bridging tools for English and German. The contributions in this step (tool creation) are shown in Figure 6.1. Parts of this research have been published in Rösiger (2018b), Poesio et al. (2018) and Rösiger et al. (2018b).

6.1. A rule-based bridging system for English

This section describes the reimplementing and adaptation of a rule-based bridging resolver proposed by Hou et al. (2014). As this system was never made publicly available,

¹<https://github.com/InaRoesiger/BridgingSystem>

6. Bridging resolution

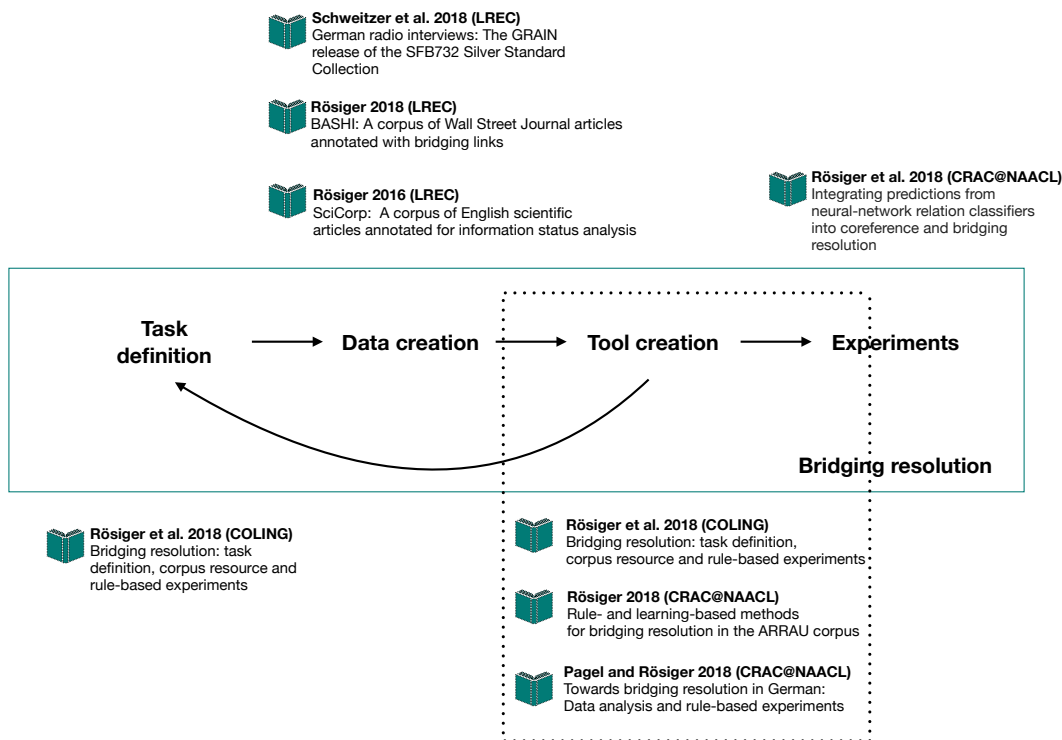


Figure 6.1.: Contribution and workflow pipeline for bridging: tool creation

we think it is a valuable effort to reimplement the system and provide it as a baseline that can then be further adapted to other domains or enriched with new knowledge sources, as we will do in the linguistic validation experiments step, where we will use the system to assess how semantic relations can help bridging resolution. We describe the original system in the next section, together with details on where the reimplementations differ from the original system. We also include a couple of additional experiments, where we compare the use of predicted and gold markables and investigate the effect of coreference information. We report that filtering out gold or even just predicted coreferent anaphors before bridging resolution significantly helps improve bridging resolution.

Experimental setup The system was designed for the corpus ISNotes (Markert et al., 2012). Hou et al. (2014) split the corpus into a development (10 documents) and test set (40 documents). The rules were optimised on the development set and the performance of the system reported on the test set. Unfortunately, the concrete development/test

split is not specified. We report numbers for our own test-development-split² as well as for the whole corpus.

6.1.1. Reimplementation

While most work on bridging resolution has focused on one of the subtasks, i.e. either identifying an expression as a bridging anaphor or finding an antecedent for one bridging anaphor, Hou et al. (2014) tackled the task of full bridging resolution. She designed eight hand-crafted rules which are based on linguistic intuitions about bridging. Most of the rules are very specific, aiming at high precision, while two rules are designed to capture more general bridging cases, thus increasing the recall.

The reimplementation comprises all three components of the original paper: preprocessing, rule adaptation and postprocessing. During preprocessing, markables are extracted, which are then passed on to the eight rules which predict bridging anaphor and antecedent pairs. In the postprocessing step, the rules are applied in order of descending precision.

Preprocessing We extract NPs as our predicted markables. We also extract the markables of the information status annotation as our set of gold markables. These form the initial set of anaphors and antecedents.

In the predicted setting, by extracting NPs only, we miss 13 out of the 663 gold anaphors and 79 out of the 663 antecedents. An analysis of the missing markables yielded the following missing candidates:

- Anaphors:
 - Constituents with the tag NML (embedded modifying noun phrases, left-branching), embedded in NPs:

(1) Crude oil prices have exploded during the last few weeks.

The market ...

(NP (NML (NN crude) (NN oil)) (NNS prices))

- Antecedents:
 - Pronouns: *we, our, his, her, ...*

²The 10 dev docs are: wsj1101, wsj1123, wsj1094, wsj1100, wsj1121, wsj1367, wsj1428, wsj1200, wsj1423, wsj1353.

6. Bridging resolution

- Other POS tags: *Anti-abortion (JJ)*, *AIDS (no label)*
the reason for these annotations is that annotators were not limited to NPs or other pre-defined categories when determining the antecedent.
- Verbal antecedents or clauses: we only focus on nominal antecedents.

As our system is designed for nominal antecedents only, we cannot help losing verbs, adjectives or clauses that are labelled as the non-nominal antecedent. Other nominal antecedents, like the ones of the NML category, should, however, be extracted. Thus, NML is added as a markable. Now we find all but one out of the 663 anaphors, but 74 antecedents are still not found. A few pronouns have been tagged as bridging antecedents, so we add them as potential antecedents. By adding personal and possessive pronouns as antecedent candidates, we can reduce the number of missed antecedents to 65. The remainder of the antecedents is non-nominal, i.e. either verbal, clausal or of another non-nominal category.

Certain NPs are removed from the list of potential anaphors in order not to suggest too many candidates, namely NPs which have a complex syntactic structure (i.e. that have embedded mentions) and NPs which have comparative markers (this is due to the exclusion of comparative anaphora from the category bridging in the corpus ISNotes).³ Contrary to Hou et al. (2014), we filter out pronouns as anaphor candidates as they should, in principle, always be coreference anaphors rather than bridging anaphors. We follow Hou et al. (2014)’s suggestion to exclude NPs whose head appeared before in the document, as these cases are typically involved in coreference chains. We also experiment with filtering out predicted and gold coreference anaphors before applying the rules.

After filtering out mentions that have embedded mentions (complex NPs) and NPs with clear comparative markers, 92 of 663 anaphors are no longer available as candidates. After filtering out definite NPs that have the same head as a previous NP, 128 of the 663 gold anaphors are not included anymore in the list of candidates.

To sum up, after the filtering step we have lost 65 of the antecedents and 128 of the anaphors. This means that with the current filtering strategy the best possible recall is around 70%.

Rules

Each rule is applied separately to the list of extracted markables and proposes pairs of bridging anaphors and antecedents. Table 6.1 gives an overview of the rules implemented.

³The list taken from Hou et al. (2014) is: similar, another, such, other, related, different, additional, comparable, same, further, extra.

| Rule | Example | Anaphor | Antecedent search | Window |
|------|---|-------------------------------|--------------------------------------|--------|
| 1 | A white woman's house ← The basement | building part | semantic connectivity | 2 |
| 2 | She ← Husband David Miller | relative | closest person NP | 2 |
| 3 | The UK ← The prime minister | GPE job title | most frequent GEO entity | - |
| 4 | IBM ← Chairman Baker | professional role | most frequent ORG NP | 4 |
| 5 | The firms ← Seventeen percent | percentage expression | modifying expression | 2 |
| 6 | Several problems ← One | number/indefinite pronoun | closest plural, subject/object NP | 2 |
| 7 | Damaged buildings ← Residents | head of modification | modifying expression | - |
| 8 | A conference ← Participants | arg-taking noun, subj pos. | semantic connectivity | 2 |

Table 6.1.: Overview of rules in Hou et al. (2014)

Each rule can have its own parameters on for example the allowed distance between the anaphor and the antecedent. Two measures are computed independently of the actual bridging resolver and are needed as input for several rules: semantic connectivity and the argument-taking ratio.

Computing the semantic connectivity The semantic connectivity goes back to the “NP of NP” pattern in Poesio et al. (2004) and was extended to a more general preposition pattern in Hou et al. (2014). The semantic connectivity between two words can be approximated by the number of times two words occur in a “noun (N) preposition (PREP) noun” pattern in a big corpus. This means that two nouns like *window* and *room* have a high semantic connectivity because they often occur as *windows in the room*, whereas other nouns do not appear often in such a construction and are therefore not highly semantically connected. The Dunning root log-likelihood ratio (Dunning, 1993) is computed as a measure of the strength of association. To compute the measure, we need to calculate the counts shown in Table 6.2. For an example computation and more details, please refer to Hou (2016b).

In contrast to Hou et al. (2014), we do not limit prepositional patterns to the three most common prepositions for a noun but count every N PREP N pattern. Also, we allow for optional adjectives and determiners in the N PREP N pattern. Following Hou

6. Bridging resolution

| | Noun 1 | Not noun 1 | Total |
|------------|--------|------------|-------|
| Noun 2 | a | b | a+b |
| Not noun 2 | c | d | c+d |
| Total | a+c | b+d | |

Table 6.2.: Contingency table for the Noun1 + preposition + Noun2 pattern

et al. (2014), we take the GigaWord corpus (Parker et al., 2011) as a big corpus (1200 M tokens) as the basis for the computation of the scores. The result is a list with noun pairs and their respective connectivity score, in a tabular text format. The scores have not been normalised (to values between 0 and 1) because we are only using them to find the pair with the highest score, not some relative score or threshold.

| Noun pair | Score |
|---------------------|-------|
| wife - husband | 28.6 |
| husband - wife | 30.7 |
| husband - daughter | 14.1 |
| husband - carpet | 2.8 |
| husband - Wednesday | -10.3 |

Table 6.3.: Exemplary semantic connectivity scores

One problem with the Wall Street Journal is that the corpus is not lemmatised. Some nouns are mapped onto gold senses, and those are always lemmatised, which means that we can copy the lemmatisation from these annotations. For all other nouns, this is not available. Our solution is to look for senses of nouns, and use these where possible (e.g. *child* for *children*). For nouns which do not have a sense mapping, we save all word forms and their lemmatisations as they were tagged in the GigaWord corpus. We use these word form - lemma pairs also when applying the rules.

- (2) children → child
- (3) husband’s → husband

If we do not find a mapping and it is not contained in our scores, we use a simple approximation for default pluralisation: we add or remove an “s” to/from the word to see whether scores exist for this slightly modified form.

Computing the argument-taking ratio The argument-taking ratio of a mention’s head reflects how likely a noun is to take arguments (Hou et al., 2014). This can be used for bridging resolution, as we assume that the bridging anaphor is lacking an implicit argument in the form of the antecedent. If it has a low argument-taking ratio, then the likeliness of an expression to be a bridging anaphor is also low. For example, the lemma *child* is often used without arguments, when we are generically speaking about *children*. *Brainchild*, however, seems to be an expression that is exclusively used with an argument, e.g. in *the brainchild of . . .* .

| Noun | Score |
|------------|-------|
| child | 0.21 |
| childhood | 0.83 |
| brainchild | 1 |
| husband | 0.9 |

Table 6.4.: Exemplary argument-taking ratios

The argument-taking ratio is calculated by taking the head frequency in the NomBank annotation divided by the head’s total frequency in the WSJ corpus. The argument-taking scores are normalised to values between 0 and 1. Again, we perform the techniques described above to deal with lemmatisation.

Rule 1: building part NPs Rule 1, called building part NPs, is designed to capture cases of meronymy that have to do with buildings, as in the following example.

- (4) At age eight, Josephine Baker was sent by her mother to a white womans house to do chores in exchange for meals and a place to sleep a place in **the basement** with coal.

For this, a list of 45 nouns which specify building parts (e.g. *windows*, *basement*) is taken from the General Inquirer lexicon (Stone et al., 1966). For an anaphor to be added to the list of bridging anaphors proposed by this rule, the head form has to be on the building list and may not contain any nominal pre-modification. Then for each potential anaphor, the NP with the strongest semantic connectivity is chosen as the antecedent within the same sentence and the previous two sentences.

We additionally exclude NPs containing a PP, as in Example (5), and exclude NPs in the idiom *leaves room for* as they are metaphorical uses that do not have to do with actual building parts.

6. Bridging resolution

(5) the windows in the room

Rule 2: relative person NP Rule 2 is meant to capture bridging relations between a relative (*husband*) and its antecedent (*she, the wife*). For this, a list of 110 nouns is extracted from WordNet which contains relatives, e.g. *husband, cousin or granddaughter*. One issue is that some of these nouns are often used generically (e.g. *children*). To overcome this, the argument-taking ratio, a measure for the likelihood of a noun to take an argument, is computed.

According to Hou et al. (2014), for an anaphor to be added to the list of bridging anaphors, the anaphor’s head must appear on the relative person list and the argument-taking ratio of its head must be bigger than 0.5 and must not contain nominal or adjectival premodification. As the antecedent, the closest non-relative person NP among all mentions preceding the anaphor from the same sentence as well as from the previous two sentences is chosen.

(6) She ... **Husband David Miller**

In our reimplementation, we first created a relative list by listing all sorts of relatives that came to our mind. The list contains 102 entries. The anaphor must have an argument-taking ratio larger than 0.5 and must not be modified by an adjective or a noun and must not contain an embedded PP or be followed by a PP. As the antecedent, we choose the closest proper name that is not an organisation (does not have **ORG** in the named entity column), named entity tagged person (**PER**) or personal pronoun except those with lemma *they* or *you*.

Rule 3: GPE job title NPs This rule aims at job titles that revolve around a geopolitical entity. Hou et al. (2014) states that “in news articles, it is common that a globally salient geopolitical entity (hence GPE, e.g., *Japan or the U.S.*) is introduced in the beginning, then later a related job title NP (e.g., *officials or the prime minister*) is used directly without referring to this GPE explicitly”.

(7) USA ... **the president**

Hou et al. (2014) set up a list of 12 job titles (*president, governor, minister, etc.*). The anaphor is added to the list of potential anaphors if it does not contain a country adjective such as *US*. As the antecedent, the most frequent GPE is chosen. In case of a tie, the closest NP is chosen.

We take the job list from Hou (2016b)⁴, but leave out *president* because in most cases, it is a central notion in the text and typically present in a coreference chain and thus not a bridging anaphor. Hou et al. (2014) stated that the anaphor must not contain a country adjective (e.g. *the German president*). We additionally remove mentions containing an embedded PP or followed by a PP, or an organisation (**ORG** in the named entity column). The antecedent is chosen to be the geopolitical entity with the highest frequency in the document.

Rule 4: role NPs While Rule 3 is designed to capture rather specific cases of bridging revolving around GPEs, Rule 4 aims at finding more general cases of bridging where the job titles are not restricted to GPEs, but to all organisations. For this, a list of 100 nouns which specify professional roles is extracted from WordNet (*chairman, president, professor*). For the mention to be considered a potential anaphor candidate, the anaphor’s head must be on the role list and the most salient proper name NP which stands for an organisation is chosen as the antecedent. Most salient here means most frequent in the document before the anaphor. In case of a tie, the closest NP should be chosen.

(8) IBM ... **Chairman Baker**

Our list of professional job roles (e.g. *doctor, CEO, chairman, employee, etc.*) contains 132 nouns. The head word of the anaphor must be on this list and the NP must not contain a country adjective, a PP, a proper name or an indefinite article. We choose the most frequent organisation within the same sentence or the previous two sentences as the antecedent.

Rule 5: percentage NPs Rule 5 is a rather specific rule, designed to address percentage expressions. If the anaphor is a percentage expression, the antecedent is predicted to be the closest NP which modifies another percentage NP via the preposition *of* among all mentions occurring in the same or up to two sentences prior.

(9) 22% of the firms said employees or owners had been robbed on their way to or from work. **Seventeen percent** reported their customers being robbed.

⁴president, official, minister, governor, senator, mayor, chancellor, ambassador, autocrat, premier, commissioner, dictator, secretary

6. Bridging resolution

In our version, the head form of the antecedent must be either *percent* or %, must not be modified by the preposition *of* itself, must not be at the end of the sentence and must be in subject position. As we do not have grammatical roles in our version of the corpus, we use the approximation that a subject is followed by a verb. The antecedent must modify a percentage expression with the preposition *of* and must be in the same or in the previous two sentences. We choose the closest NP that matches these criteria.

Rule 6: other set members Rule 6 aims at finding bridging pairs that are labelled as **set** bridging in ISNotes. The idea behind this rule is that numbers and indefinite pronouns are good indicators for bridging anaphors (if they are contained in the corpus, of course, which is not the case for all corpora). In order for the NP to be considered a bridging anaphor candidate, it must be a number expression (e.g. *one*) or an indefinite pronoun (*some*) and in subject position. The antecedent is chosen to be the closest NP among all plural subject mentions preceding the potential anaphor. If non-existent, object mentions are checked.

- (10) This creates several problems. **One** is that there are not enough police to satisfy small businesses.
- (11) Reds and yellows went about their business with a kind of measured grimness. **Some** frantically dumped belongings into pillowcases.

We have compiled a list of indefinite pronouns and number expressions⁵. The anaphor must be on this list, and in subject position. We also define a number of unsuited verbal expressions⁶ as these typically occur in contexts where the subject is used generically, e.g. in Example (12).

- (12) One has to wonder ...

The antecedent is predicted to be the closest subject NP (again, we use our approximation) in the same sentence as well as in the previous two sentences. If we do not find one in subject position, we look for the closest object NP (defined as following a verb).

Rule 7: argument-taking NPs Rule 7 is a more general rule to find bridging pairs and is based on an observation in Laparra and Rigau (2013) who found that different instances of the same predicate in a document likely maintain the same argument fillers.

⁵one, some, none, many, most, two, three, four, five, ten, dozen, hundred, million, first, second, third

⁶feel, claim, fear, see, think, proclaim, may, might, argue

A common NP is considered an anaphor if the argument-taking ratio is greater than 0.5 and if it has got no nominal or adjectival premodification and does not come with determiners. The antecedent is chosen as follows: we collect syntactic modifications and arguments (nominal premodification, possessive as well as PP modification or PP arguments) for the anaphor's head lemma form. All realisations are potential antecedent candidates. As the antecedent, we choose the most recent NP from the candidate list.

- (13) Out on the streets, *some residents of badly damaged buildings* were allowed a 15 minute scavenger hunt through their possessions. ... After being inspected, buildings with substantial damage were color - coded. Green allowed **residents** to re-enter; red allowed **residents** one last entry to gather everything they could within 15 minutes.

We search for common NPs by extracting all anaphors containing the POS tag “NN” or “NNS”. In our reimplementation, the anaphor must not be modified by any noun or adjective and must not contain an embedded PP or be followed by a PP. The antecedent must be in the same sentence or in the two previous sentences and must be the closest similar modification or argument found in the document, as described above.

Rule 8: argument-taking NPs II Rule 8 is even more general than Rule 7, in that it does not only search for similar contexts in the document but generally looks for semantically related words. It uses the concepts of argument-taking and semantic connectivity to determine semantically related words. The argument-taking ratio of an anaphor must be greater than 0.5, the anaphor cannot have nominal or adjectival premodification and it must be in subject position. As the antecedent, the mention with the highest semantic connectivity is chosen.

- (14) Initial steps were taken at Polands first international environmental conference which I attended last month. [...] While Polish data have been freely available since 1980, it was no accident that **participants** urged the free flow of information.

We additionally exclude mentions as anaphors that are bare singulars, those containing *some*, a demonstrative pronoun, negation or words on the relative list (cf. Rule 2).

Post-processing Each rule proposes a number of bridging pairs, independently of the decision of other rules. We order the rules according to their precision. In case of conflicts, the rule with the higher precision is applied.

6.1.2. Performance

In this section, we compare the performance of the original system with our reimplementation.

Rule performance

Table 6.5 shows the performance of the individual rules. The numbers in brackets are taken from Hou (2016b). The precision with respect to the anaphor tells us how many of the proposed bridging pairs contain gold bridging anaphors. The precision wrt the pair stands for how many of the pairs, i.e. both anaphor and antecedent, are correct gold pairs. Of course, the precision of the pair is always lower than the precision of the anaphor. If we chose the right antecedent in all cases, the precisions would be the same. The firing rate tells us how often the rule was applied. The number in brackets in the column Rule tells us the respective rank of the rule when ordering the rules according to their precision.

As can be seen in the table, the numbers of Hou (2016b) are not always the same as ours. We achieve a higher performance for some of the rules and a lower performance for others. On average, however, the performance is comparable.

| Rule | P of anaphor | P of pair | Firing Rate |
|----------------------------------|----------------|----------------|--------------|
| Rule1 [2] building part NPs | 63.6% (75.0) | 54.5% (50.0) | 9.2% (6.1) |
| Rule2 [5] relative person NPs | 55.5% (69.2) | 44.4% (46.2) | 7.5% (6.1) |
| Rule3 [6] GPE job title NPs | 76.2% (52.6) | 61.9% (44.7) | 17.5% (19.4) |
| Rule4 [7] role NPs | 77.7% (61.7) | 59.3% (32.1) | 22.5% (28.6) |
| Rule5 [1] percentage NPs | 100.0% (100.0) | 100.0% (100.0) | 4.2% (2.6) |
| Rule6 [3] other set member NPs | 71.4% (66.7) | 50.0% (46.7) | 11.7 % (7.8) |
| Rule7 [4] argument-taking NPs I | 72.7% (53.8) | 54.5% (46.4) | 9.2 % (6.1) |
| Rule8 [8] argument-taking NPs II | 63.6% (64.5) | 36.3% (25.0) | 18.3% (25.5) |

Table 6.5.: A bridging system for English: performance of the individual rules, their precision as well as their firing rate

Overall performance

Hou et al. (2014) states a precision of 61.7%, a recall of 18.3% and an F1 score of 28.2% for anaphor detection and a precision of 42.9%, a recall of 11.9% and an F1 score of 18.6% for full bridging resolution. In both settings, they use gold markables but no coreference information. Table 6.6 contains the scores of the reimplementation for the test and the whole corpus when using gold or predicted markables. As mentioned above, we have defined a different test-development-split, which is why the results are not directly comparable. In general, however, we think that our reimplementation achieves comparable results, as our rules also achieve similar precision values and firing rates as in Hou (2016b).

As we have simply reimplemented the system from the original paper without any hand-tuning on the development set, we also report the numbers on the whole ISNotes corpus. Here, our reimplementation yields 65.9% precision, 14.1% recall and 23.2% F1 score for the task of anaphor recognition and a precision of 49.6%, a recall of 10.6% recall and an F1 score of 17.4% for full bridging resolution. Compared to the original numbers in Hou et al. (2014), we achieve higher precision, but lower recall, resulting in an overall slightly lower F1 measure. Note that we do not carry out significance tests here, as the experiments were not performed on the same datasets.

| Setting | Corpus | Anaphor recognition | | | Full bridging | | |
|---|--------------|---------------------|------|------|---------------|------|------|
| | | P | R | F1 | P | R | F1 |
| Hou (2014), gold mark. | test set | 61.7 | 18.3 | 28.2 | 42.9 | 11.9 | 18.6 |
| Reimplementation with gold markables | | | | | | | |
| | test set | 73.4 | 12.6 | 21.6 | 60.6 | 10.4 | 17.8 |
| | whole corpus | 65.9 | 14.1 | 23.2 | 49.6 | 10.6 | 17.4 |
| Reimplementation with predicted markables | | | | | | | |
| | test set | 69.3 | 12.2 | 20.7 | 57.7 | 10.1 | 17.2 |
| | whole corpus | 65.2 | 13.6 | 22.5 | 49.2 | 10.3 | 17.0 |
| Filtering out coreferent anaphors, with gold markables | | | | | | | |
| No coreference | whole corpus | 65.9 | 14.1 | 23.2 | 49.6 | 10.6 | 17.4 |
| Predicted coreference | whole corpus | 79.6 | 14.1 | 23.9 | 59.8 | 10.6 | 18.0 |
| Gold coreference | whole corpus | 79.6 | 14.1 | 23.9 | 59.8 | 10.6 | 18.0 |

Table 6.6.: Performance of the reimplementation of Hou et al. (2014), with different settings

Coreference information

As bridging anaphors are difficult to distinguish from coreference anaphors, we think it may be beneficial for the precision of our system to filter out coreference anaphors before applying the bridging system. We experiment with three settings: (i) no coreference information, (ii) predicted coreference information and (iii) gold annotated coreference information. For predicted coreference, we applied the IMS HotCoref system (Björkelund and Kuhn, 2014) with its default settings on the ISNotes corpus.⁷ We report the change in performance on the whole corpus, as there was no optimisation involved in the filtering of the coreference anaphors. In Table 6.6, it can be seen that both predicted and gold coreference significantly improve the precision of the system.⁸ Surprisingly, there is no difference between gold and predicted coreference. The same effect can be observed with predicted mentions. We also experimented with coreference information in the final bridging system (as described in Section 8), where the observed effect is much stronger.

| Setting | Precision | Recall | F1 |
|-----------------|-----------|--------|-------------|
| No coref | 49.6 | 10.6 | 17.4 |
| Predicted coref | 59.8 | 10.6 | 18.0 |
| Gold coref | 59.8 | 10.6 | 18.0 |

Table 6.7.: Performance of the bridging system with different coreference information, gold mention setting

| Setting | Precision | Recall | F1 |
|-----------------|-----------|--------|-------------|
| No coref | 49.2 | 10.3 | 17.0 |
| Predicted coref | 55.1 | 10.3 | 17.3 |
| Predicted coref | 55.1 | 10.3 | 17.3 |

Table 6.8.: Performance of the bridging system with different coreference information, predicted mention setting

⁷We made sure to exclude the ISNotes part of OntoNotes from the training data for the coreference system, of course.

⁸Again, we use the Wilcoxon signed rank test to compute significance, at the $p=0.01$ level. In this case, all comparisons were significant, which is why they are not marked. Boldface indicates the overall best results.

| Corpus | Domain | Anaphor recognition | | | Full bridging | | |
|----------------------|------------|---------------------|--------|------|---------------|--------|------|
| | | Prec. | Recall | F1 | Prec. | Recall | F1 |
| ISNotes (gold mark.) | news | 65.9 | 14.1 | 23.2 | 49.6 | 10.6 | 17.4 |
| ISNotes (pred mark.) | news | 65.2 | 13.6 | 22.5 | 49.2 | 10.3 | 17.0 |
| BASHI (pred mark.) | news | 49.4 | 20.2 | 28.7 | 24.3 | 10.0 | 14.1 |
| SciCorp (pred mark.) | scientific | 17.7 | 0.9 | 8.1 | 3.2 | 0.9 | 1.5 |

Table 6.9.: Performance of the rule-based method on other corpora. We use predicted mentions for BASHI and SciCorp as they do not contain gold markables.

6.1.3. Generalisability of the approach

Recent work on bridging resolution has so far been based on the corpus ISNotes (Markert et al., 2012), as this was the only corpus available with unrestricted bridging annotation. Hou et al. (2014)’s rule-based system currently achieves state-of-the-art performance on this corpus, as learning-based approaches suffer from the lack of available training data. To test the generalisability of the approach by Hou et al. (2014), we apply our reimplementations to the newly annotated corpora (as presented in Section 4.3).

Experimental setup

BASHI BASHI⁹ is a newspaper corpus that we annotated with bridging links according to guidelines compatible with those of the ISNotes corpus. The corpus can be used to assess the generalisability on in-domain corpora, as ISNotes and BASHI are of the same domain. As we simply apply our systems to this data, we report performance on the whole corpus.

SciCorp SciCorp¹⁰ is a corpus of a different domain, scientific text, that can be used to assess how well the system generalises to a completely different domain. Again, we report numbers on the whole corpus.

BASHI (in-domain) results

We first apply our reimplementations to a corpus of the exact same domain as ISNotes, BASHI. As can be seen in Table 6.9, the F1 score for anaphor recognition is 28.7, which is comparable with the score on ISNotes, although we observe a much lower precision on

⁹<http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/bashi.html>

¹⁰<http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/scicorp.html>

6. Bridging resolution

BASHI. Lower precision is also the reason for the overall lower score on BASHI for full bridging resolution, which means that the performance for anaphor detection is about the same, while the performance of finding the correct antecedent is worse. Still, the system performs relatively well on this data.

SciCorp (out-of-domain) results

SciCorp is an out-of-domain corpus. When applying our system, we observe that it really does not generalise well to completely different domains, as the F1 score for full bridging resolution drops to 1.46. SciCorp differs from BASHI and ISNotes with respect to the definiteness criterion: all bridging anaphors are definite. Of course, rules designed for indefinite anaphors cannot work. While we expected some of the rules designed for news text to perform poorly (e.g. building parts, relatives, job titles etc.), the rules designed to find more general cases of bridging also do not seem to predict a lot of pairs in this domain. The reason for this might lie in the coverage of the semantic connectivity and argument-taking ratio, which are applied in these general rules: only 32% of the nouns in SciCorp are represented in the argument-taking-ratio lists, and only 3.9% of the noun pairs are contained in the semantic connectivity scores. Adding some in-domain text (e.g. large PubMed/ACL corpora) to the general corpora used to create these resources would be necessary to improve performance for the general rules of the system to work. We are positive that doing some form of domain adaptation, i.e. designing specific rules for scientific text and combining them with the improved general rules, would lead to better results.

6.2. CRAC 2018: first shared task on bridging resolution

The workshop Computational models of Reference, Anaphora and Coreference (CRAC) 2018 featured a shared task on bridging resolution, based on the ARRAU dataset. This is another opportunity to test the generalisability of our reimplementation, so we also apply our system to this dataset. As these experiments involved a shared task submission, we will provide a more detailed analysis for the ARRAU corpus.

6.2.1. The ARRAU corpus

The second release of the ARRAU corpus, first published in Poesio and Artstein (2008), was used as the data basis for the shared task. It is a multi-domain corpus that aims at “providing much needed data for the next generation of coreference/anaphora resolution systems” (Uryupina et al., 2018). The current version of the dataset contains 350k tokens and 5,512 bridging anaphors. The shared task data comprises text from three domains: RST (newspaper), TRAINS (dialogues) and the PEAR stories (narrative text). Following earlier attempts on the reliable annotation of bridging (Poesio, 2004), where it became evident that better annotation quality could be achieved by limiting the annotation to the three relations `subset`, `element` and `poss`, most of the bridging relations in ARRAU are of these types, as shown in Table 6.11. Additionally, comparative anaphora are included and marked as `other`, and bridging cases which do not fit the pre-defined relations, but are obvious cases of bridging, are marked with the relation `undersp-rel`.

The newest release of the ARRAU corpus (Uryupina et al., 2018) was used as data for the first shared task on bridging at CRAC 2018. The data was obtained from the LDC and consists of training, development and test sets for the three domains newspaper, narrative text and dialogue, with most of the text being news text. As the number of bridging anaphors in the narrative and dialogue part is quite small, the shared task focused on the RST (news) domain, but we also give numbers for the other domains.

| Domain | Number of bridging anaphors |
|--------------|-----------------------------|
| RST | 3777 |
| TRAINS | 710 |
| PEAR stories | 333 |
| Total | 5512 |

Table 6.10.: Number of bridging anaphors in the single domains of the ARRAU corpus

6.2.2. Data preparation

The ARRAU corpus was published in the MMAX format, an XML-based format of different annotation layers. We converted the data into our own, CoNLL-12-style format and used the following annotation layers to extract information:

the word level, to obtain the words, document names and word number, the sentence

6. Bridging resolution

| Relation | Number of bridging relations |
|-----------------|------------------------------|
| Element | 1126 |
| Subset | 1092 |
| Underspecified | 588 |
| Subset-inv | 368 |
| Other | 332 |
| Element-inverse | 152 |
| Poss | 87 |
| Poss-inverse | 25 |
| Other-inverse | 7 |

Table 6.11.: Bridging relations in ARRAU

| S | W | Word | Pos | Coref | Bridging | Markable | Genericity |
|---|----|-------------|-----|-------------|-----------------------|----------------------|-------------|
| 3 | 1 | Plans | nns | (23 | (bridging\$1\$1-23-28 | (m\$18 | 18\$gen-no |
| 3 | 2 | that | wdt | - | - | - | - |
| 3 | 3 | give | vbp | - | - | - | - |
| 3 | 4 | advertisers | nns | (4) | - | (m\$19) | 19\$gen-yes |
| 3 | 5 | discounts | nns | (24 | - | -(m\$20 | 20\$gen-no |
| 3 | 6 | for | in | - | - | - | - |
| 3 | 7 | maintaining | vbg | - | - | - | - |
| 3 | 8 | or | cc | - | - | - | - |
| 3 | 9 | increasing | vbg | - | - | - | - |
| 3 | 10 | ad | nn | (25 (3 | - | (m\$21) (m\$22 | 21\$gen-yes |
| 3 | 11 | spending | nn | 23) 24) 25) | bridging\$1) | m\$18) m\$20) m\$22) | 22\$gen-no |
| 3 | 12 | have | vbp | - | - | - | - |
| 3 | 13 | become | vbn | - | - | - | - |

Table 6.12.: The CoNLL-12-style format used in our bridging experiments

level, to obtain sentence numbers, the part-of-speech level to extract POS tags and the phrase level to extract bridging anaphors, their antecedent, the bridging relation, coreference information, as well as the following attributes of the markables: gender, number, person, category, genericity, grammatical function and head word.

The format is given in Table 6.12, which shows the annotation of bridging anaphors, which are numbered and contain the sentence number as well as the start and end numbers of their antecedents. For example, bridging anaphor number 1 (*plans that give advertisers discounts for maintaining or increasing ad spenders*) has an antecedent which can be found in sentence 1, word 23-28. The markables are also shown, which come with a number of attributes given at the start of the markable. Due to lack of space, we only show the attribute “genericity” in the table.

A couple of special cases of bridging annotations came up during the preparation of the data.

| Domain | Number of bridging anaphors | | |
|--------|-----------------------------|------|-------|
| | Train/dev | Test | Total |
| RST | 2715 | 588 | 3303 |
| TRAINS | 419 | 139 | 558 |
| PEAR | 175 | 128 | 303 |

Table 6.13.: Number of bridging anaphors in the shared task after filtering out problematic cases

- Multiple antecedents:
our data structure only allows one antecedent per anaphor, which is why we cannot handle cases of one anaphor having multiple antecedents.
- Discontinuous markables:

(15) **those in Europe** or Asia seeking foreign stock-exchange.

In this example, the anaphor *those in Europe seeking foreign stock-exchange* was marked as a **subset** bridging case, with *costumers seeking foreign stock-exchange* as its antecedent. As mentioned above in the paragraph on the evaluation of bridging, it is controversial whether annotating parts of NPs as markables is a good annotation strategy. In the ARRAU corpus, discontinuous anaphors and antecedents were allowed. Unfortunately, our system cannot handle discontinuous markables as it takes NPs as its basic markables.

- Bridging antecedents spanning more than one sentences:
as our markable extraction module is based on extracting certain constituency categories, we cannot handle markables spanning more than one sentence.
- Empty antecedents:
some bridging anaphors do not have an annotated antecedent.

After filtering out these cases, the corpus statistics have changed, which are given in Table 6.13.

6.2.3. Evaluation scenarios and metrics

We report the performance of our systems for four different tasks.

6. Bridging resolution

Full bridging resolution This task is about finding bridging anaphors and linking them to an antecedent. Gold bridging anaphors are not given. We use gold markables.

Bridging anaphora resolution (all) This subtask is about finding antecedents for given bridging anaphors. In this setting, we predict an antecedent for every anaphor. This is the official task of the bridging shared task.

Bridging anaphora resolution (partial) This subtask is about finding antecedents for given bridging anaphors, but in this case, we only predict an antecedent if we are relatively sure that this is a bridging pair. This means that we miss a number of bridging pairs, but the precision for the predicted pairs is much higher.

Bridging anaphora detection This subtask is about recognising bridging anaphors (without linking them to an antecedent), again using gold markables.

Data splits We design rules and optimise parameters on the training/development sets of the RST domain, and report performance on the test sets.

6.2.4. Applying the rule-based system to ARRAU

When applying our reimplementaion to the complete RST dataset, the performance drops to an F1 score of 0.3 for the task of full bridging resolution, although both datasets are of the same domain (WSJ articles). We carefully analysed the reasons for the huge difference in performance between ISNotes/BASHI and ARRAU, which both contain Wall Street Journal articles and can thus not be explained with domain effects. To do so, we started with an analysis of the rules and their predicted bridging pairs. Table 6.14 shows the rules and their performance on the RST dataset.

Before discussing the difference between the annotations in ISNotes and ARRAU in the next section, we give examples of some of the pairs as proposed by the respective rules. We also state whether the example was considered wrong or correct according to the ARRAU gold annotations, which do not always reflect our opinion, as we will soon see.

| Rule | Anaphor recognition | | Bridging resolution | |
|----------------------------|---------------------|-------------|---------------------|-------------|
| | Correct pairs | Wrong pairs | Correct pairs | Wrong pairs |
| Rule 1: Building parts | 2 | 28 | 1 | 29 |
| Rule 2: Relatives | 1 | 26 | 0 | 27 |
| Rule 3: GPE jobs | 0 | 30 | 0 | 30 |
| Rule 4: Professional roles | 10 | 251 | 1 | 260 |
| Rule 5: Percentage NPs | 6 | 3 | 5 | 4 |
| Rule 6: Set members | 8 | 4 | 4 | 8 |
| Rule 7: Arg-taking I | 3 | 38 | 0 | 41 |
| Rule 8: Arg-taking II | 14 | 163 | 4 | 173 |

Table 6.14.: Applying Hou et al. (2014) on the RST part of the ARRAU corpus: rule performance

Rule 1: building parts

- (16) Once inside, she spends nearly four hours measuring and diagramming each room in the 80-year-old house [...] She snaps photos of **the buckled floors** ... (correct)
- (17) And now Kellogg is indefinitely suspending work on what was to be a 1 billion cereal plant. The company said it was delaying **construction** ... (wrong)

Rule 2: relatives

- (18) I heard from **friends** that state farms are subsidized ... (wrong)

Rule 3: GPE jobs

- (19) The fact that New England proposed lower rate increases [...] complicated negotiations with **state officials** (wrong)

It is probably controversial whether *state officials* should be annotated as bridging, as it can also be a generic reference to the class. However, in this case, it is neither annotated as generic nor as bridging.

6. Bridging resolution

Rule 4: professional roles

- (20) Meanwhile the National Association of Purchasing Management said its latest survey indicated [...] . **The purchasing managers**, however, also said that orders turned up in October ... (correct)
- (21) A series of explosions tore through the huge Phillips Petroleum Co._{pred} plastics plant near here_{gold}, injuring more than a hundred and [...]. There were no immediate reports of deaths, but **officials** said a number of workers ... (different antecedent/antecedent overlap)

Rule 5: percentage expressions

- (22) Only 19% of the purchasing managers reported better export orders [...]. And 8% said export orders were down ... (correct)

Rule 6: set members

- (23) Back in 1964, the FBI had five black agents. **Three** were chauffeurs for ... (correct)
- (24) ... a substantial number of people will be involved.
Some will likely be offered severance package ... (wrong)

Rule 7: argument-taking I

- (25) In ending Hungary's part of the project, **Parliament** authorized Prime Minister Miklos Meneth ... (wrong)
- (26) Sales of information-processing products_{pred} increased and accounted for 46% of total sales_{gold}. In audio equipment, **sales** rose 13 % to ... (different antecedent)

Rule 8: argument-taking II

- (27) As aftershocks shook the San Francisco Bay Area, rescuers searched through rubble for survivors of Tuesday's temblor, and **residents** picked their way through ... (correct)
- (28) Lonnie Thompson, a research scientist at Ohio State_{pred} _{gold} who dug for and analyzed the ice samples. To compare temperatures over the past 10,000 years,

researchers analyzed ...
 (different antecedent/antecedent overlap)

Conclusion We soon realised that the annotations differ quite a lot with respect to the understanding of the category bridging. We noticed that besides predicting wrong pairs, the original system would suggest bridging pairs which are fine from the point of view on bridging as annotated in ISNotes, but are not annotated in the ARRAU corpus, such as Example (29).

- (29) As competition heats up in Spain's crowded bank market, [...].
The government directly owns 51.4% and ...

Additionally, it would miss a lot of annotated bridging pairs, which are of a different type, such as in Example (30) or (31). As these often involve mentions with matching heads, they are filtered out as anaphor candidates in the preprocessing step of the system.

- (30) Her husband and older son [...] run a software company. Certainly life for her has changed considerably since the days in Kiev, when she lived with her parents, her husband and **her two sons** in a 2 1/2-room apartment. (*relation: element-inverse*).
- (31) Dennis Hayes and Dale Heatherington, two Atlanta engineers, were co-developers of the internal modems that allow PCs to share data via the telephone. IBM, the world leader in **computers** ... (*relation: subset-inverse*)

This is why the performance is so poor: a lot of reasonable bridging pairs which are not annotated were predicted, while the system missed almost all instances that have been annotated as bridging in the corpus, using a different concept of bridging which we will discuss in the next section.

The differences between ISNotes and ARRAU are very fundamental and need to be discussed in more detail. Hence, we will go back to the first step in the pipeline, task definition, and present a categorisation scheme that explains these differences.

6.3. A refined bridging definition

At this point, we are taking a step back (or two steps, to be more precise) and go back to the task definition. Some of the issues in bridging and bridging resolution have been

6. Bridging resolution

controversial for a long time, e.g. the question of definiteness. The difference between the annotations in ISNotes and ARRAU, however, are not yet covered in previous discussions about the phenomenon.

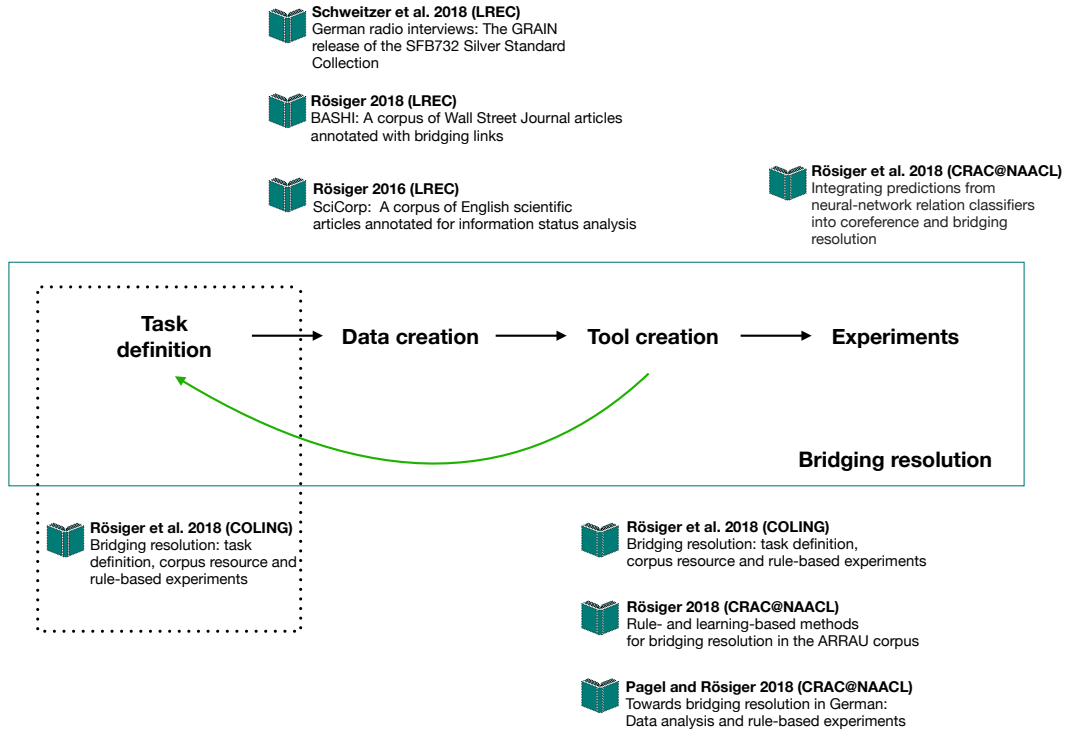


Figure 6.2.: Contribution and workflow pipeline for bridging: task definition (reloaded)

We introduce the concept of referential and lexical bridging, inspired by the two-level RefLex annotation scheme by Baumann and Riester (2012). The two terms describe two different phenomena which are currently both defined and annotated as bridging. This refined characterisation of bridging is the result of a collaboration with Arndt Riester and has been presented in Rösiger et al. (2018b).

6.3.1. Referential bridging

Referential bridging describes bridging at the level of referring expressions, i.e. we are considering noun phrases that are truly anaphoric, in the sense that they need an antecedent in order to be interpretable, like in Example (32). As such, (referential) bridging anaphors are non-coreferent, context-dependent expressions.

- (32) The city is planning a new town hall and **the construction** will start next week.

Referential bridging is often a subclass of (referential) information status annotation. We claim that there are two types of referential bridging: the first (and most frequent) type are expressions which require for their interpretation the antecedent as an implicit argument, e.g. *the construction of the new town hall* in Example (32). When uttered out of context, their referent is unidentifiable. The second type involves anaphoric subset expressions, as shown in Example (33).

- (33) I saw some dogs yesterday. **The small pug** was the cutest.

Again, context-dependence is taken as the main criterion for classifying this as referential bridging. The **subset** type is however different from the first type of referential bridging, as we are not filling an argument slot (**the small pug of some dogs*), but only expressing the fact that the expression is only interpretable because we have mentioned the set *some dogs* before and *the small pug* is a subset of this group.

Referential bridging anaphors are typically short, definite expressions (*the construction, the door*), and several accounts explicitly restrict bridging to definites, e.g. Poesio and Vieira (1998), Nedoluzhko et al. (2009), Grishina (2016), Rösiger (2016) or Riester and Baumann (2017), while others also allow for indefinite bridging, e.g. Löbner (1998) or Markert et al. (2012), with the consequence that some studies have linked indefinites as bridging anaphors (e.g. in ISNotes and others). Although having held different views on this issue, we now think that indefinite expressions can indeed – in some cases – be referential bridging anaphors, for example in Example (34) or Example (35), where the (partitive) expressions *one employee (of Starbucks)* or *leaves (of the old oak tree)* are introduced.

- (34) Starbucks has a new take on the unicorn frappuccino. **One employee** accidentally leaked a picture of the secret new drink.
- (35) Standing under the old oak tree, she felt **leaves** tumbling down her shoulders.

However, while short, definite expressions signal identifiability and are thus either anaphoric expressions or familiar items, it is much harder to decide which indefinite expressions are bridging anaphors, since indefinite expressions are prototypically used to introduce new discourse referents and principally do not need an antecedent/argument

6. Bridging resolution

in order to be interpretable. This is, for example, also reflected in the higher inter-annotator-agreement for definite than for indefinite bridging anaphors (Rösiger, 2018a).

Thus, despite the interpretational uncertainty surrounding indefinites, we take linguistic anaphoricity/context-dependence to be the defining criterion for referential bridging. Semantic relations like meronymy will be addressed in the next section under the notion of lexical bridging. It is important to concede, however, that the reason why certain definite or indefinite expressions function as bridging anaphors (while others do not) is typically due to some kind of semantic proximity between antecedent and anaphor. However, the specific relation we are dealing with may be rather abstract, vague and difficult to define, as Example (34) shows.

6.3.2. Lexical bridging

Baumann and Riester (2012) use the term “lexical accessibility” to describe lexical semantic relations, such as meronymy or hyponymy, at the word or concept level (e.g. *house* – *door*). It is important to bring to mind that lexical relations are defined as part of the intrinsic meaning of a pair of concepts, thus, abstracting away from specific discourse referents: it is the words *house* and *door* which stand in a meronymic relation, not two actual physical objects or their mental images, although typically the referents of a holonym-meronym combination will, at the same time, stand in a physical **whole-part** relation. Since this physical relation has often been taken as one of the defining criteria for bridging, e.g. by Gardent et al. (2003), Nissim et al. (2004), Nedoluzhko et al. (2009) or Grishina (2016), we suggest using the term lexical (or lexically induced) bridging for this phenomenon.

The referents of the proper nouns *Europe* and *Spain* are in a **whole-part** relation,¹¹ and the referring expressions can thus be considered a case of lexical bridging. However, the expression *Spain* is not anaphoric, since its interpretation does not depend on the “antecedent” *Europe*. **Whole-part** is probably the prototypical pre-defined relation, and it is a straightforward concept to annotate in the case of nouns denoting physical objects. However, it is less applicable in connection with abstract nouns, which is why many additional relations have been suggested, including, for instance **thematic role in an event**, **attribute of an object** (like *price*), **professional function in an organisation** (like *president*), **kinship** (like *mother*), **possessed entity** and so on.

¹¹Note that for proper nouns (names), like *Spain*, there is a one-to-one mapping between the word and its referent in the real world, which is not the case for common nouns, cf. Kripke (1972).

And yet, few schemes get by without an **other** category for the many examples which cannot be naturally classified into one of the assumed classes.

It should be noted that lexical and referential bridging are two different concepts with completely different properties: one deals with the question of pragmatic anaphoricity (or grammatical saturation) of an expression, the other with lexical proximity between two words and the relation between entities in the real world, although the two types of bridging often co-occur within one and the same pair of expressions, such as in Example (36), where we have a relation of meronymy between the content words *sea urchin(s)* and *spine(s)*, but also an anaphoric relation between the referring expressions *most sea urchins* and *the spines*, i.e. a case of referential bridging.

(36) In most sea urchins, touch elicits a prompt reaction from **the spines**.

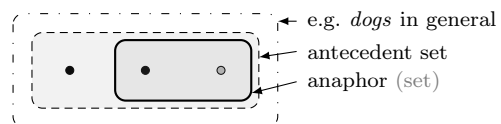
The second release of the ARRAU corpus (Uryupina et al., 2018), as used in the first shared task on bridging resolution, for example, contains instances of both referential and lexical bridging, with the majority of the bridging links being purely lexical bridging pairs, i.e. most expressions labelled as bridging are actually not context-dependent.

6.3.3. Subset relations and lexical givenness

Another relation often brought up in connection with (lexical) bridging is the **subset** or **element-of** relation, which is the most common relation in ARRAU.¹² In principle, an expression referring to an element or a subset of a previously introduced group can be of the referential type of bridging, like in Example (37), where the anaphor is interpreted as *the small pug (from the prementioned group of dogs)*, but this is not always the case, as Example (38) shows, where the bridging anaphor is not context-dependent.

(37) I saw some dogs yesterday. **The small pug** was the cutest.

(38) Newsweek said it will introduce the Circulation Credit Plan, which awards space credits to advertisers on renewal advertising. The magazine will reward with page bonuses **advertisers who in 1990 meet or exceed their 1989 spending, [...]**

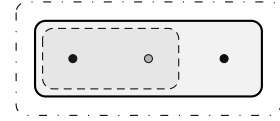


¹²The graphics in this section were provided by Arndt Riester.

6. Bridging resolution

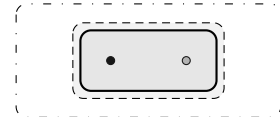
The subset relation can sometimes be reversed, as shown in Example (39), where, again, no context-dependence is involved.

- (39) I saw a small pug yesterday. I like
many dogs.



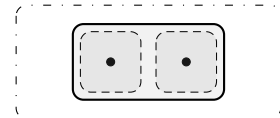
It should be noted, however, that **subset/element-of** pairs also have much in common with coreference pairs, since the lexical relation between their head nouns tends to be hypernymy, synonymy or plain word repetition (lexical relations which are summarised as *lexical givenness* in Baumann and Riester, 2012) or hyponymy (i.e. *lexical accessibility*). Note that, although the antecedent and anaphor expressions in Example (40) stand in a hypernym-hyponym relation (or reverse), their respective referent is the same. Hence, these cases do not exemplify bridging but coreference.

- (40) a. I saw a dog yesterday. **The small pug** was very cute.
b. I saw small pugs yesterday.
The dogs were very cute.



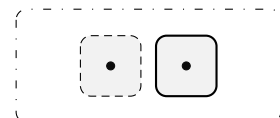
Note that **element-of** bridging is also conceptually very close to the phenomenon of aggregation/summation, in which the group entity follows a list of elements, and which also counts as a case of coreference.

- (41) I saw a pug and a Yorkshire terrier. **The dogs** were very cute.



A final case, which is treated as a special class of information status in Markert et al. (2012) and annotated as a subclass of bridging in ARRAU, are so-called **comparative** or **other-anaphors**. The head noun of the anaphor must be lexically given (Riester and Piontek, 2015, 242f.) and the two expressions are marked as two contrastive elements from the same alternative set (Rooth, 1992). Comparative anaphors can be considered cases of referential bridging where the implicit argument is the implicit or explicit alternative set, i.e. *another dog (from a specific or unspecific set dogs)*.

- (42) I saw a small pug two days ago and **another dog** yesterday.



6.3.4. Near-identity

While many approaches distinguish only between coreferent anaphors, which refer to the same referent as their antecedent, and bridging anaphors, which refer to a different referent, Recasens and Hovy (2010a) and Recasens et al. (2012) have introduced a third concept, the concept of near-identity, which has been picked up by others, e.g. Grishina (2016). Near-identity is defined to hold between an anaphor and an antecedent whose referents are almost identical, but differ in one of four respects: name metonymy, meronymy, class or spatio-temporal functions.

- (43) On homecoming night Postville feels like Hometown, USA, but a look around this town of 2,000 shows its become a miniature Ellis Island . . . For those who prefer **the old Postville**, Mayor John Hyman has a simple answer.

We believe that the introduction of this additional category in between coreference and bridging introduces more uncertainty and, therefore, potentially makes the annotation process more difficult. Example (43), for instance, is structurally analogous to comparative anaphors.

6.3.5. Priming and bridging

Another issue that we observed in the GUM corpus was that sometimes a referring expression is annotated as bridging because the entity has been “primed”, i.e. something from the context has raised our expectations so that we can now easily build a bridge to the before mentioned entity. Consider Example (44), where *the Dark Knight* refers to a rather popular Batman movie.

- (44) The Batman movies ... **The Dark Knight** is my favourite.¹³

Of course, the context of *the Batman movies* makes it more likely that *The Dark Knight* is mentioned in the following text. Still, *The Dark Knight* as a title of a movie is not context-dependent, and in our opinion either of the information status category `unused-known` or `unused-unknown`, depending on the reader’s knowledge. As such, it is a case of a non-anaphoric subset relation. Softening the border between the category `unused` and `bridging` by introducing the concept of an expression that has been primed by some previous context does in our opinion again result in a less clear bridging definition.

¹³Example by Amir Zeldes, personal communication

Apart from these cases of “primed” bridging, GUM contains mostly referential bridging in the form of argument filling or referential subset relations. We also found some cases of aggregation annotated as bridging, which we see as a special case of coreference.

6.4. Shared task results

ARRAU seems to contain a rather small fraction of referential bridging pairs, and a large number of lexical pairs. This is probably because the focus of the annotation was set on the pre-defined relations, such as **subset**.

The following example, where the rule-based system has identified a gold bridging anaphor shows the different views with respect to the antecedent chosen: the gold annotations tell us that *Jan Leemans, research director* is a subset of *researchers*, whereas the predicted antecedent tells us that he is the *research director at Plant Genetic Systems*, reflecting the argument slot filling type of referential bridging.

(45) At Plant Genetic Systems_{pred}, researchers_{gold} have isolated a pollen-inhibiting gene that [...] . **Jan Leemans, research director**, said ...

6.4.1. Rules for bridging in ARRAU

With the modular approach of the rule-based system, one can define new rules to also capture lexical bridging and lexical givenness. We add a number of rather specific rules, which are meant to increase precision, but also include more general rules to increase recall. The rules have been developed on the training and development set of the RST domain of the corpus. We also leave in three rules of the original rule-based system: building parts (Rule 1), percentage expressions (Rule 5) as well as set members (Rule 6). The final performance of the adapted system (F-score of 19.5) is given in Table 6.15. The new rules are presented in the following.

While this adaptation was done to achieve high results on the shared task data, we argue that this is generally not a good way to achieve progress in bridging resolution. As ARRAU contains a mix of referential and lexical bridging and lexical givenness, it should not be used as a data basis for a general bridging system. As referential and lexical bridging are two different phenomena, they should not be modelled in a mixed bridging system. We suggest creating different systems for the two tasks, either by using corpora that only contain one type (such as ISNotes for referential bridging) or by

| Corpus | Domain | Anaphor recognition | | | Full bridging resolution | | |
|-------------------------------------|--------|---------------------|--------|------|--------------------------|--------|-------------|
| | | Prec. | Recall | F1 | Prec. | Recall | F1 |
| ISNotes (gold markables) | news | 65.9 | 14.1 | 23.2 | 49.6 | 10.6 | 17.4 |
| ISNotes (pred markables) | news | 65.2 | 13.6 | 22.5 | 49.2 | 10.3 | 17.0 |
| BASHI (pred markables) | news | 49.4 | 20.2 | 28.7 | 24.3 | 10.0 | 14.1 |
| ARRAU (original, gold markables) | news | 13.3 | 0.9 | 1.7 | 2.2 | 0.2 | 0.3 |
| ARRAU (adapted, gold markables) | news | 29.2 | 32.3 | 30.8 | 18.5 | 20.6 | 19.5 |

Table 6.15.: Performance of the rule-based method on other corpora. We use predicted mentions for BASHI and SciCorp as they do not contain gold markables.

labelling the bridging relations according to their type and treating them separately if you use one corpus that contains both.

Comparative anaphora Contrary to ISNotes, the ARRAU corpus contains comparative anaphors, which are labelled with the relation **other**. For a markable to be considered a comparative anaphor, it must contain a comparative marker¹⁴, e.g. *two additional rules*, *the other country*, etc. We then search for the closest markable which is of the same category than the anaphor and whose head matches its head in the last seven sentences. If this search is not successful, we search for an antecedent of the same category as the anaphor in the same and previous sentence. If this fails too, we search for a markable with the same head or a WordNet (WN) synonym appearing before the anaphor.

(46) the issue ... **other issues in memory**

We exclude a couple of very general terms, such as *things* or *matters* as potential anaphors, as they are typically used non-anaphorically, such as in Example (47).¹⁵

(47) Another thing is that ...

Subset/Element-of bridging This is a rather general rule to capture mostly lexical bridging and lexical givenness cases of the relations **subset/element**.

As the anaphor is typically more specific than the antecedent (except for cases of the relation **subset-inverse/element-inverse**), it must be modified by either an adjective,

¹⁴other, another, similar, such, related, different, same, extra, further, comparable, additional

¹⁵The full list is: *thing, matter, year, week, month*.

6. Bridging resolution

a noun or a relative clause. We then search for the closest antecedent of the same category with matching heads in the last three sentences.

(48) computers ... **personal computers**

If this fails, we check whether the head of the anaphor is a country. If so, we look for the closest antecedent with *country* or *nation* as its head in the same sentences or the previous five sentences. This is rather specific but helps find many pairs in the news domain.

(49) countries... **Malaysia**

If this also fails, we take the closest WordNet synonym of the same category within the last three sentences as the antecedent. Again, we use our small list of general terms to exclude rather frequent general expressions, which are typically not of the category bridging.

Time subset For this rule, we list a number of time expressions, such as *1920s*, *80s*, *etc.* The anaphor must be of the category **time** and must be one of those time expressions. We then search for the closest antecedent of the same category in the last seven sentences for which the decade number matches.

(50) 1920s ... **1929**

(51) the 1950s ... **the early 1950s**

One anaphora We search for expressions where *one* is followed by a common noun. We then remember the common noun part of the expression and search for the closest plural entity of the same category whose common noun part matches the common noun part of the anaphor. Taking into account all words with a common noun tag turned out to work better than just comparing the heads of the phrases.

(52) board members ... **one board member**

If this rule does not apply, we look for anaphor candidates of the pattern *one of the N* and again search for the closest plural entity for which the common noun part of the expressions matches.

(53) the letters ... **one of the letters**

As in a few of the other rules, we exclude a couple of very general terms as they typically do not refer back to something that has been introduced before.

Locations In the RST data, a lot of cities or areas are linked to their state/country. We can find these bridging pairs with the WordNet relation `partHolonym`. To be considered an anaphor, the markable must be of the category `space` or `organization` whose size is three words or less (as to exclude modification and arguments). We then search for the closest antecedent of the same category that is in a WN `partHolonym` relation with the anaphor.

(54) California ... **Los Angeles**

(55) Lebanon ... **Beirut**

Same heads This rule is very similar to the subset/element-of rule, but is designed to find more cases that have not yet been proposed by the subset/element-of rule. For a markable to be considered an anaphor, it must be a singular, short NP (containing four words or less). We then search for the closest plural expression of the same category whose head matches the head of the anaphor or that is in a WordNet synonym relation with the anaphor's head, in the last five sentences.

(56) Democrats ... **a democrat**

If this fails, we look at singular markables with a maximal size of three words which contain an adjective as anaphor candidates, and then search for a plural antecedent of the same category whose head matches the head of the anaphor or that is in a WordNet synonymy relation with the anaphor's head, in the last seven sentences.

(57) the elderly ... **the young elderly**

(58) market conditions ... **current market conditions**

If this also fails, we look for `inverse` relations, i.e. a plural anaphor and a singular antecedent of the same category and matching heads/WN synonym in the last seven sentences.

(59) an automatic call processor that **Automatic call processors**

6. Bridging resolution

Persons In this rather specific rule, we search for expressions containing an apposition which refer to a person, e.g. *David Baker, vice president*. For this, the anaphor candidate must match such a pattern and be of the category **person**. As the antecedent, we choose the closest plural person NP whose head matches the head of the apposition.

(60) Specialists ... **John Williams, a specialist**

The rest This rule is also very specific and aims to resolve occurrences of *the rest*, which, in many cases, is annotated as a bridging anaphor. We thus search for occurrences of *the rest* and propose as an antecedent a number expression within the last three sentences.

(61) 90 % of the funds ... **The rest**

Proposing antecedents for all remaining anaphors For the task of bridging anaphora resolution, i.e. choosing an antecedent for a given anaphor, we need to force the system to propose an antecedent for every bridging anaphor.

This is why we include a couple of rules, which are applied in the order presented here and which propose an antecedent for every anaphor which has not yet been proposed as an anaphor by the other rules.

- Pronoun anaphors:

The anaphor must be a pronoun of the category **person**. As the antecedent, we chose the closest plural person NP in the last two sentences.

(62) At a recent meeting of manufacturing executives, everybody I talked with was very positive, he says. Most say **they** plan to ...

This is in a way a strange annotation, as pronouns should in theory always be coreference anaphors, not bridging anaphors. An alternative annotation would be to link *they* back to *most*, and *most* as a bridging anaphor to *manufacturing executives*.

- WordNet synonyms in the last three sentences.

(63) The purchasing managers ... **250 purchasing executives**

- Cosine similarity greater than 0.5 in the last seven sentences.

This rule is meant to find more general related cases of bridging. For the cosine similarity, we take the word2vec pre-trained vectors (Mikolov et al., 2013).

(64) “Wa” is Japanese for team spirit and Japanese ballplayers have miles and miles of it. **A player’s commitment** to practice ...

- The anaphor is a person and the antecedent is the closest organisation in the last two sentences.
- First word head match: choose the closest antecedent within the last two sentences, where the anaphor and antecedent both start with a proper noun.
- Same category in the last three sentences, choose the closest.

(65) ... that have funneled money into his campaign. After **his decisive primary victory over Mayor Edward I. Koch**

- Global headmatch/WordNet synonyms: “global” in this case means that we search for an antecedent in the whole document.
- Global same category.
- Choose the closest NP as a fallback plan.

6.4.2. A learning-based method

To compare the performance of the rule-based system with a learning-based method, we set up an SVM classifier¹⁶, which we provide with the same information as the rule-based system.

The classifier follows a pair-based approach similar to Soon et al. (2001), where the instances to be classified are pairs of markables. For training, we pair every gold bridging anaphor with its gold antecedent as a positive instance. As a negative instance, we pair every gold bridging anaphor with a markable that occurs in between the gold anaphor and gold antecedent.¹⁷ During testing, we pair every markable except the first one in the document with all preceding markables. As the classifier can classify more than one

¹⁶Using Weka’s SMO classifier with a string to vector filter

¹⁷This is a common technique in coreference resolution, done in order to reduce the number of negative instances and help the imbalance issue of having more non-coreferent/non-bridging cases than coreferent/bridging ones.

6. Bridging resolution

antecedent-anaphor-pair as bridging for one anaphor, we choose the closest antecedent (closest-first decoding).

As the architecture of the machine learning is not designed to predict at least one antecedent for every given bridging anaphor (it can classify all pairs of antecedent-anaphor for one anaphor as “not bridging”), we cannot report results for bridging anaphora resolution (all). However, we report results for partial bridging anaphora resolution, where, during training, we pair the gold bridging anaphors with all preceding markables, instead of pairing all markables with all preceding markables as in the full bridging scenario.

We define the following features. Features marked with a ? are boolean features.

Markable features: words in the markable, gold head form, predicted head form, noun type (proper, pronoun, nominal), category, determiner (def, indef, demonstr, bare), number, gender, person, nested markable?, grammatical role, genericity, partial previous mention?, full previous mention?, containing a comparative marker?, containing an adjective?, containing one?, containing a number?, lengths in words.

Pair features: distance in sentences, distance in words, head match?, modifier/argument match?, WordNet synonym?, WordNet hyponym?, WordNet meronym?, WordNet partHolonym?, semantic connectivity score, highest semantic connectivity score in document?, cosine similarity.

6.4.3. Final performance

| | Bridging recognition | | | Anaphora-res.-all acc | Anaphora-res.-partial | | | Full bridging | | |
|------------|----------------------|------|------|--------------------------|-----------------------|------|------|---------------|------|------|
| | P | R | F1 | | P | R | F1 | P | R | F1 |
| RST | | | | | | | | | | |
| rule-based | 29.2 | 32.5 | 30.7 | 39.8 | 63.6 | 22.0 | 32.7 | 18.5 | 20.6 | 19.5 |
| ML-based | - | - | - | - | 47.0 | 22.8 | 14.8 | 17.7 | 20.3 | 18.6 |
| PEAR | | | | | | | | | | |
| rule-based | 75.0 | 16.0 | 26.4 | 28.2 | 69.2 | 13.7 | 22.9 | 57.1 | 12.2 | 20.1 |
| ML-based | - | - | - | - | 26.6 | 5.7 | 9.4 | 5.47 | 12.5 | 7.61 |
| TRAINS | | | | | | | | | | |
| rule-based | 39.3 | 21.8 | 24.2 | 48.9 | 66.7 | 36.0 | 46.8 | 27.1 | 21.8 | 24.2 |
| ML-based | - | - | - | - | 56.6 | 23.6 | 33.3 | 10.3 | 14.6 | 12.1 |

Table 6.16.: Performance of the different systems on the tests sets of the single domains of ARRAU, using gold markables and using gold bridging anaphors in the two bridging anaphora resolution settings.

Table 6.16 shows the results of the modified rule-based approach and the learning-based approach for all tasks. It can be seen that the rule-based approach significantly outperforms the learning-based one in every setting.¹⁸ Surprisingly, in spite of the fact that the rules were designed on the training/dev sets of the RST domain, the performance for the PEAR and TRAINS domain is even better in most settings. However, this might be an effect of TRAINS and PEAR being small datasets.

Recently, the official scorer for the evaluation of the shared task has become available, which differs from our internal evaluation in the handling of some of the special cases. Table 6.17 compares our internal scores against the scores of the official scorer. In most cases, as we ignored the special cases, the scores of the official scorer are lower. However, there are also some cases where the official score is lower. In some cases, this also leads to different results, for example for the PEAR domain, the scores of the learning-based approach outperform the scores of the rule-based approach, although, with our internal scorer, the difference between the scores is quite large. This again shows the need for a refined evaluation metric. As there were no other participants in the shared task, the results in Table 6.16 are the best published results on the ARRAU datasets so far.

| | Anaphor recognition | | | Anaphora-res.-all | | | Anaphora-res.-partial | | | Full bridging | | |
|-----------------|---------------------|------|------|-------------------|------|------|-----------------------|------|------|---------------|------|------|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| RST | | | | | | | | | | | | |
| Rule (internal) | 29.2 | 32.5 | 30.7 | 39.8 | 39.8 | 39.8 | 63.6 | 22.0 | 32.7 | 18.5 | 20.6 | 19.5 |
| Rule (official) | - | - | - | 36.5 | 35.7 | 36.1 | 58.4 | 20.6 | 30.5 | 16.8 | 13.2 | 14.8 |
| ML (internal) | - | - | - | - | - | - | 47.0 | 22.8 | 30.7 | 17.7 | 20.3 | 18.6 |
| ML (official) | - | - | - | - | - | - | 51.7 | 16.2 | 24.7 | 12.6 | 15.0 | 13.7 |
| PEAR | | | | | | | | | | | | |
| Rule (internal) | 75.0 | 16.0 | 26.4 | 28.2 | 28.2 | 28.2 | 69.2 | 13.7 | 22.9 | 57.1 | 12.2 | 20.1 |
| Rule (official) | - | - | - | 30.5 | 28.2 | 29.3 | 62.5 | 11.3 | 19.1 | 53.1 | 4.8 | 8.8 |
| ML (internal) | - | - | - | - | - | - | 26.6 | 5.7 | 9.4 | 5.47 | 12.5 | 7.61 |
| ML (official) | - | - | - | - | - | - | 37.5 | 4.2 | 7.6 | 23.6 | 7.3 | 11.2 |
| TRAINS | | | | | | | | | | | | |
| Rule (internal) | 39.3 | 21.8 | 24.2 | 48.9 | 48.9 | 48.9 | 66.7 | 36.0 | 46.8 | 27.1 | 21.8 | 24.2 |
| Rule (official) | - | - | - | 47.5 | 47.3 | 47.4 | 64.4 | 36.0 | 46.2 | 28.4 | 11.3 | 16.2 |
| ML (internal) | - | - | - | - | - | - | 56.6 | 23.6 | 33.3 | 10.3 | 14.6 | 12.1 |
| ML (official) | - | - | - | - | - | - | 63.2 | 12.8 | 21.3 | 19.0 | 11.0 | 13.9 |

Table 6.17.: Performance of the different systems on the tests sets of ARRAU, using gold markables (and gold bridging anaphors in the anaphora resolution settings). We report performance using the official and our own internal scorer.

Table 6.18 shows the rules and their performance in the final system for full bridging resolution. As we only applied the rules on the test set after having developed the rules,

¹⁸Significance computed using the Wilcoxon signed ranked test, at the p=0.05 level.

6. Bridging resolution

some rules are included which do not predict any pairs because they predicted pairs in the training/dev setting (on which the system was designed).

| Rule | Anaphor recognition | | | Full bridging resolution | | |
|-------------------|---------------------|-------|-----------|--------------------------|-------|-----------|
| | Correct | Wrong | Precision | Correct | Wrong | Precision |
| 1: Building parts | 0 | 0 | - | 0 | 0 | - |
| 2: Percentage | 1 | 0 | 100.0 | 1 | 0 | 100.0 |
| 3: Set members | 1 | 1 | 50.0 | 0 | 2 | 0.0 |
| 4: Comp anaphora | 44 | 16 | 73.3 | 26 | 34 | 43.3 |
| 5: Subset/element | 57 | 247 | 18.8 | 34 | 270 | 11.2 |
| 6: Time subset | 3 | 6 | 33.3 | 3 | 6 | 33.3 |
| 7: One anaphora | 0 | 0 | - | 0 | 0 | - |
| 8: Locations | 25 | 11 | 69.4 | 22 | 14 | 61.1 |
| 9: Head matching | 72 | 236 | 23.4 | 42 | 266 | 13.6 |
| 10: The rest | 1 | 1 | 50.0 | 0 | 2 | 0.0 |
| 11: Person | 10 | 1 | 90.9 | 8 | 3 | 72.7 |

Table 6.18.: Performance of the single rules for full bridging resolution on the test set of the RST dataset, using gold markables

6.5. A rule-based bridging system for German

Most of the work on bridging resolution, with its subtasks of anaphor detection and antecedent selection, has focused on English (e.g. Hou et al., 2014; Markert et al., 2012; Rahman and Ng, 2012). For German, Grishina (2016) has presented a corpus of 432 bridging pairs as well as an in-depth analysis on some properties of bridging, e.g. on the distance between anaphors and their antecedents and on the distribution of bridging relations. Apart from Cahill and Riestler (2012)’s work on bridging anaphor detection as a subclass in information status classification and Hahn et al. (1996)’s early work on bridging resolution, there have been no automatic approaches to bridging resolution in German.

German corpora containing bridging annotations have been presented in Section 4.2. Apart from the Coref pro corpus (Grishina, 2016), which has recently been made available, there is only the SemDok corpus by Bärenfänger et al. (2008), which is not openly available. To the best of our knowledge, DIRNDL is currently the largest German dataset containing bridging annotations.

Hence, we think it is a valuable effort to adapt the bridging system described for English to German. While the adaptation process addresses specificities of German, it also needs to take into account the properties of the available training data. This section

presents the adaptation to German and experiments on bridging anaphor detection and full bridging resolution. As the annotation on the newly created corpus GRAIN has only recently been completed, experiments on bridging in GRAIN are not featured in this thesis. However, the GRAIN corpus has been used as the data basis in a recent Master thesis (Pagel, 2018). Our joint results have been published in Pagel and Rösiger (2018). In this thesis, we focus on bridging resolution in the DIRNDL corpus.

6.5.1. Adaptation to German

Related work

Corpora Table 6.19 compares the corpora containing bridging annotations in German. As can be seen, DIRNDL is the largest resource, with 655 bridging pairs, followed by the Coref pro corpus and GRAIN. Unfortunately, not much is known about the SemDok corpus, which seems to be currently unavailable. As the Coref pro and GRAIN corpora only became available after our experiments have been performed, we based our adaptation process on the DIRNDL corpus.

| Corpus | Available | Genre | Bridging pairs | Anaphors | Other properties |
|-----------|-----------|--------------------------|----------------|----------|------------------------|
| DIRNDL | Yes | radio news | 655 | definite | - |
| Coref pro | Yes | news, narrative, medical | 432 | definite | near-identity involved |
| SemDok | No | scientific+news | ? | all NPs | - |
| GRAIN | Yes | interviews | 274 | definite | - |

Table 6.19.: Overview of German corpora annotated with bridging

A quick recap on DIRNDL The DIRNDL corpus (Eckart et al., 2012; Björkelund et al., 2014), a corpus of radio news, contains bridging annotations as part of its information status annotation (on transcripts of the news), following older guidelines of the RefLex scheme (Baumann and Riester, 2012). Overall, 655 bridging pairs have been annotated. Apart from the manual information status annotation, other linguistic annotation layers (POS-tagging, parsing, morphological information) have been created automatically.

Computational approaches Cahill and Riester (2012) presented a CRF-based automatic classification of information status, which included bridging as a subclass. However, they did not state the accuracy per class, which is why we cannot derive any

6. Bridging resolution

performance estimation for the task of bridging anaphor detection. They stated that bridging cases “are difficult to capture by automatic techniques”, which confirms intuitions from information status classification for English, where bridging is typically a category with rather low accuracy (Markert et al., 2012; Rahman and Ng, 2012; Hou, 2016a). Hahn et al. (1996) and Markert et al. (1996) presented a resolver for bridging anaphors, back then called textual ellipsis or functional anaphora, in which they resolved bridging anaphors in German technical texts based on the centering theory and a knowledge base. The corpus and the knowledge base as well as the overall system are, however, not available, which makes a comparison with our system difficult.

Bridging definition in RefLex

As both the DIRNDL and GRAIN corpus were annotated according to the RefLex scheme (Baumann and Riester, 2012; Riester and Baumann, 2017), we repeat the main idea of this scheme, as well as its implications for bridging anaphors.

RefLex distinguishes information status at two different dimensions, namely a referential and a lexical dimension. The referential level analyses the information status of referring expressions (i.e. noun phrases) according to a fine-grained version of the given/new-distinction, whereas the lexical level analyses the information status at the word level, where content words are analysed as to whether the lemma or a related word has occurred before.

Bridging anaphors are a subclass of referential information status. On the referential level, indefinite expressions are considered to be discourse-new and are thus treated as expressions of the information status category **new**. Therefore, the bridging anaphors in our data are always definite. This is a major difference between the annotations in DIRNDL and GRAIN compared to the ISNotes annotations. This fact needs to be considered during the adaptation process but is of course not a specificity of German but rather a guideline decision.

In RefLex, **bridging-contained** is a separate information status class, where the anaphor is an argument of the antecedent, either in a prepositional phrase or a possessive construction, e.g. in *the approach’s accuracy* or *the accuracy of the approach*. In this thesis, we do not cover these cases.

Unlike ISNotes, which features gold annotations, DIRNDL and GRAIN were processed using automatic NLP tools. Systems trained on automatic annotations typically achieve lower performance, as errors during other pre-processing steps are propagated to the bridging resolution step.

Finally, RefLex suggests annotating PPs rather than their embedded NPs. This has to be reflected during markable extraction.

Experimental setup

DIRNDL revision One issue that we observed is that the DIRNDL annotations grew while the RefLex guidelines were still optimised. As a result, not all rules that are nowadays stated in the guidelines have been implemented correctly. Firstly, many cases are of the type shown in Example (66) and (67)¹⁹.

- (66) DE: Der Iran ... **an seinem Atomprogramm**
 EN: Iran ... **their nuclear programme**
- (67) DE: Der Iran ... **an deren Atomprogramm**
 EN: Iran ... **whose nuclear programme**

These cases, where the anaphor contains a possessive or a demonstrative pronoun, are typical cases of the category `bridging-contained` and should not be labelled as bridging according to the final version of the guidelines. Secondly, in cases where the annotators could not find a suitable antecedent, they did not annotate one. As a result, some anaphors do not have an antecedent. Also, although indefinite expressions should not be considered bridging anaphor candidates, there are some indefinite expressions that have been labelled as bridging anaphors, e.g. in Example (68).

- (68) DE: Die USA haben die verschärfte UNO-Resolution des UNO-Sicherheitsrates begrüsst. US-Staatssekretär Burns sprach **von einer aussagekräftigen Zurechtweisung**.
 EN: The US have welcomed the tightened UNO resolution ...
 US state secretary Burns called it **a meaningful rebuke**.

Thus, an automatic revision of the DIRNDL data was necessary to make the annotations more consistent. We automatically filtered out the following bridging anaphor candidates using part-of-speech patterns:

- Indefinite bridging anaphors (and their antecedents);
 - expressions with an indefinite article, *ein Gedanke (a thought)*;

¹⁹DIRNDL is not a parallel corpus, the translations are only included for readers that do not understand German.

6. Bridging resolution

- indefinite number expressions, *23 Deutsche (23 Germans)*;
- negated indefinite expressions, *kein Deutscher (no German)*;
- adverbs are taken into account, i.e. *rund 23 Deutsche (about 23 Germans)* is also filtered out.

- Bridging anaphors of the type **bridging-contained**, as in *sein Atomprogramm (their nuclear programme)*, *nach seinem Beginn (after its start)*. Adverbs are again taken into account, i.e. *erst nach ihrer Bergung (only after their rescue)* is also changed to the type **bridging-contained**;
- Anaphors without an antecedent.

Note that other, more complicated patterns, have not been taken into account, e.g. the bridging-contained cases (which are still marked as bridging), as in Example (69), where the information that *the north and south of the country* refers to *Sri Lanka* is already established by linking *the country* and *Sri Lanka* as a coreferent pair. The information is thus contained in the markable and it is not necessary to link this as a bridging case.

- (69) DE: Seit Wochen geht die Luftwaffe von Sri Lanka gegen Rebellenstellungen **im Norden und Süden des Landes** vor.
- EN: For weeks, the air forces of Sri Lanka have bombed rebellion posts in **the north and south of the country**.

In very obviously wrongly marked cases (like in the example above), the label has been changed by hand. This affects only a few of the bridging cases.

An open question remains what should be done with markables of the category **generic**. In DIRNDL, these cases have their own information status label **generic**. The new corpus GRAIN and the new annotation guidelines do not contain a category **generic**, so this has been changed into an attribute, i.e. another information status category is annotated which is given the attribute **generic**. Thus, DIRNDL contains about 1500 cases of generic markables, so a re-annotation effort is rather costly. As we will later see, as a result, some reasonable candidates proposed by the system, e.g. *die Jugend - Europa* in Example (70), are considered wrong because they are annotated as **generic**.

- (70) DE: Sie kommen aus allen 27 Mitgliedstaaten und tauschen ihre Vorstellung von der Zukunft Europas aus. Die EU-Kommission bezeichnet das Treffen in Rom als Auftakt für einen neuen Dialog zwischen den

europäischen Institutionen und **der Jugend**.

EN: ... they are sharing their views on the future of Europe.

... a new dialogue between the European institutions and **the Youth**.

The newly improved annotations have been made available on the DIRNDL webpage. For optimisation, we use the development set and we report performance on the test set, if not indicated otherwise. We also report the performance on the whole DIRNDL corpus.

Preprocessing and rules for German

Markables RefLex suggests annotating PPs rather than their embedded NPs, in order to handle merged forms of determiners and prepositions. Therefore, we extract NPs (if not embedded in a PP) and PPs as our predicted markables. We extract all markables with information status annotation as our set of gold markables.

Filtering of bridging anaphor candidates As potential bridging anaphor candidates, we filter out a number of noun types, as they are not considered bridging anaphors:

- Pronouns: all pronouns are excluded as they are typically either pleonastic or coreferent with an already introduced entity.
- Indefinite expressions: all indefinite markables should, as stated in the guidelines, not be bridging anaphor candidates. We use a set of definite determiners to determine the definiteness of the markables.
- Proper names: proper names are also definite, but are not suited as bridging anaphors as they typically occur as expressions of the category `unused/mediated-general`. NPs containing embedded proper names can, of course, be of the category bridging and should not be excluded.
- Markables whose head has appeared before in the document are excluded. This is meant as an approximation for coreference anaphors.
- NPs that have embedded NPs are excluded. In practice, this leads to the exclusion of long NPs that have embedded markables, e.g. in Example (71).

(71) DE: unter dem Deckmantel der zivilen Nutzung der Nuklearenergie
 EN: under the guise of civilian use of nuclear energy

6. Bridging resolution

These expressions are typically of the information status category **unused-unknown**.

Filtering of bridging antecedent candidates When using predicted markables, it sometimes happens that overlapping markables are extracted. To overcome this, we filter out embedded named entities (NEs) in NPs or PPs from the set of potential antecedents, but only if the NP or PP differs from the NE solely in the form of a determiner, preposition or a pre-modifying noun, as in the following examples:

(72) Der Iran

(73) Im Iran

(74) Bundesaußenminister Steinmeier (*Foreign secretary Steinmeier*)

Not excluded are embedded NPs in other constructions, for example involving genitives, e.g. in Example (75).

(75) auf Wunsch Spaniens (*at Spain's discretion*)

Rules

We have implemented and adapted to German all eight rules as proposed by Hou et al. (2014). The input to the rules are the extracted markables. Each rule then proposes bridging pairs, independently of the other rules. The rules have been described in detail in the reimplementation description of the English rule-based system.

A distributional lexical resource for German

Similar to the English bridging system, some of the rules require a distributional lexical resource, which is described in the following.

Computing the semantic connectivity

The concept of semantic connectivity was introduced in the reimplementation of the English bridging resolver. In a nutshell, the semantic connectivity between two words can be approximated by the number of times two words occur in an “N PREP N” pattern.

This means that two nouns like *Sand* and *Strand* (*sand and beach*) have a high semantic connectivity because they often occur as *Sand am Strand* (*sand on the beach*), whereas other nouns do not often appear in such a construction and are therefore not highly semantically connected.

We take the SdeWaC corpus (Faaß and Eckart, 2013), a web corpus of 880 M tokens, to compute the semantic connectivity for all combinations of nouns that occur in this prepositional pattern in the corpus. This way, we not only compute the numbers for nouns in DIRNDL, but also for other nouns, making the approach applicable for new texts.

In contrast to English, German has many one-word compounds, like *Hüpfkind* (*jumping kid*), *Schreikind* (*screaming kid*). Many of these are infrequent, thus leading to sparsity issues. To overcome this, we apply the compound splitter Compost (Cap, 2014), and compute the semantic connectivity for the heads of the respective compounds. This reduces the number of pairs from 12,663,686 to 8,294,725.

Argument-taking ratio

The argument-taking ratio is a measure that describes the likelihood of a noun to take an argument. In the English bridging resolver, this was computed with the help of the NomBank annotations. These manual annotations list, for every occurrence in the WSJ corpus, the arguments of the nouns. To compute the argument-taking ratio, one then simply has to divide the number of NomBank annotations for one noun by the total frequency of the noun in the corpus. This is only possible because both the ISNotes and the NomBank annotation were performed on the same corpus. For other languages, we need to derive the number of cases in which the noun takes an argument automatically. To do this, we define these patterns of modification/argumenthood:

1. PP-postmodification/PP argument :

N_{target} PREP (Det) (ADJ)* N
Türen im Haus (*doors in the house*)

2. NPgen arguments:

N_{target} (Det) (ADJ)* N
die Kinder der Frau (*the woman's kids*)

3. Possessive pre-modification:

POSS N_{target}
Ihr Ehemann (*her husband*)

We then divide the frequency of a noun in these constructions by the total frequencies of the noun in a large corpus. Again, we use the SdeWaC corpus to derive the argument-taking ratio scores. As in the computation of the semantic connectivity scores, we run

6. Bridging resolution

into sparsity issues due to infrequent compounds. Thus, we also apply the compound splitter, to get more stable ratios. The argument-taking ratios are compiled for the head of the noun if a compound split exists. This reduces the number of nouns from 5,527,197 to 2,335,293.

In the following section, we describe the rules and how we adapted them to German.

Rule 1: building parts The anaphor is a part of a building (e.g. *window, room, etc.*) and is not pre-modified by a common or proper noun. The antecedent is selected as the one with the highest semantic connectivity in the same or the previous two sentences.

(76) im Zimmer ... **Die Fenster** (*in the room ... the windows*)

We translated the nouns on the building list to German but found that there is no noun in the DIRNDL corpus that is present on the building list, i.e. this rule is not particularly suited for our domain. It is left in anyway as it could be relevant for other data.

Rule 2: relative person NPs The anaphor is contained in a list of relative nouns (e.g. *child, son, husband, etc.*), its argument-taking ratio is greater than 0.4 (meaning that it is not used generically, i.e. in *children like toys*, but typically appears with an argument (*husband of ...*). It is not modified by an adjective or a noun and does not contain an embedded PP or is not followed by a PP.

Antecedents must be in the same sentence or the two previous ones and must be either a proper noun and not a location, or a named entity tagged as a person, or a personal pronoun except second person *du (you)*.

(77) Martha ... **Ihr Mann** (*Martha ... her husband*)

Rule 3: GPE job titles The anaphor is on a list of official job titles for a country (e.g. *commissioner, secretary, etc.*). It does not contain a country adjective as in *der argentinische Außenminister (the Argentinian foreign secretary)* and does not contain/ is not followed by a PP or an organisation.

The antecedent is the most salient geopolitical entity in the document. Salience is determined by frequency in the document. In case of ties, the closest is chosen.

(78) Deutschland ... **Der Außenminister** (*Germany ... the foreign secretary*)

Rule 4: professional roles

(79) IBM ... **CEO Peter Müller**

(80) der SPD ... **Der Vorstand** (*SPD ... the executive board*)

The head of the anaphor appears on a list of professional roles, (like *manager, doctor*) and does not contain a country adjective, a PP, a proper name or an organisation. The most salient antecedent is chosen within the last four sentences. Salience is determined by frequency in the document.

Rule 5: percentage expressions

(81) 10% der Deutschen ... 5% (*10% of all Germans ... 5%*)

The anaphor is a percentage expression containing % or “Prozent”. As antecedent, the modifier expression of another percentage expression is chosen, e.g. *der Deutschen* in *10% der Deutschen*. This rule is not applicable to DIRNDL as these percentage expressions are indefinite.

Rule 6: other set members This rule is not applicable to our data as it is designed for indefinite anaphora. It is left unimplemented in the resolver, in case one wants to implement it for other corpora.

Rule 7: argument-taking ratio I The anaphor is a common noun phrase (non-modified and without arguments) with an argument-taking ratio over 0.4. The antecedent is determined by finding the closest similar modification in the document. For details, refer to Section 6.1.1.

Rule 8: argument-taking ratio II The anaphor is a definite, non-modified expression without arguments in subject position (where it is likely to either be coreferent or bridging) with an argument-taking ratio over 0.4. The antecedent is chosen as the entity with the highest semantic connectivity in the last three sentences.

New rules

In addition to adapting the rules from the English system to German, we also added two new rules, which are tailored to our domain of radio news.

6. Bridging resolution

Rule 9: country part-of It is common in our data that a country is introduced into the discourse and then some aspect related to the country or a part of the country is picked up later as a bridging anaphor.

(82) Australien ... **Die Regierung** (*Australia ... the government*)

(83) Japan ... **Die Westküste** (*Japan ... the west coast*)

Therefore, we introduce a new rule: if the anaphor is a non-demonstrative definite expression without adjectival or nominal pre-modification and without a PP modification or argument that occurs on our list of country parts, we search for the most salient country. Saliency is determined by frequency in the document, with the exception of the subject in the very first sentence, which overrides frequency in terms of saliency. The list of country parts consists of terms like *Regierung* (*government*), *Einwohner* (*residents*), etc.

Rule 10: high semantic connectivity Rule 10 is similar to Rule 8 in Hou et al. (2014), but without the constraint that the anaphor has to be in subject position. However, it must be a non-modified NP or PP without any arguments. If the semantic connectivity score to a previously introduced mention is higher than a certain threshold (15.0 in our experiments), it is proposed as the antecedent. The antecedent must appear in the last four sentences. The feature is designed to capture more general cases of bridging, which can be found by looking for a high semantic connectivity between the anaphor and the antecedent.

Post-processing

The rules are ordered and applied according to their precision. Due to PPs being markables in DIRNDL, it is sometimes the case that the antecedent is in principle correct, but because of errors in syntactic parsing or other constraints, the resolver chose a slightly different span, e.g. without the preposition or the determiner. We count these cases, where the difference consists only of a determiner or preposition, as correct. For example,

(84) *Ägypten* if embedded in *in Ägypten* should also count as correct.

| Rule | Anaphor recognition | | | Full bridging resolution | | |
|---------|---------------------|-------|-----------|--------------------------|-------|-----------|
| | Correct | Wrong | Precision | Correct | Wrong | Precision |
| Rule 4: | 3 | 0 | 100.0 | 1 | 2 | 33.3 |
| Rule 8: | 26 | 31 | 45.6 | 13 | 44 | 22.8 |
| Rule 9: | 29 | 2 | 93.5 | 21 | 10 | 67.7 |
| Rule 10 | 57 | 40 | 58.8 | 25 | 72 | 25.7 |

Table 6.20.: Bridging resolution on DIRNDL: precision of the firing rules

| Corpus | Anaphor recognition | | | Full bridging resolution | | |
|------------------|---------------------|------|------|--------------------------|------|------|
| | P | R | F1 | P | R | F1 |
| Whole corpus | 61.2 | 17.6 | 27.3 | 31.9 | 9.2 | 14.2 |
| Test corpus | 60.6 | 21.3 | 31.4 | 38.3 | 13.6 | 20.1 |
| Train/dev corpus | 62.6 | 16.1 | 25.6 | 30.0 | 7.7 | 12.3 |

Table 6.21.: Bridging resolution on DIRNDL: overall performance

6.5.2. Performance

Table 6.20 shows the performance of the single rules when being applied to DIRNDL. From the original English system, only Rule 4 (GPE job titles) and the very general Rule 8 (which is based on semantic connectivity) fire. Our new rules also propose pairs. Rule 9 is rather specific and therefore has a high precision, while Rule 10 proposes a lot of pairs with mediocre precision, as it was designed to increase recall.

Most of the rules transferred from the English bridging resolver do not predict any bridging pairs in our data. For some cases, this can be explained by the different bridging definitions and guidelines, such as the fact that there are no indefinite bridging anaphors in our data. Rule 6, for example, which is designed to resolve anaphors containing a number expression or indefinite pronouns, cannot propose any correct pairs.

Of course, ISNotes, the corpus on which the experiments in the English bridging resolver were based on, and DIRNDL are also of slightly different domains (news text in ISNotes vs. radio news in DIRNDL), which might explain some of the differences.

Table 6.21 presents the performance of the overall system for anaphor detection and full bridging resolution. Overall, we achieve an F1 score of 14.2% for full bridging resolution with a precision of 31.9% and 9.2% recall. Surprisingly, the performance on the test corpus is better than on the development set.

Gold vs. predicted markables

| Setting | Precision | Recall | F1 |
|--------------------|-----------|--------|------|
| Predicted mentions | 29.9 | 9.2 | 14.0 |
| Gold mentions | 31.9 | 9.2 | 14.2 |

Table 6.22.: Bridging resolution on DIRNDL: predicted vs. gold mentions

Table 6.22 shows the scores for full bridging resolution for predicted and gold markables. As can be seen, the precision is slightly lower for predicted mentions. However, as the annotations on DIRNDL were performed on an earlier version of automatic syntactic annotations, the difference is only small and not statistically significant in this case.

Bridging resolution with gold coreference

| Setting | Precision | Recall | F1 |
|-----------------------|-----------|--------|-------------|
| No coreference | 21.4 | 9.2 | 12.8 |
| Predicted coreference | 22.4 | 9.2 | 13.0 |
| Gold coreference | 31.9 | 9.2 | 14.2 |

Table 6.23.: Bridging resolution with different types of coreference information in DIRNDL (using gold markables)

In the bridging system above, gold coreferent entities are removed from the list of potential anaphors. In a purely automatic system, this information is of course not available but could be approximated using a state-of-the-art coreference resolver. To test the difference it makes when we use predicted vs. gold vs. no coreference information at all, we experiment with different coreference settings. To test the effect of coreference information, we also run the system without filtering out coreferent anaphors. For the predicted version, we used the coreference system IMS HotCoref DE as described in Section 5, applying the default model trained on TüBa-D/Z on our data. In Table 6.23, we show that, as expected, the precision and F1 score are significantly higher in the

setting with coreference.²⁰ Predicted coreference, however, still improves precision (and the final F1 score) a little bit.

Error analysis

We found that there are a number of issues that affect the performance of the system. They are discussed in the following.

Preprocessing There are a couple of cases where a markable does not have an NP or PP annotated in the automatic constituency parse (due to parsing errors). No annotated NP means that we do not have the expression available as a markable in our experiments.

Abstract anaphors Bridging anaphors typically have nominal antecedents, but in some cases, they refer back to a VP or clausal antecedent. In these cases, we cannot find the right antecedent as our system only considers nominal antecedents.

Span mismatch When using predicted markables, there are some cases where there is an overlap between the gold and the predicted antecedent, but in the evaluation, they are considered wrong. More sophisticated evaluation metrics for bridging resolution would help here.

(85) auch in der Hauptstadt Tokio_{annotated span PP} (*also in the capital Tokio*)

Wrong annotations We found some annotations in DIRNDL that are against the RefLex guidelines, e.g. where annotators have marked single nouns or NPs that are embedded in PPs instead of PPs. These cannot be resolved correctly by the system.

Information status category generic As mentioned above, the information status category **generic** is present in DIRNDL, but not in the newest guidelines. This means that some bridging anaphors are labelled as **generic** rather than **bridging** as they are generic entities. In the newest version, **generic** is an attribute that bridging NPs (and other categories) can have.

²⁰We compute significance using the Wilcoxon signed rank test (Siegel and Castellan, 1988) at the 0.05 level.

Indefinites Many of the rules in Hou (2016b) focus on (indefinite) **part-whole** relations: in DIRNDL, these indefinite **part-whole** cases are not annotated on the referential level, so they are not contained as bridging pairs. They are, however, included in the lexical layer of RefLex, which could in principle also be used in the bridging experiments. However, the **part-whole** annotations here are based on the word level (in contrast to the NP level in the referential layer), where anaphoricity is not a criterion. For example, *wheel* would be marked as a part of a **car** on the word level, and we do not know whether this was actually a context-dependent case. Thus it is non-trivial to infer bridging pairs from the lexical-level part-whole annotations, as we do not know which of the part-whole relations between two words also contains an anaphoric relation between the two markables.

6.6. Conclusion

We have implemented a state-of-the-art bridging resolver for English. In our experiments, we have made a couple of observations. First, filtering out coreference anaphors before resolving bridging anaphors helps to increase the performance as coreference and bridging anaphors are difficult to distinguish. When applying the system on BASHI, we found that the bridging resolver as described in Hou et al. (2014) generalises well to other in-domain data, if they contain similar bridging annotations. In experiments with SciCorp we found that most of the rules are rather domain-specific and do not generalise well to other domains. The two more general rules also do not work as well on other domains because the two resources on which they are based, the argument-taking ratio list and the semantic connectivity scores, are computed on the basis of GigaWord, which surprisingly does not seem to contain many of the words that appear in the scientific articles. Adding some domain-specific data to these resources would certainly make the approach more applicable to the new domains.

When working with the ARRAU corpus we realised that there are very different understandings of bridging that have not been addressed as such in previous research. Our bridging characterisation thus distinguishes referential bridging, lexical and subset bridging as three rather different types of bridging that also have different properties. Non-identical anaphoricity is the main criterion for referential bridging, while lexical and subset bridging can also occur with non-anaphoric expressions.

After this theoretical contribution, we have focused on setting up a well-performing system on the ARRAU corpus, since the first shared task on bridging used this dataset

for the evaluation of the submitted systems. Therefore, we have implemented many new rules to also deal with lexical and non-anaphoric subset bridging. As our system was the only participating system, our results on ARRAU are the best published results on this dataset so far.

Finally, we have extended the bridging resolver to German by adapting the eight rules as well as implementing two new rules. The system was tested on the DIRNDL corpus and achieved similar results than the English resolver on ISNotes and BASHI. Again, the positive effect of removing coreference anaphors could be shown.

Overall, the performance of the openly available systems lies between 14 and 18% F1 score, on newspaper text. Of course, this means that there is still a lot of room for improvement. One of our linguistic validation experiments will thus be about integrating automatically predicted semantic relations into bridging resolution.

Part III.

Linguistic validation experiments

7. Using prosodic information to improve coreference resolution

Research Question 4: Linguistic validation experiments

With tools and data being available, do theoretical assumptions about the tasks hold true on actual data? Can we use the theoretical notions to improve the tools?

Now that we have developed tools and created data for coreference and bridging resolution, we can use the tools to validate theoretical claims about the task that have been made in theoretical or experimental studies. We will present two experiments that give examples of how the tools can be used. If the theoretical assumptions hold true, the tools' performances should benefit from the inclusion of the newly integrated information.

In our first experiment, described in this chapter, we examine the effect of prosodic features on coreference resolution in spoken discourse. We test features from different prosodic levels and investigate which strategies can be applied to include prosodic information in coreference. We also perform experiments on whether including prosodic boundaries and determining whether the accent is the nuclear accent is beneficial for the task.

We perform experiments using manually annotated and automatically predicted prosodic information. Our study deals with German data, but the prosodic properties are comparable to other West Germanic languages, like English or Dutch. Figure 7.1 shows our contributions in this step. Parts of this research have been published in Rösiger and Riester (2015) and Rösiger et al. (2017).

7.1. Motivation

In Example (1), taken from Umbach (2002), the question for the coreference resolver, besides linking the anaphoric pronoun *he* back to *John*, is to decide whether *an old*

7. Using prosodic information to improve coreference resolution

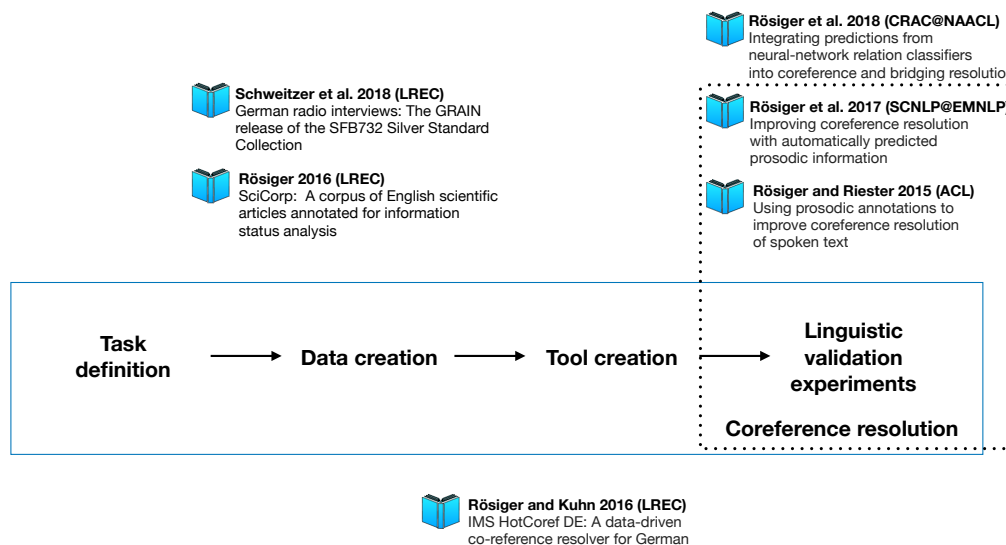


Figure 7.1.: Contribution and workflow pipeline for coreference: validation, part 1

cottage and *the shed* refer to the same entity. The problem here is that the transcript of this little snippet is ambiguous: even for humans (without further context), it remains unclear whether *the shed* is only a part of *the cottage* or whether the two expressions are used as synonyms.

- (1) {John}₁ has {an old cottage}₂.
Last year {he}₁ reconstructed {the shed}_?.

Almost all work on coreference resolution is based on text, although there exist a few systems for pronoun resolution in transcripts of spoken text (Strube and Müller, 2003; Tetreault and Allen, 2004). It has been shown that there are differences between written and spoken text that lead to a drop in performance when coreference resolution systems developed for written text are applied on spoken text (Amoia et al., 2012). For this reason, it may help to use additional information available from the speech signal, for example prosody.

In West-Germanic languages, such as English and German, there is a tendency for coreferent items, i.e. entities that have already been introduced into the discourse (their information status is *given*), to be deaccented, as the speaker assumes the entity to be salient in the listener’s discourse model (cf. Terken and Hirschberg (1994); Baumann and Riestler (2013); Baumann and Roth (2014)). We can make use of this fact by provid-

ing prosodic information to the coreference resolver. Example (2), this time marked with prominence information, shows that prominence can help us resolve cases where the transcription is potentially ambiguous. The accented syllables in the example are capitalised. Coreferent anaphors are marked in bold face.

- (2) {John}₁ has {an old cottage}₂.
 a. Last year {he}₁ reconstructed {the SHED}₃.
 b. Last year {he}₁ reconSTRUCted **the shed**₂.

The pitch accent on *shed* in (2-a) leads to the interpretation that *the shed* and *the cottage* refer to different entities, where the shed is a part of the cottage (they are in a bridging relation). In contrast, in (2-b), *the shed* is deaccented, which suggests that *the shed* and *the cottage* corefer.

7.2. Background

Pitch accents Pitch accents are changes in fundamental frequency, often combined with an increased intensity or longer duration. In West-Germanic languages, accentuation is used as a means to emphasise something. There are different shapes that describe the change in fundamental frequency, such as a rise or a fall. Figure 7.2 shows the change in fundamental frequency for one exemplary pitch accent type. The shapes are typically described with the help of the so-called ToBI labels, Tones and Break Indices, where the accent type categories consist of (sequences of) high and low targets, H and L. The GToBI(S) guidelines for German by Mayer (1995) for example distinguish the following categories: H*L, L*H,!H*L,H*,L*,HH*L and L*HL.

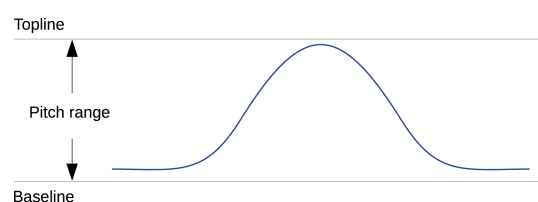


Figure 7.2.: One exemplary pitch accent shape

GToBI(S) stands in the tradition of autosegmental-metrical phonology, cf. Pierrehumbert (1980), Gussenhoven (1984), Féry (1993), Ladd (2008), Beckman et al. (2005). Speakers mainly make use of pitch accents and prosodic phrasing. The annotations distinguish intonation phrases, terminated by a major boundary (%), and intermediate

7. Using prosodic information to improve coreference resolution

phrases, closed by a minor boundary (-), as shown in Figure 7.3. As such, they make up a hierarchy: intonation phrases (IP), terminated by a major boundary (%) contain intermediate phrases (ip), which are closed by a minor boundary (-).

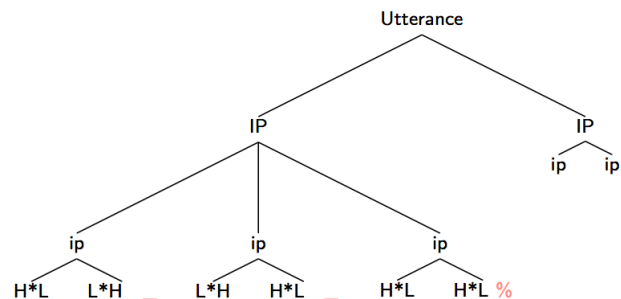


Figure 7.3.: The relation between phrase boundaries and intonation and intermediate phrases

The available pitch accent and boundary annotations allow us to automatically derive a secondary layer of prosodic information which represents a mapping of the pitch accents onto a prominence scale in which the nuclear (i.e. final) accents of an intonation phrase (n2) rank as the most prominent, followed by the nuclear accents of intermediate phrases (n1) and pre-nuclear (i.e. non-final) accents which are perceptually the least prominent. To put it simply, the nuclear accent is the most prominent accent in a prosodic phrase while pre-nuclear accents are less prominent. See Figure 7.4 for the relation between nuclear accents and boundary annotations.

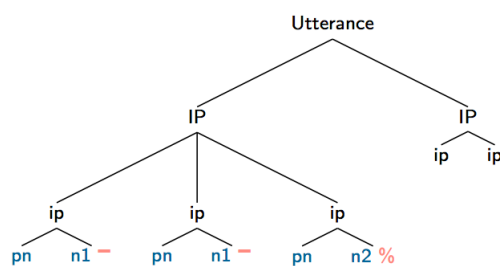


Figure 7.4.: The relation between boundary tones and nuclear and pre-nuclear accents

Pitch accents and coreference Many theoretical and experimental studies have shown that there is a tendency for coreferent items, i.e. entities that have already been introduced into the discourse, to be deaccented, as the speaker assumes the entity to be salient

(3) Anaphoric complex NP (DIRNDL sentences 9/10):

- 9: Im Mittelpunkt steht eine von der Ratspräsidentin, Bundeskanzlerin Merkel, vorbereitete “Berliner Erklärung”.
- 10: Die Präsidenten [...] wollen [den TEXT über die ZIEle und ZUkunft der EU] unterzeichnen.
 the presidents [...] want [the text about the aims and future the EU] sign
 ((L*H L*(H-) (H*L H*L H*L -)%)
 pn n1 pn pn

*Central is the ‘Berlin Declaration’ that was prepared by the president of the Council of the EU, Chancellor Merkel.
 The presidents want to sign [the text about the aims and future of the EU.]*

(4) Non-anaphoric complex NP (DIRNDL sentences 2527/2528):

- 2527: Der Prozess um den Tod eines Asylbewerbers aus Sierra Leone in Polizeigewahrsam ist [...] eröffnet worden.
- 2528: [Wegen KÖRperverletzung mit Todesfolge und fahrlässiger Tötung] MÜSsen ...
 [Due assault with lethal consequence, and reckless homicide] must
 ((H*L L*(H-) (H*L -)%)
 pn n1 n2

*The trial about the death of an asylum seeker from Sierra Leone during police custody has started.
 Charges include [assault with lethal consequence, and reckless homicide], ...*

in the listener’s discourse (cf. Terken and Hirschberg (1994); Schwarzschild (1999); Crutenden (2006) for English or Baumann and Riester (2013); Baumann and Roth (2014); Baumann et al. (2015) for German).

While we expect the difference between the presence or absence of pitch accents to influence the classification of short NPs like in Example (1), we do not expect complex NPs to be fully deaccented. For complex NPs, we nevertheless hypothesise that the prosodic structure of coreferential NPs will turn out to significantly differ from the structure of discourse-new NPs such as to yield a measurable effect. Examples (3) and (4) show the prosodic realisation of two expressions with different information status. In Example (3), the complex NP *the text about the aims and future of the EU* refers back to *the Berlin Declaration*, whereas in Example (4), the complex NP *assault with lethal consequences and reckless homicide* is not anaphoric. The share of prenuclear accents is higher in the anaphoric case, which indicates lower overall prominence.

7.3. Related work

Baumann and Riester 2013 Baumann and Riester (2013) examined the question whether different types and degrees of givenness trigger different prosodic markings. The paper discusses the prosodic realisation of referential expressions in annotated corpora

7. *Using prosodic information to improve coreference resolution*

of read and spontaneous speech, with a focus on the relation between information status and accent position as well as accent type.

Their starting point is based on the two-level RefLex scheme. They claim that givenness can occur on (i) the referential level: coreference with an antecedent already introduced into the discourse (referential givenness) or (ii) the lexical level: availability of a lexical unit in the discourse (lexical givenness).

They study the prosodic realisations of different referential and lexical category combinations and confirm the relevance of both the referential and lexical level. The data on read speech shows a tendency of a stepwise increase in prosodic prominence from given to new items. For spontaneous speech, the results are less clear.

As this thesis is concerned with anaphoricity and focuses on coreference and bridging anaphors, we will only analyse the relation between prosody and referential givenness, although we do think that givenness on the lexical level also plays a role, as already discussed in combination with comparative anaphors in Section 6.3.3

Amoia’s study on coreference in written and spoken text Amoia et al. (2012) described an empirical study of coreference in English spoken vs. written text, in which they aimed at defining specific parameters that classify differences in genres of spoken and written texts such as the preferred segmentation strategy, the maximal allowed distance or the length and the size of the coreference chains.

They also performed a precision-based evaluation on two corpora, one containing spontaneous interviews and one containing popular science texts, using the deterministic coreference system in the Stanford CoreNLP pipeline (Lee et al., 2011). The system achieved a MUC precision of 51% on spoken text, while on written text it achieved 64%. This confirms the results of previous work where coreference systems differ in their performance to process spoken vs. written text and that they perform better on written text, as this is also the type of text on which they typically are developed.

We think that improving the performance on spoken text by including prosodic features is thus a worthwhile effort.

7.4. Experimental setup

Data We use the DIRNDL corpus (Eckart et al., 2012; Björkelund et al., 2014) as it contains both manual coreference and manual prosody labels. We adopt the official train,

test and development split¹ designed for research on coreference resolution. The recorded news broadcasts in the DIRNDL corpus were spoken by 13 male and 7 female speakers, in total roughly 5 hours of speech. The prosodic annotations follow the GToBI(S) standard for pitch accent types and boundary tones (Mayer, 1995).

In the experiments where we use automatic predictions, we make use of two class labels of prosodic events: all accent types (marked by the standard ToBI *) grouped into a single class (pitch accent presence) and the same for intonational phrase boundaries (marked by %).

In the experiments based on manual prosodic information, we make use of both the simplified scheme and the fine-grained GToBI labels and phrase boundaries.

System and baseline We use the IMS HotCoref DE coreference resolver as a state-of-the-art coreference resolver for German, as described in Section 5. The standard features are text-based and consist mainly of string matching, part of speech, constituent parses, morphological information and combinations thereof.

As we aim at coreference resolution applicable to new texts, particularly in the setting using automatically predicted prosodic information, all annotations used to create the text-based features are automatically predicted using NLP tools. When training the system on the concatenation of the train and the development set of DIRNDL, as described in Section 5.2.6, we achieve a CoNLL score of 46.11. This will serve as a baseline in the following experiments.

7.5. Prosodic features

Our prosodic features mainly aim at definite descriptions, where it is difficult for the resolver to decide whether the potential anaphor is actually anaphoric or not. In these cases, accentuation is an important means to distinguish between given entities (often deaccented) and other categories (i.e. bridging anaphors or new information) that are typically accented, particularly for entities whose heads have a different lexeme than their potential antecedent. Pronouns are not the case of interest here, as they are (almost) always coreference anaphors.

Some of the features only take into account the absence or type of the pitch accent while others additionally employ prosodic phrasing. To get a better picture of the effect of these features, we implement, for each feature, one version for all noun phrases and

¹<http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/dirndl.en.html>

7. Using prosodic information to improve coreference resolution

a second version only for short noun phrases (<4 words). As explained above, this is to take into account the fact that longer phrases are rarely completely deaccented and are thus different from short NPs.

Two main features

The following two main features are tested in both the automatic and the manual setting.

Pitch accent presence focuses on the presence of a pitch accent, disregarding its type. If one accent is present in the markable, the Boolean feature gets assigned the value `true`, and `false` otherwise.

Nuclear accent presence is a Boolean feature comparable to pitch accent presence. It gets assigned the value `true` if there is a nuclear (`n2` or `n1`) accent present in the markable. In contrast to the first feature, this feature makes use of prosodic phrasing and takes the greater prominence of nuclear accents into account.

In the setting using manual prosodic information, we test a number of additional features.

Other features ignorant of phrase boundaries

Pitch accent type corresponds to the pitch accent types that are present in the GToBI(S) based annotations, as shown in Table 7.1. The types describe the shape of the change in fundamental frequency.

| Description | Label |
|---------------|-------|
| Fall | H*L |
| Rise | L*H |
| Downstep fall | !H*L |
| High target | H* |
| Low target | L* |
| Early peak | HH*L |
| Late peak | L*HL |

Table 7.1.: ToBI types in GToBI(S)

In case there are several ToBI types present, we look at the last label in the markable. As the ToBI types are not predicted in the automatic setting, we can only test the feature using manually annotated prosodic information.

Other features including phrase boundary information

The following set of features takes into account the degree of prominence of pitch accents, which encodes information about prosodic phrasing. How to determine and compare the overall prominence of complex NPs is an ongoing research question. The features described below are meant to test what works in an applied setting.

Nuclear accent presence (n2) is a variant of nuclear accent presence, where the boolean feature gets assigned the value `true` if there is a nuclear accent of type `n2` present in the markable. This is meant to be able to judge the helpfulness of the distinction between `n1` and `n2` accents. As this distinction is not contained in the automatic setting, it can only be tested using manual information.

Nuclear accent type looks at the different degrees of accent prominence. The markable gets assigned the type `n2`, `n1`, `pn` if the last accent in the phrase matches one of the types (and `none` if it is deaccented).

Nuclear bag of accents treats accents like a bag-of-words approach treats words: if one accent type is present once (or multiple times), the accent type is considered present. This means we get a number of different combinations ($2^3 = 8$ in total) of accent types that are present in the markable, e.g. `pn` and `n1` but no `n2` for Example (3), and `pn`, `n1` and `n2` for Example (4).

Nuclear: first and last includes linear information while avoiding an explosion of combinations. It only looks at the (degree of the) first pitch accent present in the markable and combines it with the last accent.

7.6. Manual prosodic information

We present a study on German spoken text that uses manual prosodic marking to show the principled usefulness of prosodic features for coreference resolution. In the long run and for application-based settings, of course, we do not want to rely on manual annotations. The features based on manual prosodic information investigate the potential of prominence information and are meant to motivate the use of automatic prosodic features, which we will also explore.

7. Using prosodic information to improve coreference resolution

To the best of our knowledge, this is the first work on coreference resolution in spoken text that tests the theoretical claims regarding the interaction between coreference and prominence in a general, state-of-the-art coreference resolver.

The manual prosodic information is taken from the DIRNDL corpus. We test all the features described above.

7.7. Automatically predicted prosodic information

Practical applications on spoken language need to rely on automatically predicted prosodic information, as manual labels are not only expensive but not applicable in an automatic pipeline setup.

In this section, we annotate the prosodic information automatically, thus omitting any manual annotations from the feature set. We predict pitch accents (and phrase boundaries) using a convolutional neural network (CNN) model from acoustic features extracted from the speech signal. We assess the quality of these annotations before we include them in the coreference resolver.

This part of the experiment was a collaboration between projects A6 and A8 of the SFB-732. The CNN classifier experiments and the resulting prosodic accents were provided by Sabrina Stehwien. The results of the collaboration were published in Rösiger et al. (2017).

In this section, we describe the prosodic event detector used in this work. It is a binary classifier that is trained separately for either pitch accents or phrase boundaries and predicts for each word, whether it carries the respective prosodic event.

We apply a CNN model, illustrated in Figure 7.5². The input to the CNN is a matrix spanning the current word and its right and left context word. The input matrix is a frame-based representation of the speech signal. The signal is divided into overlapping frames for each 20 ms with a 10 ms shift and is represented by a 6-dimensional feature vector for each frame.

We use acoustic features as well as position indicator features following Stehwien and Vu (2017) that are simple and fast to obtain. The acoustic features were extracted from the speech signal using the OpenSMILE toolkit (Eyben et al., 2013). The feature set consists of 5 features that comprise acoustic correlates of prominence: smoothed fundamental frequency (f0), root-mean-square (RMS) energy, loudness, voicing probability and Harmonics-to-Noise Ratio. The position indicator feature is appended as an extra

²Graphic provided by Sabrina Stehwien.

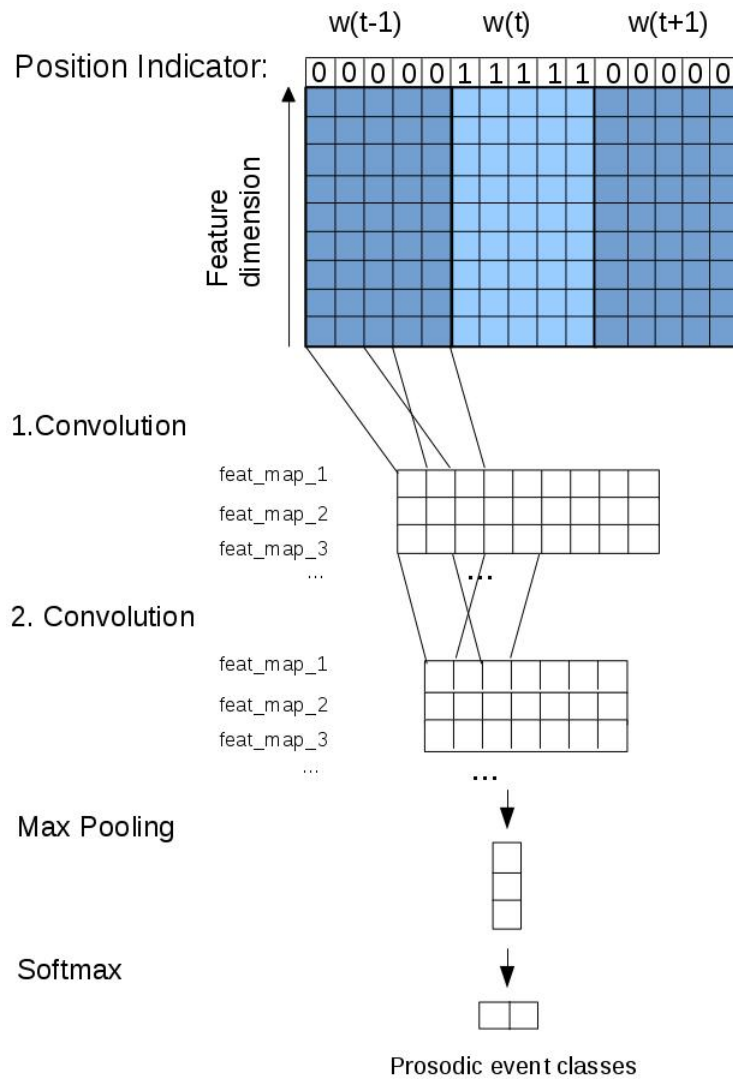


Figure 7.5.: CNN for prosodic event recognition with an input window of 3 successive words and position indicating features.

7. Using prosodic information to improve coreference resolution

feature to the input matrices (see Figure 7.5) and aids the modelling of the acoustic context by indicating which frames belong to the current word or the neighbouring words.

We apply two convolution layers in order to expand the input information and then use max pooling to find the most salient features. In the first convolution layer, we ensure that the filters always span all feature dimensions. All resulting feature maps are concatenated to one feature vector which is fed into the two-unit softmax layer.

Predicting prosodic labels on DIRNDL

We predict prosodic events for the whole DIRNDL corpus used in this paper. To simulate an application setting, we train the CNN model on a different dataset. Since the acoustic correlates of prosodic events, as well as the connection between sentence prosody and information status, are similar in English and German, we train the prosodic event detector on English data and apply the model to the German DIRNDL corpus.³ The data used to train the model is a 2.5 hour subset of the Boston University Radio News Corpus (Ostendorf et al., 1995) that contains speech from 3 female and 2 male speakers and that includes manually labelled pitch accents and intonational phrase boundary tones. Hence, both corpora consist of read speech by radio news anchors. The prediction accuracy on the DIRNDL anaphora corpus is 81.9% for pitch accents and 85.5% for intonational phrase boundary tones. The per-class accuracy is 82.1% for pitch accents and 37.1% for phrase boundaries. Despite these low-quality phrase boundary annotations, we believe that, as a first step, their effectiveness can still be tested. The speaker-independent performance of this model on the Boston dataset is 83.5% accuracy for pitch accent detection and 89% for phrase boundary detection. We conclude that the prosodic event detector generalises well to the DIRNDL dataset and the obtained accuracies are appropriate for our experiments.

7.8. Results and discussion

We test our prosodic features by adding them to the feature set used in the baseline. We define short NPs to be of length 3 or shorter. In this setup, we apply the feature only to short NPs. In the all NP setting, the feature is used for all NPs. The ratio of short vs. longer NPs in DIRNDL is roughly 3:1. Note that we evaluate on the whole test set

³Rosenberg et al. (2012) report good cross-language results of pitch accent detection on this dataset.

in both cases. We report how the performance of the coreference resolver is affected in three settings:

- (a) trained and tested on manual prosodic labels (short gold),
- (b) trained on manual prosodic labels, but tested on automatic labels (short gold/auto) (this simulates an application scenario where a pre-trained model is applied to new texts) and
- (c) trained and tested on automatic prosodic labels (short auto).

We predict the presence of a pitch accent and use phrase boundaries to derive nuclear accents, which are taken to be the last (and perceptually most prominent) accent in an intonation phrase. We do not predict the pitch accent type (e.g. fall H*L or rise L*H) as this distinction is generally difficult to model in the automatic setting. We will perform experiments based on manual labels using pitch accent type as a feature later.

We hypothesise the following:

Short NPs Since long, complex NPs almost always have at least one pitch accent, the presence and the absence of a pitch accent is more helpful for shorter phrases.

Long NPs For long, complex NPs, we look for nuclear accents that indicate the phrase’s overall prominence. If the NP contains a nuclear accent, it is assumed to be less likely to take part in coreference chains.

Table 7.2 shows the effect of the pitch accent presence feature on our data. All features perform significantly better than the baseline.⁴ As expected, the numbers are higher if we limit this feature to short NPs. We believe that this is due to the fact that the feature contributes most when it is most meaningful: on short NPs, a pitch accent makes it more likely for the NP to contain new information, whereas long NPs almost always have at least one pitch accent, regardless of their information status.

We achieve the highest performance using manual annotations (gold), followed by the version that has been trained on manual annotations and tested on automatically predicted prosodic labels (gold/auto), with a score that is not significantly worse than

⁴We compute significance using the Wilcoxon signed rank test (Siegel and Castellan, 1988) at the 0.01 level.

7. Using prosodic information to improve coreference resolution

the gold version. This is important for applications as it suggests that the loss in performance is small when training on gold data and testing on predicted data. As expected, the version that is trained and tested on predicted data performs worse but is still significantly better than the baseline. Hence, prosodic information is helpful in all three settings. It also shows that the assumption on short NPs is also true for automatic labels.

Table 7.3 shows the effect of adding nuclear accent presence as a feature to the baseline. Again, we report results that are all significantly better than the baseline. The improvement is largest when we apply the feature to all NPs, i.e. also including long, complex NPs. When restricted to only nuclear accents, the presence of an accent feature will receive the value `true` for only a few of the short NPs that would otherwise have been assigned `true` in terms of general pitch accent presence. Therefore, nuclear pitch accents do not provide sufficient information for a majority of the short NPs. For long NPs, however, the presence of a nuclear accent is more meaningful, as these tend to always have at least one accent.

The performance of the nuclear accent presence feature follows the pattern present for pitch accent presence: `gold > gold/auto > auto`. Again, automatic prosodic information contributes to the system’s performance.

The highest CoNLL score when using automatic labels is 50.64, as compared to 53.99 with gold labels. To the best of our knowledge, these are the best results reported on the DIRNDL anaphora dataset so far.

| Baseline | 46.11 | |
|----------------------|--------------|---------|
| + Accent | short NPs | all NPs |
| + Presence gold | 53.99 | 49.68 |
| + Presence gold/auto | 52.63 | 50.08 |
| + Presence auto | 49.13 | 49.01 |

Table 7.2.: Performance of pitch accent presence (in CoNLL score)

More detailed experiments based on manual annotations

We perform some additional experiments where we further investigate the use of prosodic boundaries and the use of certain ToBI or nuclear types. As the prediction quality of the boundaries was rather low (37.1% precision) and ToBI types are difficult to predict automatically, we base these experiments on manual annotations.

| Baseline | 46.11 | |
|----------------------|-----------|--------------|
| + Nuclear accent | short NPs | all NPs |
| + Presence gold | 48.63 | 52.12 |
| + Presence gold/auto | 48.46 | 51.45 |
| + Presence auto | 48.01 | 50.64 |

Table 7.3.: Performance of nuclear accent presence (in CoNLL score)

Table 7.4 examines the effect of the respective new features in terms of the CoNLL scores. Features that achieved a significant improvement over the baseline are marked with a star.

As can be seen, features based on GToBI(S) accent type (pitch accent type) did not result in any significant improvements.

| Baseline | 46.11 | |
|---|-----------|---------|
| + Accent | short NPs | all NPs |
| + Pitch accent type | 45.31 | 46.23 |
| + Nuclear accent type (n1 vs. n2 vs. pn vs. none) | 47.17 | 46.79 |
| + Nuclear accent type (n1/n2 vs. pn vs. none) | 48.55* | 45.24 |
| + Nuclear accent presence (n2) | 46.69 | 48.88* |
| + Nuclear bag of accents | 46.09 | 48.45* |
| + Nuclear first+last | 46.41 | 46.74 |

Table 7.4.: Additional features based on manual prosodic information (gold setting)

In terms of features that are phonologically more informed, the picture is less clear. Distinguishing between prenuclear and nuclear accents (nuclear accent type) is a feature that works best for short NPs where there is only one accent. A significant increase in performance was achieved by distinguishing nuclear (n1/n2) vs. prenuclear accents. Distinguishing n1 and n2 accents did not lead to significant improvements.

Nuclear accent presence of an n2 accent, on the other hand, works well for all NPs, but not as well as the more general nuclear presence in the main experiments.

The nuclear bag of accents feature works quite well, too: this is a feature designed for NPs that have more than one accent and so it works best for complex NPs. The feature Nuclear first+last did not lead to significant improvements.

7. Using prosodic information to improve coreference resolution

Overall, these features perform worse than the two main features accent presence and nuclear accent presence. Still, it becomes clear that one has to be very careful in terms of how the prosodic information is used. In general, the presence of an accent works better than the distinction between certain accent types, and including intonation boundary information also contributes to the system’s performance when applying the feature to all NPs, including complex NPs.

As the ratio of short vs. longer phrases in DIRNDL is 3:1, applying the feature only to short NPs without boundary information leads to the highest overall result (53.99). However, depending on the ratio of short and long NPs in other data, including the boundaries to also better treat complex NPs can be beneficial. The best version including prosodic boundaries and applying the feature to all NPs leads to a CoNLL score of 52.12.

To conclude, the overall best score was achieved by looking at the presence of an accent for short phrases. Here, the presence alone is a beneficial information to determine whether the markable is a coreference anaphor. The second best score was achieved by determining whether there is a nuclear accent contained in the markable, where these were not limited to short NPs.

For the two main features, the most important point is also that prosodic information was beneficial in every setting, whether it was based on manual or automatically predicted prosodic information.

Analysis

In the following section, we discuss two examples from the DIRNDL dataset that provide some insight as to how the prosodic features helped coreference resolution in our experiments.

The first example is shown in Figure 7.6. The coreference chain marked in this example was not predicted by the baseline version. With prosodic information, however, the fact that the NP *der Koalition (the coalition)* is deaccented helped the resolver to recognise that this was given information: it refers to the recently introduced antecedent *der Großen Koalition (the grand coalition)*. This effect clearly supports our assumption that the absence of pitch accents helps for short NPs.

An additional effect of adding prosodic information that we observed concerns the length of antecedents determined by the resolver. In several cases, e.g. in Example (5), the baseline system incorrectly chose an embedded NP (1A) as the antecedent for a pronoun. The system with access to prosodic information correctly chose the longer NP

EXPERTEN {der Großen KOALITION}₁ haben sich auf [...] ein Niedriglohn-
Experts (of) the grand coalition have themselves on a low wage

Konzept VERSTÄNDIGT. Die strittigen Themen [...] sollten bei der nächsten
concept agreed. The controversial topics shall at the next

Spitzenrunde {der Koalition}₁ ANGESPROCHEN werden.
meeting (of) the coalition raised be.

*EN: Experts within the the grand coalition have agreed on a strategy to address [problems associated with] low income. At the next meeting **the coalition** will talk about the controversial issues.*

Figure 7.6.: The relation between coreference and prominence: example from the DIRNDL dataset with English translation. The candidate NP (anaphor) of the coreference chain in question is marked in boldface, the antecedent is underlined. Pitch accented words are capitalised.

(1B).⁵ Our analysis confirms that this is due to the accent on the short NP (on *Phelps*). The presence or absence of a pitch accent on the adjunct NP (on *USA*) does not appear to have an impact.

(5) {{Michael PHELPS}}_{1A} aus den USA}}_{1B}. **{Er}**₁ ...
Michael Phelps from the USA. He ...

Further work is necessary to investigate the feature interaction and the impact on the length of the predicted antecedent.

7.9. Conclusion and future work

We have shown that enhancing the text-based feature set for a coreference resolver, consisting of e.g. automatic part-of-speech (POS) tags and syntactic information, with pitch accents and prosodic phrasing information helps to improve coreference resolution of German spoken text.

Our results on the basis of manual prosodic labelling show that the presence of an accent is a helpful feature in a machine-learning setting. Including prosodic boundaries and determining whether the accent is the nuclear accent also increases results when applying the feature to all NPs (including complex NPs).

⁵The TüBA-D/Z guidelines state that the maximal extension of the NP should be chosen as the markable.

<http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-coreference-manual-2007.pdf>

7. *Using prosodic information to improve coreference resolution*

We show that using prosodic labels that have been obtained automatically also significantly improves the performance of a coreference resolver. In this work, we predict these labels using a CNN model and use these as additional features. Despite the quality of the predicted labels being slightly lower than the gold labels, we are still able to achieve significant improvements. This encouraging result also confirms that not only is prosodic information helpful to coreference resolution but that it also has a positive effect even when predicted by a system. We interpret this as a promising result, which motivates further research on the integration of coreference resolution and spoken language.

As a first step, our results on German spoken text are promising and we expect them to be generalisable to other languages with similar prosody.

8. Integrating predictions from neural-network relation classifiers into coreference and bridging resolution

Research Question 4: Linguistic validation experiments

With tools and data being available, do theoretical assumptions about the tasks hold true on actual data? Can we use the theoretical notions to improve the tools?

The second validation experiment concerns both coreference and bridging resolution and investigates the question if automatically predicted semantic knowledge can be used to improve coreference and bridging resolution.

The most difficult cases in NP coreference are those which require semantic knowledge to infer the relation between the anaphor and the antecedent, as in Example (1), where we need to know that *Malaria* is a *disease*.

(1) Malaria is a mosquito-borne infection. **The disease** is transmitted via a bite ...

Bridging resolution always requires semantic information. For example, in order to resolve *the windows* in Example(2) to *the room*, we need to know that a room typically has windows. The relation can also be rather abstract, as shown in Example (3).

(2) I went into the room. **The windows** were broken.

(3) Over the first few weeks, Mancuso FBI has sprung straight from the headlines. **The opening show** featured a secretary of defense designate accused of womanizing.

The semantic relation information necessary for anaphora resolution is typically integrated into a system through a knowledge base, by relying on WordNet, Wikipedia or

8. Neural-net relation predictions for coreference and bridging resolution

similar resources (cf. Vieira and Poesio (2000), Ponzetto and Strube (2007), a.o.). Up to date, few approaches have tried to integrate automatically induced information about semantic relations (e.g. Poesio et al. (2002); Feuerbach et al. (2015)). In the current study, we suggest state-of-the-art neural-network classifiers trained on relation benchmarks to predict semantic relations between noun pairs, and integrate the relation predictions into existing systems for coreference and bridging resolution. Two experiments with representations differing in noise and complexity improve our bridging but not our coreference resolver. This work was a collaboration between projects A6 and the SemRel project headed by Sabine Schulte im Walde. The neural-net experiments as well as the resulting relation predictions were provided by Maximilian Köper and Kim-Anh Nguyen. Contributions in the respective pipelines are shown in Figure 8.1 and 8.2. Parts of this research have been published in Rösiger et al. (2018a).

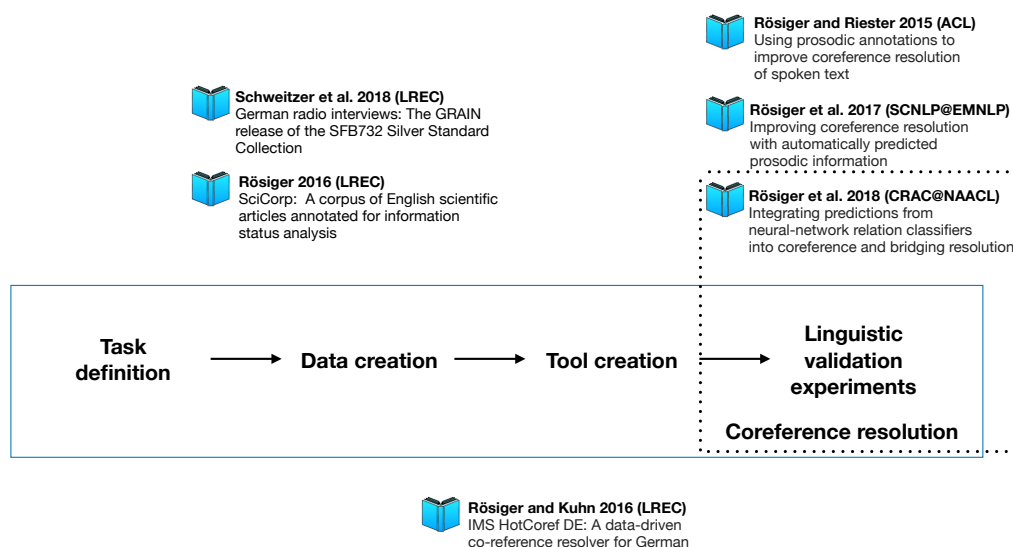


Figure 8.1.: Contribution and workflow pipeline for coreference: validation, part 2

8.1. Relation hypotheses

Coreference signals a relation of identity, so we assume that coreference resolution should benefit from relations that link identical or highly similar entities. Obviously, synonymy is a member of this set of relations, as exemplified in Example (4):

- (4) I live on Shortland Street. **The road** will be closed for repair work next week.

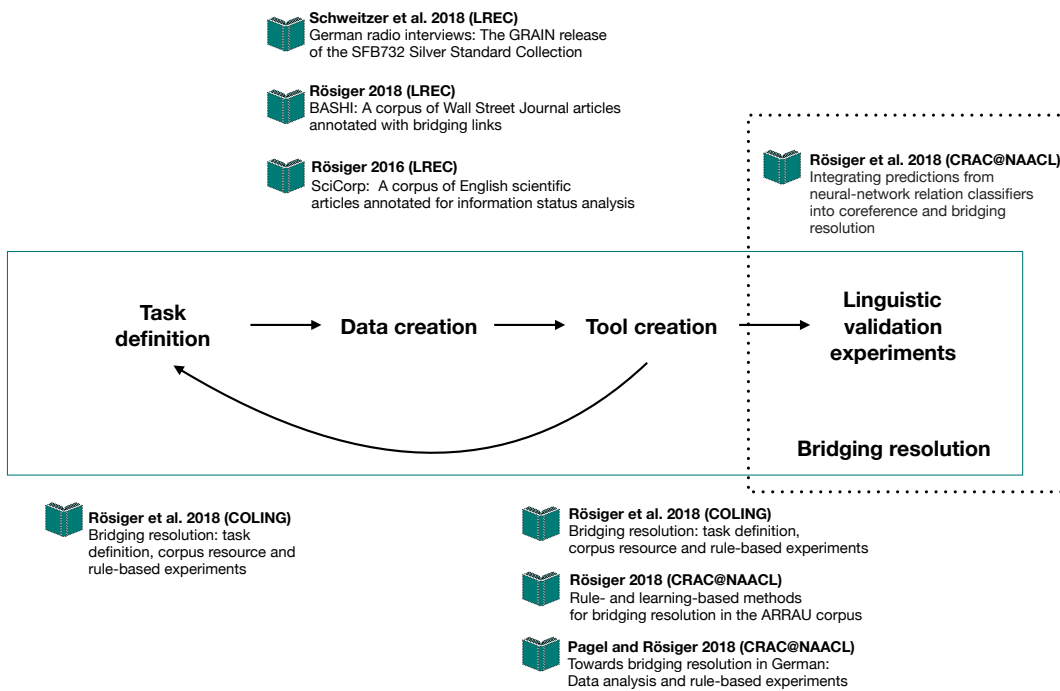


Figure 8.2.: Contribution and workflow pipeline for bridging: validation

Hyponymy can also be used to refer to a previously introduced entity, as in Example (5):

- (5) My neighbour's dog has been getting on my nerves lately. **The stupid animal** kept barking all night.

Note that the direction of this relation is important, as we can introduce a hyponym and then later refer to it via a hypernym, but not vice versa. Although, in news text, you might find a certain writing style which allows for hypernyms to later be referred to via a hyponym, e.g. in Example (6).

- (6) Today we are celebrating a great athlete. **The Olympic swimmer** has always been one of our personal favorites.

The relations between a bridging anaphor and its antecedent are assumed to be more diverse. The prototypical bridging relation is represented by meronymy:

- (7) My car broke down yesterday. It turned out to be a problem with **the engine**.

However, other relations come into play, too, such as attribute-of and part-of-event (Hou, 2016b).

8.2. Experimental setup

This section describes the data, tools and evaluation metrics used in the two experiments.

Data We base our experiments on the OntoNotes corpus (Weischedel et al., 2011). For bridging, we use the ISNotes corpus (Markert et al., 2012). In order to obtain candidate pairs for semantic relation prediction, we consider all heads of noun phrases in the OntoNotes corpus and combine them with preceding heads of noun phrases in the same document. Due to the different corpus sizes, the generally higher frequency of coreferent anaphors and the transitivity of the coreference relation, we obtained many more coreference pairs (65,113 unique pairs) than bridging pairs (633 in total, including 608 unique pairs).

Bridging resolver We base our experiment on our bridging resolver presented in Section 6.1. It contains eight rules which all propose anaphor-antecedent pairs, independently of the other rules. The rules are applied in order of their precision. Apart from information on the connectivity of two nouns, which is derived from counting how often two nouns appear in a “*noun₁ preposition noun₂*” pattern in a large corpus, the tool does not contain information about semantic relations.

Coreference resolver We use the IMS HotCoref resolver (Björkelund and Kuhn, 2014) as a coreference resolver, because it allows an easy integration of new features. While its performance is slightly worse than the state-of-the-art neural coreference resolvers, the neural resolvers rely on word embeddings, which already implicitly contain semantic relations.

Evaluation metrics For coreference resolution, we report the performance as CoNLL score, version 8.01 (Pradhan et al., 2014). For bridging resolution, we report performance in precision, recall and F1. For bridging evaluation, we take coreference chains into account during the evaluation, i.e. the predicted antecedent is considered correct if it is in the same coreference chain. We apply train-development-test splits, use the training and development set for optimisation, and report performance on the test set.

8.3. First experiment

8.3.1. Semantic relation classification

We used the publicly available relation resource BLESS (Baroni and Lenci, 2011), containing 26,546 word pairs across the six relations `co-hyponymy/coordination`, `attribute`, `meronymy`, `hypernymy`, and `random` (no relation). As classification method, we relied on the findings from Shwartz and Dagan (2016), and used a plain distributional model combined with a non-linear classifier (neural network) with only word representations. As many of our target word pairs rarely or never occurred together in a shared sentence, we could not integrate intervening words or paths as additional features.

We took the publicly available 300-dimensional vectors from ConceptNet (Speer et al., 2017), combined the word representations with the semantic relation resources, and trained a feed-forward neural network for classification. The input of the network is simply the concatenation of the two words, and the output is the desired semantic relation. At test time we present two words and output the class membership probability for each relation. In addition, we provide information about the semantic similarity by computing the cosine.

We relied on the training, test and validation split from Shwartz and Dagan (2016). The hyper-parameters were tuned on the validation set and obtained the best performance by relying on two hidden layers with 200 and 150 neurons respectively. As activation function, we applied rectified linear units (ReLU). We set the batch size to 100 and used a dropout rate of 20%.

In Figure 8.3¹, we present the neural-net classifier used in the first experiment. As can be seen, only the concatenated word representations are used as input. For the pair *dog-aquarium*, the classifier’s output is a high membership degree for the class `random`, i.e. it considers the two pairs to be non-related. Another output that is computed directly from the word representations is the cosine similarity, which is low in this case.

Figure 8.4 shows the output of the same neural net for the pair *dog-animal*. The output contains a high cosine similarity and a high membership degree for the relation `hypernymy`.

Intrinsic Evaluation To validate that the semantic relation classification works to a sufficient degree, we performed an intrinsic evaluation. On the test set from Shwartz

¹The neural-net illustrations in this section were provided by Maximilian Köper.

8. Neural-net relation predictions for coreference and bridging resolution

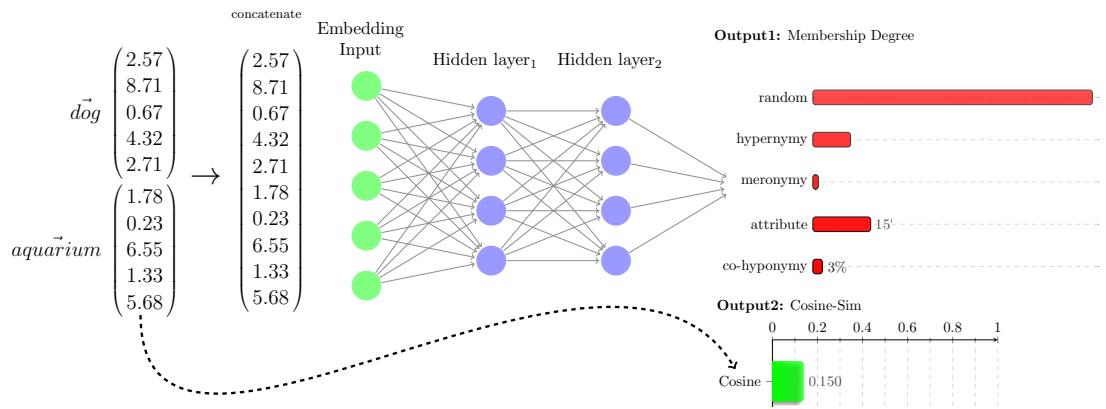


Figure 8.3.: Neural net relation classifier: example of a non-related pair

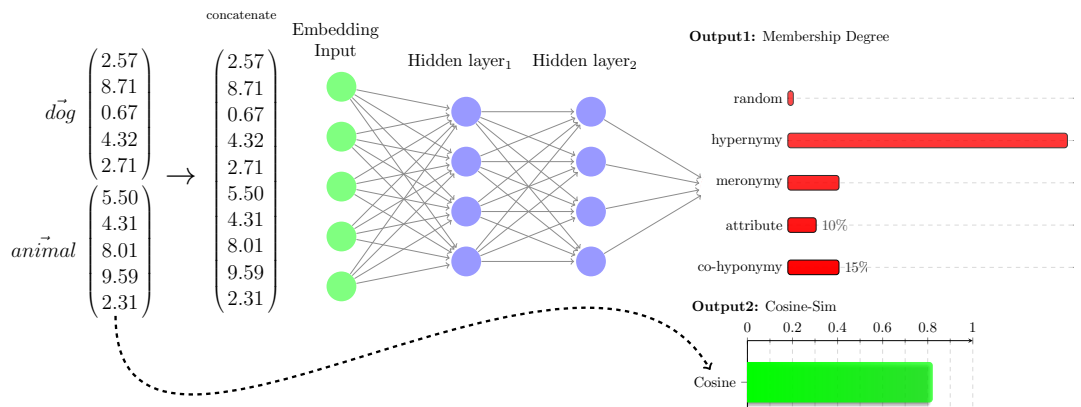


Figure 8.4.: Neural net relation classifier: example of a hypernym pair

and Dagan (2016), our model achieved an accuracy of 87.8%*, which is significantly² better than the majority class baseline (i.e. the random class with 45%). Shwartz and Dagan report a weighted average F-score of 89, which is only marginally better than our reimplementation (88).

While this performance seems very good and confirms the quality of our reimplementation, the work by Levy et al. (2015) pointed out that such supervised distributional models often just memorise whether a word is a prototypical example for a certain relation. Indeed, we found many of these cases in our dataset. For example, the term ‘gas’ appeared $\frac{9}{10}$ times in a meronym relation in training and $\frac{4}{4}$ times as a meronym in the test set. To counter this effect we conducted a second evaluation where we made sure that the training and test set contained different terms.

With an accuracy of 58.6%* and a weighted mean F-score of .52, the performance of this second evaluation was still significantly better than the majority class baseline but considerably worse than the reported results on the BLESS train/test split with lexical overlap. Still, we assume that this evaluation provides a more realistic view of the relation classification. Results per relation are given in Table 8.1. It can be seen that the model is skewed towards the majority class (**random**), whereas in particular the **hypernym** relation seems to be difficult. Here we observed many false decisions between **coordination** and **hypernymy**.

| Relation | P | R | F1 |
|----------|------|------|------|
| Random | 63.7 | 93.8 | 75.9 |
| Coord | 46.6 | 41.2 | 43.7 |
| Attri | 68.9 | 18.7 | 29.4 |
| Mero | 31.1 | 22.4 | 26.0 |
| Hyper | 25.0 | 0.4 | 0.7 |

Table 8.1.: Results of the intrinsic evaluation on BLESS (without lexical overlap)

8.3.2. Relation analysis

Before using the predicted relations for coreference and bridging resolution, we analysed the distribution of relations across the bridging and coreference pairs annotated in our corpora, as well as across all other, non-related pairs. Table 8.2 shows the average cosine similarities (COS) of these pairs. As expected, the average cosine similarity is

²We used the χ^2 test * with $p < 0.001$.

8. Neural-net relation predictions for coreference and bridging resolution

highest for coreference pairs and a little lower for bridging pairs, but still much higher in comparison to all other pairs. In the rows below cosine similarity, we give the averages of the output probabilities of the classifier for each relation. **Random** represents the class for non-related pairs without a relation. Such non-related pairs have indeed a high score for not being in a relation, whereas coreference and bridging pairs have lower scores in this category. Non-related random pairs have a high score for not being in a relation, whereas coreference and bridging pairs have lower scores in this category. Both coreference and bridging pairs have high **meronym** values, which is surprising for the coreference pairs. Bridging pairs also have a higher **coordination** value (i.e. **co-hyponymy**), and a slightly higher value for **hypernymy**.

| | Coref pairs | Bridging pairs | Other pairs |
|--------|-------------|----------------|-------------|
| COS | 0.26 | 0.19 | 0.05 |
| Random | 0.39 | 0.49 | 0.78 |
| Coord | 0.22 | 0.13 | 0.03 |
| Attri | 0.07 | 0.07 | 0.06 |
| Mero | 0.22 | 0.23 | 0.10 |
| Hyper | 0.09 | 0.07 | 0.02 |

Table 8.2.: Average cosine similarities and relation classifier probabilities for coreferent and bridging pairs in comparison to other pairs of nouns, experiment 1

8.3.3. Relations for bridging resolution

| Baseline | - | - | - | - | - | - | - | - | 59.82 | 10.58 | 18.0 |
|------------|---------------------------------|-------|-----------|--------|-------|-------------------------------------|-------|-----------|--------|--------------|------|
| Relation | <i>without cosine threshold</i> | | | | | <i>with cosine threshold of 0.2</i> | | | | | |
| | Correct | Wrong | Precision | Recall | F1 | Correct | Wrong | Precision | Recall | F1 | |
| Coord | 5 | 41 | 45.57 | 11.37 | 18.20 | 5 | 32 | 48.3 | 11.37 | 18.41 | |
| Attri | 3 | 46 | 43.48 | 11.06 | 17.63 | 2 | 8 | 56.56 | 10.9 | 18.28 | |
| Mero | 14 | 101 | 35.69 | 12.80 | 18.84 | 14 | 36 | 50.00 | 12.80 | 20.38 | |
| Hyper | 2 | 7 | 57.02 | 10.90 | 18.3 | 2 | 4 | 58.47 | 10.9 | 18.38 | |
| Not random | 17 | 105 | 35.90 | 13.27 | 19.37 | 15 | 54 | 45.3 | 12.95 | 20.15 | |

Table 8.3.: Correct and wrong bridging pairs which are found by the additional semantic rule, with and without additional cosine threshold constraint (> 0.2)

As short, unmodified NPs are generally considered useful bridging anaphor candidates, because they often lack an antecedent in the form of an implicit modifier, we add the following new rule to our bridging resolver: search for an unmodified NP, in the form

| Threshold | Correct | Wrong | P | R | F1 |
|-----------|---------|-------|-------|-------|--------------|
| 0.15 | 16 | 56 | 44.20 | 12.64 | 19.66 |
| 0.20 | 14 | 36 | 50.00 | 12.80 | 20.38 |
| 0.25 | 10 | 26 | 52.03 | 12.16 | 19.72 |
| 0.30 | 2 | 22 | 50.74 | 10.90 | 17.95 |

Table 8.4.: Effect of the cosine threshold constraint, for the relation meronymy

of “*the N*”, e.g. in *the advantages*. As bridging antecedents typically appear in a rather close window (Hou, 2016b), we search for an antecedent within the last three sentences. As bridging pairs have a higher cosine value than non-related pairs, we experiment with an additional cosine similarity constraint: if the pair is in a certain relation and the cosine similarity is greater than 0.2, it is proposed.

Table 8.3 shows the results for the different relations as well as the versions with and without a cosine similarity threshold, which are explored further in Table 8.4. Note that both tables do not give absolute numbers of correct and wrong bridging pairs, but only the bridging pairs which were proposed by the newly added semantic rule.

Meronymy seems to be the best predictor for bridging, with a significant gain of 2.38% in F1 score³, followed by the not-random version. The precision slightly decreased, but since the rule was designed to increase recall, this is acceptable. In the best setting (meronymy, cosine threshold of 0.2) we now find 14 additional correct pairs, for example:

- (8) IBM said it expects industrywide efforts to become prevalent because semiconductor manufacturing has become so expensive. A state-of-the-art plant cost 40 million in the mid-1970s but costs 500 million today because **the technology** is so complex.

We also find 36 more wrong pairs, for example:

- (9) In the 1980s, the Justice Department and lower federal courts that enforce the Voting Rights Act have required state legislatures and municipal governments to create the maximum number of “safe” minority election districts – districts where minorities form between 65% and 80% of **the voting population** .

³We compute significance using the Wilcoxon signed rank test (Siegel and Castellan, 1988) at the 0.05 level.

8. Neural-net relation predictions for coreference and bridging resolution

The reasons for a wrongly proposed candidate pair can be two-fold: (i) the relation predicted by the classifier is wrong and there is no actual relation between the two words or (ii) the relation predicted by the classifier is correct but the anaphoricity criterion is not given. As ISNotes contains solely referential bridging pairs, a meronymy relation alone is not sufficient for the annotation of a bridging pair.

8.3.4. Relations for coreference resolution

We used the following features in the resolver:

- *Random as the highest class*: a boolean feature which returns `true` if the random class got assigned the highest value of all the relations.
- *Cosine binned into low, middle, high*: this is a binned version of cosine similarity. We experimented with two different bins, the first one $\{0-0.3, 0.3-0.49, >0.49\}$, the second one $\{0-0.3, 0.3-0.6, >0.6\}$
- *Relation with the highest value*: a multi-value feature with 6 potential values: `none`, `mero`, `coord`, `attri`, `hyper` and `random`. The class with the highest value is returned.

We added one feature at a time and analysed the change in CoNLL score. The results are not shown in detail, as the score decreased in every version. For coreference resolution, where the baseline performance is already quite high, the additional semantic information thus does not seem to improve results. This is in line with Björkelund and Kuhn (2014), where integrating a WordNet synonym/hypernym lookup did not improve the performance, as well as Durrett and Klein (2013), where increased semantic information was not beneficial either.

8.4. Second experiment

The first experiment had a few major shortcomings. First, we did not have lemmatised vectors, and as a result, singular and plural forms of the same lemma had different values. Sometimes, this led to the wrong analysis, as in Example (10), where the singular and plural versions of *novel* make different predictions, and where a lemmatised version would have preferred the correct antecedent:

| Word 1 | Word 2 | COS | coord | attri | mero |
|------------|--------|------|-------------|-------|-------------|
| characters | novel | 0.35 | 0.69 | 0.02 | 0.27 |
| characters | novels | 0.43 | 0.28 | 0.05 | 0.38 |

- (10) In novels of an earlier vintage_{predicted}, David would have represented excitement and danger; Malcom, placid, middle-class security. The irony in this novel_{gold} is that neither man represents a “safe” middle class haven - : Nora’s decision is between emotional excitement and emotional security, with no firm economic base. **The characters** confront a world in which it seems increasingly difficult to find a “middle way” between the extremes of success and failure,

Second, many proper nouns were assigned zero values, as they were not covered by our vector representations. These pairs thus could not be used in the new rule. Third, the relations in the benchmark dataset BLESS do not completely match our hypotheses, as synonymy for example is not included. We thus designed a second experiment to overcome these shortcomings.

8.4.1. Semantic relation classification

To address the problem with out-of-vocabulary words we relied on fasttext (Bojanowski et al., 2017), which uses subword information to create representations for unseen words. We created 100-dimensional representations by applying a window of 5 to a lemmatised and lower-cased version of DECOW14 (Schäfer, 2015). The semantic relations were induced from WordNet (Fellbaum, 1998), by collecting all noun pairs from the relations: *synonymy*, *antonymy*, *meronymy*, *hyponymy*, *hypernymy*. To obtain a balanced setup, we sampled 2,010 random pairs from each relation, and in addition, we created random pairs without relations across files. Hyper-parameters of the neural network were identical to the ones used in the first experiment, as shown in Figure 8.5⁴.

Intrinsic evaluation We obtained a similar performance as before, an accuracy of 55.8%* (exp1: 58.6) and a mean weighted f-score of 55 (exp1: 52). Results per relation are shown in Table 8.5. Interestingly, the performances with respect to the individual relations differ strongly from the first experiment. In this second experiment, with

⁴Again, this graphic was provided by Maximilian Köper.

8. Neural-net relation predictions for coreference and bridging resolution

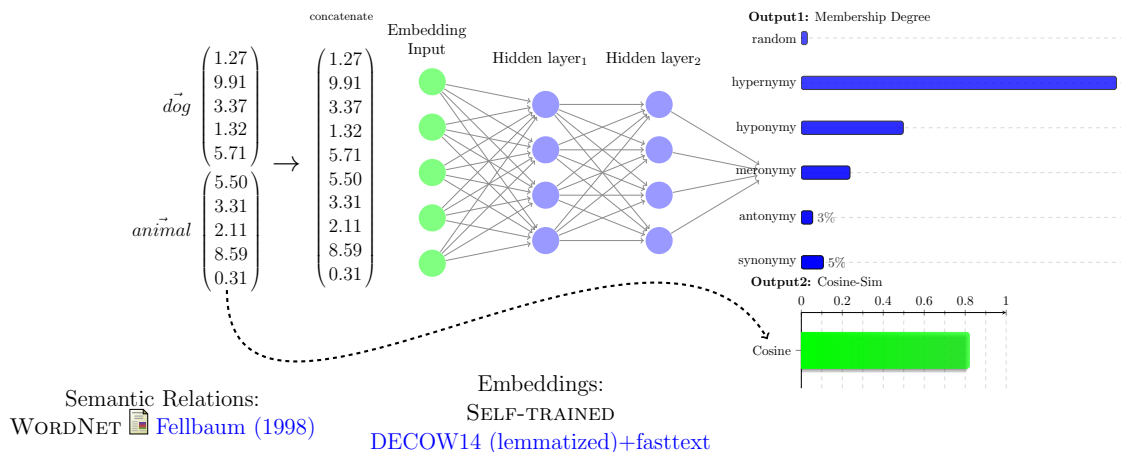


Figure 8.5.: Neural net relation classifier in the second experiment

balanced relations, *meronymy* and *antonymy* are well-detected whereas *random* performs inferior.

| Relation | P | R | F1 |
|----------|------|------|------|
| Random | 56.7 | 39.0 | 46.2 |
| Ant | 70.0 | 83.4 | 76.3 |
| Syn | 46.3 | 46.5 | 46.4 |
| Mero | 62.1 | 69.5 | 65.6 |
| Hyper | 48.9 | 49.1 | 49.0 |
| Hypo | 47.5 | 47.6 | 47.6 |

Table 8.5.: Results of the intrinsic evaluation on WordNet

8.4.2. Relation analysis

Table 8.6 shows that –unexpectedly– the probabilities of the coreference and bridging pairs in comparison to other pairs differ much less than in the first experiment.

8.4.3. Relations for coreference and bridging resolution

The two setups for integrating the relation classification into bridging and coreference resolution were exactly the same as in the first experiment. The outcome is, however, a little disappointing. The baseline system for bridging resolution was only improved in one condition, for the relation *meronymy* and with a cosine threshold of 0.3, reaching

| | Coref pairs | Bridging pairs | Other pairs |
|--------|-------------|----------------|-------------|
| COS | 0.38 | 0.31 | 0.22 |
| Random | 0.13 | 0.15 | 0.21 |
| Mero | 0.18 | 0.15 | 0.17 |
| Hyper | 0.25 | 0.23 | 0.23 |
| Hypo | 0.20 | 0.27 | 0.19 |
| Syn | 0.16 | 0.15 | 0.15 |
| Ant | 0.08 | 0.06 | 0.05 |

Table 8.6.: Average relation classifier probabilities and cosine similarities for coreferent and bridging pairs in comparison to other pairs of nouns, experiment 2

F1=18.92 (in comparison to F1=20.38 in the first experiment). Regarding coreference resolution, we did not obtain any improvements over the baseline, as in the first experiment.

These results correspond to the less clear differences in the relation analysis (cf. Table 8.6) but are unexpected because in our opinion the setup for experiment 2 in comparison to the setup for experiment 1 was clearly improved regarding the task requirements.

8.5. Final performance of the bridging tool

While the coreference resolver could not benefit from the additional semantic knowledge, the bridging tool’s performance increases, as shown in the previous experiments.

The final performance of the bridging resolver is given in Table 8.7. We also show that the added features work best if we include coreference information (gold or predicted), as illustrated in Table 8.8.

| | Anaphor recognition | | | Full bridging resolution | | |
|----------------------------|---------------------|------|------|--------------------------|------|-------------|
| | P | R | F1 | P | R | F1 |
| Without semantic relations | 79.6 | 14.1 | 23.9 | 59.8 | 10.6 | 18.0 |
| With predicted meronymy | 71.6 | 18.3 | 29.2 | 50.0 | 12.8 | 20.4 |

Table 8.7.: Final performance of the English bridging system

| Setting | Precision | Recall | F1 |
|-----------------------|-----------|--------|-------------|
| No coreference | 32.0 | 12.8 | 18.3 |
| Predicted coreference | 47.9 | 12.8 | 20.2 |
| Gold coreference | 50.0 | 12.8 | 20.4 |

Table 8.8.: Final performance of the English bridging system with different coreference information, gold mention setting

8.6. Discussion and conclusion

As the data for which we predicted the relations do not contain labelled relations that match the categories in our hypotheses, it is difficult to assess how well the classifiers work on this data. Despite the fact that we applied state-of-the-art methods, annotating at least a small part of the data would be necessary to assess the quality of the predictions. Our analysis shows that while some of our hypotheses have been confirmed, e.g. that meronymy is the most important relation for bridging, which can be used to improve the performance of a bridging resolver, the distribution of the relations in actual corpus data seems to be more complex than our initial hypotheses suggested, as we find for example also cases of meronymy in the coreference pairs.

For some of the relations, the missing direction can be problematic. In Example (11), the goal of the bridging resolver is to find an antecedent for *the city's*. *The city* itself has not been introduced before, only *the Marina neighborhood* (the gold antecedent). As we do not have a direction encoded in our data, we have a high `meronymy` score for *resident - city*, although the part-of-relation clearly exists the other way around: one can introduce a city and then talk about parts of it, e.g. the residents, but not the other way around. This information is unfortunately not given in the data. The pair *neighborhood - city* has a low `meronymy` score and a high score for `coord` (co-hyponymy).

- (11) In the hard-hit Marina neighborhood_{gold}, life after the earthquake is often all too real, but sometimes surreal. Some scenes: Saturday morning, a resident_{predicted} was given 15 minutes to scurry into a sagging building and reclaim what she could of her life's possessions. Saturday night she dined in an emergency shelter on salmon steaks prepared by chefs from one of **the city's** four-star restaurants.

As the performance for coreference resolution is already quite high, the predicted relations did not improve the performance. For bridging resolution, however, the per-

formance is typically low, and further work on finding general cases of bridging seems promising.

9. Conclusion

9.1. Summary of contributions

The aim of this thesis is to improve coreference and bridging resolution, both on the theoretical and the computational level. In this section, we summarise the contributions presented in this thesis.

A refined bridging definition The first contribution is one for which the need became evident while performing bridging experiments on available corpora, where our bridging resolver did not generalise well to other corpora due to very different types of bridging annotated in these resources. We introduced the term **referential bridging** to cover two types of bridging on the level of referring expressions: (i) argument slot filling (*the wheel (of the car)*) and (ii) referential subset expressions (*the small pug (out of the previously mentioned group of dogs)*). In both cases, context-dependence is the main criterion for referential bridging. This is not the case for **lexical or lexically induced bridging**, where we have an anaphoric or non-anaphoric expression that stands in some relation with a previously introduced entity. This relation typically exists either on the word level or models a real-world relation based on the relation on the concept level (*Europe - Spain*). One special case that has sometimes been annotated as bridging are **non-referential subset cases**, where the non-anaphoric expression is a subset or a superset of a previously introduced entity (*computers - small computers*). These are cases of lexical givenness, as the head word is considered lexically given.

Three new bridging corpora To overcome the lack of available data with compatible bridging definitions, we have annotated three medium-sized corpora, one newspaper (in-domain) corpus of about 57k tokens called BASHI and one scientific (out-of-domain) corpus of 61k tokens called SciCorp. For German, we have annotated a radio interview corpus containing 20 interviews of about 10 minutes each, called GRAIN. SciCorp

9. Conclusion

and GRAIN were also annotated with coreference information, while BASHI already contained coreference since we used articles from the OntoNotes corpus.

A state-of-the-art coreference resolver for German Our adaptation of an English data-driven coreference resolver to German mainly focused on features designed to capture specificities of German. The adapted tool achieves state-of-the-art performance on the German benchmark dataset TüBa-D/Z and enables us to do linguistic validation experiments on the use of prosodic features for coreference resolution. Table 9.1 shows the performance on TüBa-D/Z version 8 in comparison to other resolvers.¹

| System | CoNLL gold | CoNLL regular |
|------------------------|---------------|------------------|
| IMS HotCoref DE (open) | 63.61 | 48.61 |
| CorZu (open) | 58.11 | 45.82 |
| BART (open) | 45.04 | 39.07 |
| SUCRE (closed) | 51.55 | 36.32 |
| TANL-1 (closed) | 20.39 | 14.17 |

Table 9.1.: Comparison of different German coreference systems

A state-of-the-art bridging resolver for English Based on existing work on rule-based bridging resolution and motivated by the lack of an openly available bridging resolver, we have developed a system for full bridging resolution for English that achieves state-of-the-art performance. We have shown that filtering out gold or automatically predicted coreference before performing the bridging resolution step improves performance. Coreference information is helpful because bridging and coreference anaphors are difficult to distinguish, as they are both typically short, often definite expressions. Table 9.2 shows the performance of the bridging resolver as well as the effect of coreference.

| Setting | Precision | Recall | F1 |
|-----------------------|-----------|--------|-------------|
| No coreference | 49.6 | 10.6 | 17.4 |
| Predicted coreference | 59.8 | 10.6 | 18.0 |
| Gold coreference | 59.8 | 10.6 | 18.0 |

Table 9.2.: Performance of the English bridging system

¹The performance on the newest TüBa-D/Z version 10 is presented in Section 5.2.3.

The resolver can generalise well to other corpora, if they contain referential bridging as annotated in the corpus ISNotes, on which the original system was designed. We have also proposed extensions of the system that can also handle lexical bridging and lexical givenness and compared the system against a learning-based approach.

A state-of-the-art bridging resolver for German We have developed a rule-based bridging system for German that is the first publicly available bridging resolver for German, which achieves state-of-the-art performance on the DIRNDL corpus. We show that, again, filtering out gold or automatically predicted coreference anaphors improves performance, as presented in Table 9.3.

| Setting | Precision | Recall | F1 |
|-----------------------|-----------|--------|-------------|
| No coreference | 21.4 | 9.2 | 12.8 |
| Predicted coreference | 22.4 | 9.2 | 13.0 |
| Gold coreference | 31.9 | 9.2 | 14.2 |

Table 9.3.: Performance of the German bridging resolver (on DIRNDL)

Prosodic information improves coreference resolution Our linguistic validation experiments have proven that both manually annotated and automatically predicted prosodic information improves coreference resolution. We showed that the presence of a pitch accent is a useful feature in a learning-based setting, and that including prosodic boundaries and inferring nuclear accents improves the performance for complex NPs. Surprisingly, the drop in performance was small when training the system on gold pitch accents and applying it on automatically predicted pitch accents in unseen texts. This is a promising result and shows that this strategy can also be used in an application scenario. Table 9.4 shows the effect of our two main features, pitch accent presence and nuclear accent presence.

| Baseline | 46.11 | |
|---------------------------|--------------|--------------|
| + Accent presence | short NPs | all NPs |
| + gold | 53.99 | 49.68 |
| + gold/auto | 52.63 | 50.08 |
| + auto | 49.13 | 49.01 |
| + Nuclear accent presence | | |
| + gold | 48.63 | 52.12 |
| + gold/auto | 48.46 | 51.45 |
| + auto | 48.01 | 50.64 |

Table 9.4.: Performance of pitch accent and nuclear accent presence (in CoNLL score)

Automatically predicted meronymy improves bridging resolution We have shown that meronymy as predicted by a state-of-the-art neural-net relation classifier improves bridging resolution, as shown in Table 9.5. Our results indicate that the often made assumption that meronymy is the prototypical bridging relation holds true in our data, as it was the only relation with which we could improve our bridging resolver. As the bridging antecedent and anaphor are generally thought to be related, i.e. they have a high similarity in a word vector space, adding a cosine similarity threshold also improved results.

| | Anaphor recognition | | | Full bridging resolution | | |
|----------------------------|---------------------|------|------|--------------------------|------|-------------|
| | P | R | F1 | P | R | F1 |
| Without semantic relations | 79.6 | 14.1 | 23.9 | 59.8 | 10.6 | 18.0 |
| With predicted meronymy | 71.6 | 18.3 | 29.2 | 50.0 | 12.8 | 20.4 |

Table 9.5.: Final performance of the English bridging system

9.2. Lessons learned

Over the course of this thesis, our understanding of the problems involved in coreference and bridging resolution have continuously grown. The main contributions have already been presented in the last section and many of the smaller, more detailed issues as well as the suggested solutions are already contained in the individual chapters. In this section, we want to reflect on some of the more meta-level lessons we learned during the preparation of this thesis.

Is the concept of bridging too vague to be modelled? Yes and no.

When we started our experiments on bridging, the type of bridging we had in mind was referential bridging as annotated in ISNotes, where non-identical context dependence is the main criterion for bridging. When working on available corpora such as ARRAU, we noticed that there were many different phenomena annotated as bridging. We think that our introduction of the concepts referential and lexical bridging helps to make the bridging definition clearer and we hope that this very important distinction will also be acknowledged in the creation of new data or validation annotation checks for existing corpora. With the current state of bridging annotations, taking several corpora in order to have more data is not a good idea, as the corpora contain all kinds of contradictory phenomena, which is problematic.

If we concentrate on one type of bridging, namely referential bridging, the annotations in the corpora that only have this type of annotation (ISNotes or BASHI) show an inter-annotator-agreement (κ) of about 0.6. Thus, we think that the definition of referential bridging is clear enough to be modelled automatically, although of course the kappa values will always only be moderate (or borderline substantial), as the annotations depend very much on subjective interpretations of the text. The largest problem remains that the corpora annotated with this type of bridging contain only a few bridging pairs.

This is different in corpora such as ARRAU, where the focus was set on specific pre-defined relations independently of context-dependence. These are typically easier and faster to annotate and, as a result, the corpus also contains much more so-called bridging pairs. However, since referential bridging, lexical bridging and non-anaphoric subset relations are not distinguished in this corpus, we would argue that the bridging definition there is too vague to be modelled and that the annotations should be enriched with a layer which tells us which expressions are actually anaphoric and which are not. The non-anaphoric lexical bridging and subset cases are a completely different task and have much in common with the prediction of semantic relations between words, which is an NLP task in its own right, that has received a lot of attention over the last years.

Should we think more about annotation decisions and limitations before doing corpus annotation? Yes, definitely.

Before deciding to make serious annotation limitations such as to limit bridging anaphors to definite NPs or to only annotate nominal antecedents, we should reflect more on the consequences of these limitations. Antecedents can for example always be labelled as non-nominal and filtered out, if desired, but if they are not annotated in the first place,

9. Conclusion

the cost of re-annotating them later will be much higher. When combining several corpora, taking out the cases when they are labelled as such is much easier than working with corpora where certain things are not annotated. In the worst case, the result is non-compatible corpus resources. At the start of this thesis, we were heavily influenced by theoretical studies assuming that indefinite NPs can be interpreted without context because they introduce new information. As a result, and also because we thought this would make the annotation process easier, we decided to restrict bridging anaphors in SciCorp and GRAIN to definite NPs. It turned out that it was sometimes difficult for the annotators to decide which markables were definite expressions (for example in cases involving bare singulars), so this decision complicated rather than facilitated the annotation process. In the meantime, many examples have convinced us that indefinite expressions can also be bridging anaphors (*Starbucks – an employee*) and we would suggest not to make such restrictions when annotating anaphoric phenomena. More generally, introducing meaningful extra labels that one might not think are necessary as of now might help the compatibility or later use of the corpus resources.

Are linguistic validation experiments still helpful in the age of neural-net models?

In our case, yes.

While contributing an interesting perspective from the viewpoint of an applied setting on the theoretical claims, both our experiments also managed to improve the performance of the respective tools: our coreference resolver could be improved by including prosodic information, and our bridging resolver benefitted from automatically predicted meronyms. This shows that in the age of neural-net models based on word embeddings, linguistic information can still help enable state-of-the-art resolvers.

Should coreference and bridging be learned jointly? Learning coreference and bridging jointly remains an interesting and promising idea, that is unfortunately difficult to put into practice due to the lack of data for bridging.

Despite the fact that this thesis is concerned with both coreference and bridging resolution, you might have noticed that the tasks are treated in separate chapters and that there is not much interaction except the occasional filtering out of coreference anaphors before performing bridging resolution.

The idea to model coreference and bridging in the same framework and to learn them in a joint setting was one of the main ideas we had at the start of preparing this thesis. Learning these two rather closely related anaphoric tasks jointly makes sense

because they are two sides of anaphora resolution that involve similar steps: first, one has to determine that an expression is anaphoric and in a second step, the best fitting antecedent has to be selected. Coreference and bridging anaphors are also similar in appearance and are as a result often confused by an automatic system, as they are often definite expressions and in any case typically short. Filtering out coreference anaphors before performing bridging resolution has proven to improve results in our experiments. We are confident that removing bridging anaphors from the set of potential coreference anaphor candidates would also increase results for coreference resolution. The antecedent search principles applied in both tasks are also similar and there is a huge overlap in terms of factors that come into play and determine the salience of an antecedent, such as recency or grammatical roles. Grishina (2016) presented interesting correlations between coreference clusters and bridging anaphors, for example the fact that 56% of all the clusters have associated bridging markables and that there is a difference in terms of the average size of a cluster that is connected to a bridging markable (6.1 markables) and a non-bridging cluster (2.4 markables). In their data, the largest bridging cluster contained 22 markables, while the largest non-bridging cluster only contained 9 markables. This means that a cluster connected to a bridging markable is usually larger than an average cluster. These differences could be exploited in a joint learning setting. In terms of evaluating bridging, there is also the dependency that predicted bridging antecedents do not have to be the exact gold antecedent, as long as they are both in the same coreference chain.

We assume that the latent tree approach as used in our coreference resolution system is particularly suited for this approach. Figure 9.1 shows the structure for Example (1). As explained in Section 3.1, the approach assumes a hidden structure underlying the data in the form of latent trees. Besides the coreference pairs, one could also integrate the bridging pairs in the latent tree approach with a different type of relation (the two bridging anaphors are marked in green and the bridging relation is illustrated with green arrows). This way, we can not only learn from the relation between the bridging anaphor and its antecedent (as in the pair-based learning approach in our bridging chapter), but can also make use of information coming from the coreference cluster of the antecedent. For example, *the great hall* could be in a coreference cluster with other head words such as *room* or *atrium*. This could help establish that *the little door* is a part of *the great hall*, as *room* and *door* are rather prototypical examples of a **whole-part** relation. Note that we do not show all mentions and coreference chains, but focus on a few to make our point.

9. Conclusion

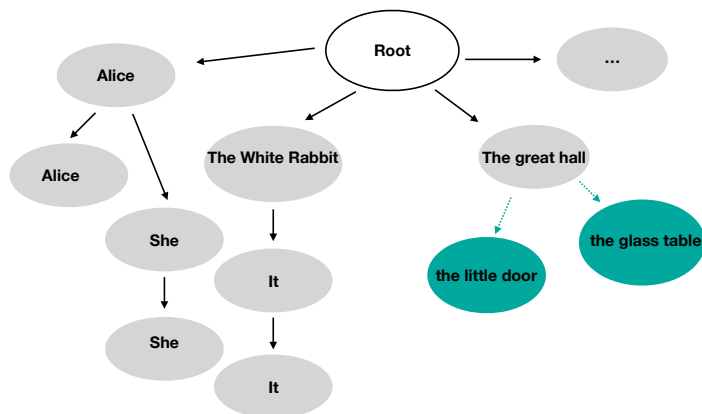


Figure 9.1.: A data structure based on latent trees for the joint learning of coreference and bridging

- (1) It was the White Rabbit, trotting slowly back again, and looking anxiously about as it went, as if it had lost something; and Alice heard it muttering to itself [...] Alice guessed in a moment that it was looking for the fan and the pair of white kid gloves, and she very good-naturedly began hunting about for them, but they were nowhere to be seen – everything seemed to have changed since her swim in the pool, and the great hall, with the glass table and the little door, had vanished completely.

We performed experiments using this data structure and the ISNotes corpus in some prototypical tests. However, with the 633 bridging pairs in ISNotes, the outcome was a bit disappointing: the data is just too small to learn the complex dependencies that exist between the two tasks and as a result, we could not make the approach work, which is why this experiment is also not included in the main chapters of this thesis. We assume that for a positive effect on coreference resolution, it would need a very large bridging corpus. As the state-of-the-art for bridging is not as advanced as for bridging resolution, improvements for bridging resolution could probably be made with a smaller bridging corpus, but still, we would need a corpus of a much larger size than the ones which are currently available. Even when combining the ISNotes corpus with the newly created BASHI corpus, the datasize is still too small.

In order to improve bridging resolution, there are also other interactions which could be exploited. In previous research on coreference resolution, e.g. in Lassalle and Denis (2015), anaphoricity and coreference was learned jointly. A similar thing could be ex-

exploited for bridging resolution, for example to jointly learn anaphoricity and certain semantic relations, as these are the two requirements for referential bridging.

When larger bridging corpora become available, learning coreference and bridging jointly seems like a promising direction of future research. Of course, one could then also use a different data structure than the one presented here, for example in a deep learning approach. With the current lack of bridging data, the answer to this question is unfortunately that learning coreference and bridging jointly with only little bridging data does not seem to work.

9.3. Future work

In this section, we discuss ideas for future work based on our contributions in this thesis.

Create a large-scale corpus for referential bridging Having created BASHI as a medium-sized corpus annotated with referential bridging, BASHI and ISNotes now contain about 1000 bridging pairs. This is enough to perform experiments using learning-based methods, but for being able to generalise well and for neural-net approaches, where we typically need about the same amount of data points as parameters in the neural net, it might still be too little data. Therefore, a large-scale corpus of referential bridging would benefit further development in bridging resolution a lot.

Apply neural-net approaches to coreference and bridging resolution During the preparation of this thesis, neural-net approaches came up in the field of coreference resolution and have replaced other learning-based approaches as the state-of-the-art (the first approach was presented in Clark and Manning (2016b)). For German, to the best of our knowledge, no one has applied this method on TüBa-D/Z. As this is a rather large benchmark dataset, the approach should also work for German. The advantage of such an approach is the rather slim feature set, which mainly consists of word embeddings and a number of rather basic features. However, systems relying heavily on lexical features such as word embeddings should also be used with caution, as Moosavi and Strube (2017) warned that they generalise badly to unseen data, as there is often an overlap in terms of lexical material in the training and test set, which the lexical features then just memorise.

Find better ways to generalise in bridging resolution One of the main issues in state-of-the-art bridging resolution is that the rules or the learning applied does not generalise well to unseen bridging pairs. In the bridging chapter, we have seen that a system designed for news text does not work for other domains due to the many specific rules contained in the resolver. Our learning-based experiments also showed that with the current features that are included, the statistical system does not seem to generalise better than the rule-based approach.

Hou (2018) has presented word embeddings based on prepositional modification patterns (*wheels of the car*), to capture semantic relatedness in the word embeddings. Due to the implicit semantic relations contained in the word embeddings, this works better than using plain prepositional modification patterns such as the ones used in our bridging resolver. The approach has so far only been used to select antecedents for given bridging anaphors, not for full bridging resolution.

If more bridging corpora annotated with referential bridging are released in the future, using word embeddings based on specific syntactic patterns is a promising direction to be applied in a neural-net setting to resolve bridging references.

Linguistic validation experiments Our two experiments were meant to motivate further use on the helpfulness of linguistic or semantic information. As our experiments were mainly pilot studies to test the principled usefulness, follow-up experiments could create more insight into how the information can be applied. For prosody, it would be interesting to include the features used for pitch accent detection directly as a feature in the machine learning, particularly in a neural-net setting, where complicated dependencies can be learned.

Of course, there are many more different types of linguistic information that could be integrated. One idea would be to investigate the relation between focus/topic and coreference and bridging. This would be interesting from a theoretical point of view, but could also benefit the tools' performances, although the current status in automatic focus/topic prediction is probably not advanced enough to be applied as predicted information. However, experiments on gold annotated focus/topic could give us an idea of how this could benefit anaphora resolution. The corpus GRAIN, for example, contains both coreference and bridging as well as focus and topic information.

Bibliography

- Alshawi, H. (1987). Memory and context for language interpretation.
- Amoia, M., Kunz, K., and Lapshinova-Koltunski, E. (2012). Coreference in spoken vs. written texts: a corpus-based analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).
- Asher, N. (1993). Reference to abstract objects in discourse.
- Asher, N. and Lascarides, A. (1998). Bridging. *Journal of Semantics*, 15(1):83–113.
- Attardi, G., Simi, M., and Dei Rossi, S. (2010). 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. Association for Computational Linguistics.
- Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.
- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1, COLING '98*, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Bärenfänger, M., Goecke, D., Hilbert, M., Lungen, H., and Stührenberg, M. (2008). Anaphora as an indicator of elaboration: A corpus study. *JLCL*, 23(2):49–73.
- Baroni, M. and Lenci, A. (2011). How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 1–10, Stroudsburg, PA, USA.
- Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- Batista-Navarro, R. and Ananiadou, S. (2011). Building a coreference-annotated corpus from the domain of biochemistry. *Proceedings of The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, page 83.

Bibliography

- Baumann, S. and Riester, A. (2012). Referential and lexical givenness: Semantic, prosodic and cognitive aspects. *Prosody and meaning*, 25:119–162.
- Baumann, S. and Riester, A. (2013). Coreference, lexical givenness and prosody in German. *Lingua*, 136:16–37.
- Baumann, S., Röhr, C., and Grice, M. (2015). Prosodische (De-)Kodierung des Informationsstatus im Deutschen. *Zeitschrift für Sprachwissenschaft*, 34(1):1–42.
- Baumann, S. and Roth, A. (2014). Prominence and coreference – On the perceptual relevance of F0 movement, duration and intensity. In *Proceedings of Speech Prosody*, pages 227–231.
- Beckman, M., Hirschberg, J., and Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In Jun, S.-A., editor, *Prosodic Typology – The Phonology of Intonation and Phrasing*, pages 9–54. Oxford University Press.
- Bergsma, S. and Lin, D. (2006). Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Sydney, Australia. Association for Computational Linguistics.
- Björkelund, A., Eckart, K., Riester, A., Schaufler, N., and Schweitzer, K. (2014). The extended DIRNDL corpus as a resource for automatic coreference and bridging resolution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC 2014, pages 3222–3228.
- Björkelund, A. and Farkas, R. (2012). Data-driven multilingual coreference resolution using resolver stacking. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Shared Task*, pages 49–55. Association for Computational Linguistics.
- Björkelund, A. and Kuhn, J. (2014). Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47–57, Baltimore, Maryland. Association for Computational Linguistics.
- Bohnet, B. and Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea. Association for Computational Linguistics.

- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Brennan, S. E., Friedman, M. W., and Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, pages 155–162. Association for Computational Linguistics.
- Broscheit, S., Poesio, M., Ponzetto, S. P., Rodriguez, K. J., Romano, L., Uryupina, O., Versley, Y., and Zanolini, R. (2010a). Bart: A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 104–107, Uppsala, Sweden. Association for Computational Linguistics.
- Broscheit, S., Ponzetto, S. P., Versley, Y., and Poesio, M. (2010b). Extending BART to Provide a Coreference Resolution System for German. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.
- Bunescu, R. (2003). Associative anaphora resolution: A web-based approach. In *Proceedings of the 2003 EACL Workshop on The Computational Treatment of Anaphora*.
- Bussmann, H. (1990). *Lexikon der Sprachwissenschaft*. Kröners Taschenausgabe. Kröner.
- Cahill, A. and Riestler, A. (2012). Automatically acquiring fine-grained information status distinctions in German. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 232–236. Association for Computational Linguistics.
- Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. In *Language Resources and Evaluation*, volume 44, pages 387–419.
- Cap, F. (2014). Morphological processing of compounds for statistical machine translation. Dissertation, Institute for Natural Language Processing (IMS), University of Stuttgart.
- Carden, G. (1982). Backwards anaphora in discourse context. *Journal of Linguistics*, 18(2):361–387.
- Caselli, T. and Prodanof, I. (2006). Annotating bridging anaphors in Italian: in search of reliability. *Relation*, 27:9–03.
- Castaño, J., Zhang, J., and Pustejovsky, J. (2002). Anaphora resolution in biomedical literature. In *Proceedings of the International Symposium on Reference Resolution for NLP*.

Bibliography

- Chomsky, N. (1981). *Lectures on Government and Binding*. Foris, Dordrecht.
- Chomsky, N. (1988). *Current issues in linguistic theory*, volume 38. Walter de Gruyter.
- Clark, H. H. (1975). Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 169–174. Association for Computational Linguistics.
- Clark, K. and Manning, C. D. (2016a). Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of The 2016 Conference on Empirical Methods on Natural Language Processing*.
- Clark, K. and Manning, C. D. (2016b). Improving coreference resolution by learning entity-level distributed representations. *Proceedings of The 54th Annual Meeting of the Association for Computational Linguistics*.
- Clear, J. H. (1993). The digital word. chapter The British National Corpus, pages 163–187. MIT Press, Cambridge, MA, USA.
- Cohen, K. B., Lanfranchi, A., Corvey, W., Jr., W. A. B., Roeder, C., Ogrena, P. V., Palmer, M., and Hunter, L. E. (2010). Annotation of all coreference in biomedical text: Guideline selection and adaptation. In *Proceedings of the Second Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2010), LREC 2010*.
- Cruttenden, A. (2006). The de-accenting of given information: a cognitive universal? In Bernini, G. and Schwartz, M., editors, *Pragmatic Organization of Discourse in the Languages of Europe*, pages 311–355. De Gruyter, Berlin.
- Daumé, H. and Marcu, D. (2009). Learning as search optimization: Approximate large margin methods for structured prediction. *Proceedings of the 22nd international conference on Machine learning*, abs/0907.0809.
- Denis, P. and Baldridge, J. (2007). A ranking approach to pronoun resolution. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, volume 158821593.
- Dinu, G., Pham, N. T., and Baroni, M. (2013). DISSECT – DIStributional SEmantics Composition Toolkit. In *Proceedings of The 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria.
- Dipper, S., Lüdeling, A., and Reznicek, M. (2013). NoSta-D: A corpus of German non-standard varieties. *Non-Standard Data Sources in Corpus-Based Research*, (5):69–76.
- Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In *Proceedings of the fourth international conference on*

- Language Resources and Evaluation*. European Language Resources Association.
- Donnellan, K. S. (1966). Reference and definite descriptions. *The philosophical review*, 75(3):281–304.
- Draudt, A.-C. (2018). Inter-Annotator Agreement von Informationsstatus-Annotationen im GRAIN-Korpus (BSc thesis).
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.
- Durrett, G. and Klein, D. (2013). Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982.
- Eckart, K., Riestler, A., and Schweitzer, K. (2012). A discourse information radio news database for linguistic analysis. In Christian Chiarcos, S. N. and Hellmann, S., editors, *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, pages 65–76. Springer.
- Eyben, F., Weninger, F., Groß, F., and Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838.
- Faaß, G. and Eckart, K. (2013). SdeWaC – a corpus of parsable sentences from the web. In *Language Processing and Knowledge in the Web*, Lecture Notes in Computer Science. Springer.
- Fang, Y. and Teufel, S. (2014). A summariser based on human memory limitations and lexical competition. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Faruqui, M. and Padó, S. (2010). Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of Die Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS) 2010*, Saarbrücken, Germany.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.
- Fernandes, E., dos Santos, C., and Milidiú, R. (2012). Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Shared Task*, pages 41–48, Jeju Island, Korea. Association for Computational Linguistics.
- Féry, C. (1993). *German Intonational Patterns*. Niemeyer, Tübingen.

Bibliography

- Feuerbach, T., Riedl, M., and Biemann, C. (2015). Distributional semantics for resolving bridging mentions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 192–199.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Fraurud, K. (1990). Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7(4):395–433.
- Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.
- Gardent, C. and Manuélian, H. (2005). Création dun corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues*, 46(1):115–140.
- Gardent, C., Manuelian, H., and Kow, E. (2003). Which bridges for bridging definite descriptions? In *Proceedings of EACL: Fourth International Workshop on Linguistically Interpreted Corpora*, pages 69–76, Budapest.
- Garvey, C. and Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, 5(3):459–464.
- Gasperin, C. and Briscoe, T. (2008). Statistical anaphora resolution in biomedical texts. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 257–264. Association for Computational Linguistics.
- Gasperin, C., Karamanis, N., and Seal, R. (2007). Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium*.
- Ge, N., Hale, J., and Charniak, E. (1998). A statistical approach to anaphora resolution. In *Sixth Workshop on Very Large Corpora*.
- Grishina, Y. (2016). Experiments on bridging across languages and genres. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: the CORBON 2016 Workshop on Coreference Resolution Beyond OntoNotes*, pages 7–15.
- Grosz, B., Weinstein, S., and Joshi, A. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*.
- Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
- Gussenhoven, C. (1984). *On the Grammar and Semantics of Sentence Accents*. Foris, Dordrecht.

- Haghighi, A. and Klein, D. (2007). Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th annual meeting of the Association of Computational Linguistics*, pages 848–855.
- Hahn, U., Strube, M., and Markert, K. (1996). Bridging textual ellipses. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1, COLING '96*, pages 496–501, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Poláková, L., Rysová, M., Sgall, P., Spoustová, J., Straňák, P., Synková, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š., and Žabokrtský, Z. (2018). Prague dependency treebank 3.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman, London.
- Hamp, B. and Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.
- Harabagiu, S. M., Moldovan, D. I., Pasca, M., Surdeanu, M., Mihalcea, R., Girju, R., Rus, V., Lacatusu, V. F., Morarescu, P., and Bunescu, R. C. (2001). Answering complex, list and context questions with lcc's question-answering server. In *Proceedings of Text Retrieval Conference (TREC)*.
- Hardmeier, C. and Federico, M. (2010). Modelling pronominal anaphora in statistical machine translation. In *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*, pages 283–289.
- Hawkins, J. A. (1978). Definiteness and indefiniteness: A study in reference and grammaticality prediction.
- Hearst, M. A. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16.
- Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.-M., Van Der Vloet, J., and Verschelde, J.-L. (2008). A coreference corpus and resolution system for Dutch.
- Hirschman, L. and Chinchor, N. (1998). Appendix F: MUC-7 Coreference Task Definition (version 3.0). In *Proceedings of the Seventh Message Understanding Conference*

Bibliography

(MUC-7).

- Hobbs, J. R. (1979). Coherence and coreference. *Cognitive science*, 3(1):67–90.
- Hobbs, J. R., Stickel, M. E., Appelt, D. E., and Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, 63:69–142.
- Hou, Y. (2016a). Incremental Fine-grained Information Status Classification Using Attention-based LSTMs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1880–1890.
- Hou, Y. (2016b). *Unrestricted Bridging Resolution*. PhD thesis.
- Hou, Y. (2018). Enhanced word representations for bridging anaphora resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 1–7.
- Hou, Y., Markert, K., and Strube, M. (2013a). Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 814–820.
- Hou, Y., Markert, K., and Strube, M. (2013b). Global inference for bridging anaphora resolution. In *Proceedings of The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–917.
- Hou, Y., Markert, K., and Strube, M. (2014). A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2082–2093.
- Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.
- Kamp, H. (1981). A theory of truth and semantic representation. *Formal semantics-the essential readings*, pages 189–222.
- Kaplan, D., Iida, R., Nishina, K., and Tokunaga, T. (2012). Slate – a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, pages 89–101.
- Karttunen, L. (1969). Discourse referents. In *Proceedings of the 1969 Conference on Computational Linguistics, COLING '69*, pages 1–38, Stroudsburg, PA, USA. Associ-

- ation for Computational Linguistics.
- Klenner, M. and Tuggener, D. (2011). An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 178–185, Hissar, Bulgaria.
- Kobayashi, N., Inui, K., and Matsumoto, Y. (2007). Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.
- Kobdani, H. and Schütze, H. (2010). Sucre: A modular system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 92–95, Uppsala, Sweden. Association for Computational Linguistics.
- Korzen, I. and Buch-Kromann, M. (2011). Anaphoric relations in the copenhagen dependency treebanks. *Corpus-based Investigations of Pragmatic and Discourse Phenomena*, 3:83–98.
- Kripke, S. (1972). Naming and necessity. In Davidson, D. and Harman, G., editors, *Semantics of Natural Language*, pages 253–355. Springer, Dordrecht.
- Krug, M., Puppe, F., Jannidis, F., Macharowsky, L., Reger, I., and Weimar, L. (2015). Rule-based coreference resolution in German historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 98–104.
- Ladd, D. R. (2008). *Intonational Phonology (2nd ed.)*. Cambridge University Press.
- Lakoff, G. (1971). The role of deduction in grammar.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.
- Lappin, S. and Leass, H. (1994). An algorithm for pronominal anaphora resolution.
- Lassalle, E. and Denis, P. (2011). Leveraging different meronym discovery methods for bridging resolution in French. In *Discourse Anaphora and Anaphor Resolution Colloquium*, pages 35–46. Springer.
- Lassalle, E. and Denis, P. (2015). Joint anaphoricity detection and coreference resolution with constrained latent structures.
- Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). 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: Shared task*, pages 28–34. Association for Computational Linguistics.

Bibliography

- Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197. Association for Computational Linguistics.
- Levy, O., Remus, S., Biemann, C., and Dagan, I. (2015). Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, USA.
- Löbner, S. (1998). Definite associative anaphora. *Manuscript: <http://user.phil-fak.uniduesseldorf.de/~loebner/publ/DAA-03.pdf>*.
- Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32. Association for Computational Linguistics.
- Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., and Roukos, S. (2004). A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*.
- Markert, K., Hou, Y., and Strube, M. (2012). Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 795–804. Association for Computational Linguistics.
- Markert, K., Nissim, M., and Modjeska, N. (2003). Using the web for nominal anaphora resolution. In *Proceedings of the 2003 EACL Workshop on the Computational Treatment of Anaphora*.
- Markert, K., Strube, M., and Hahn, U. (1996). Inferential realization constraints on functional anaphora in the centering model. In *In Proceedings of the 18 th Annual Conference of the Cognitive Science Society; La*, pages 609–614.
- Martí, M., Taulé, M., Bertran, M., and Márquez, L. (2007). AnCora: Multilingual and Multilevel Annotated Corpora.
- Mayer, J. (1995). Transcription of German Intonation. The Stuttgart System. University of Stuttgart.
- McKeown, K., Daume III, H., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P., Biran, O., Bothe, S., Collins, M., Fleischmann, K. R., et al. (2016). Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*, 67(11):2684–2696.
- Mikhaylova, A. (2014). Koreferenzresolution in mehreren Sprachen. Msc thesis, Center for Information and Language Processing, University of Munich.

- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, abs/1301.3781.
- Mirkin, S., Dagan, I., and Padó, S. (2010). Assessing the role of discourse references in entailment inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL 2010, pages 1209–1219. Association for Computational Linguistics.
- Moosavi, N. S. and Strube, M. (2016). Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.
- Moosavi, N. S. and Strube, M. (2017). Lexical features in coreference resolution: To be used with caution. *arXiv preprint arXiv:1704.06779*, abs/1704.06779.
- Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, AKBC-WEKEX '12*, pages 95–100, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Naumann, K. and Möller, V. (2006). Manual for the annotation of in-document referential relations. *University of Tübingen*.
- Nedoluzhko, A., Mírovský, J., Ocelák, R., and Pergler, J. (2009). Extended coreferential relations and bridging anaphora in the prague dependency treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*, Goa, India, pages 1–16.
- Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411. Association for Computational Linguistics.
- Ng, V. and Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Nicolov, N., Salvetti, F., and Ivanova, S. (2008). Sentiment analysis: Does coreference matter. In *AISB 2008 convention communication, interaction and social intelligence*, volume 1, page 37.
- Nissim, M., Dingare, S., Carletta, J., and Steedman, M. (2004). An annotation scheme for information status in dialogue. *Proceedings of the 4th international conference on language resources and evaluation (LREC)*.

Bibliography

- Ostendorf, M., Price, P., and Shattuck-Hufnagel, S. (1995). The Boston University Radio News Corpus. Technical Report ECS-95-001, Boston University.
- Pagel, J. (2018). Rule-based and learning-based approaches for automatic bridging detection and resolution in German (MSc thesis).
- Pagel, J. and Rösiger, I. (2018). Towards bridging resolution in German: Data analysis and rule-based experiments. In *Proceedings of NAACL-HLT: Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC)*, pages 50–60, New Orleans, US.
- Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English Gigaword fifth edition. *Philadelphia: Linguistic Data Consortium*.
- Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics.
- Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology.
- Poesio, M. (2004). Discourse annotation and semantic annotation in the gnome corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 72–79. Association for Computational Linguistics.
- Poesio, M. and Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky*, pages 76–83. Association for Computational Linguistics.
- Poesio, M. and Artstein, R. (2008). Anaphoric Annotation in the ARRAU Corpus. In *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.
- Poesio, M., Grishina, Y., Kolhatkar, V., Moosavi, N. S., Rösiger, I., Roussell, A., Uma, A., Uryupina, O., Yu, J., and Zinsmeister, H. (2018). Anaphora resolution with the arrau corpus. In *Proceedings of NAACL-HLT: Workshop on Computational Models of Reference, Anaphora and Coreference*, New Orleans, USA.
- Poesio, M., Ishikawa, T., Im Walde, S. S., and Vieira, R. (2002). Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 3rd international conference on language resources and evaluation (LREC)*.
- Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004). Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 143. Association for Computational Linguistics.

- Poesio, M., Stuckardt, R., and Versley, Y., editors (2016). *Anaphora Resolution - Algorithms, Resources, and Applications*. Theory and Applications of Natural Language Processing. Springer.
- Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.
- Poesio, M., Vieira, R., and Teufel, S. (1997). Resolving bridging references in unrestricted text. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 1–6. Association for Computational Linguistics.
- Ponzetto, S. P. and Strube, M. (2007). Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research (JAIR)*, 30:181–212.
- Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., and Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.
- Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning: Shared Task*, pages 1–40.
- Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 1–27, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Prince, E. F. (1981). Toward a taxonomy of given-new information. In *Radical Pragmatics*, pages 223–55. Academic Press.
- Prince, E. F. (1992). The zpg letter: Subjects, definiteness, and information-status. *Discourse description: diverse analyses of a fund raising text*, pages 295–325.
- Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of The 2010 Conference on Empirical Methods on Natural Language Processing*.
- Rahman, A. and Ng, V. (2009). Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 968–977. Association for Computational Linguistics.

- Rahman, A. and Ng, V. (2012). Learning the fine-grained information status of discourse entities. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 798–807, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Recasens, M. and Hovy, E. (2010a). A typology of near-identity relations for coreference (NIDENT). In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.
- Recasens, M. and Hovy, E. (2010b). BLANC: Implementing the Rand index for coreference evaluation. *Journal of Natural Language Engineering*, 16(5).
- Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 1–8, Stroudsburg, PA, USA.
- Recasens, M., Marti, M. A., and Orasan, C. (2012). Annotating near-identity from coreference disagreements. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 165–172.
- Recasens, M., Marti, M. A., and Taulé, M. (2007). Where anaphora and coreference meet. Annotation in the Spanish CESS-ECE corpus. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 504–509.
- Riester, A. and Baumann, S. (2017). The RefLex Scheme – Annotation guidelines. SinSpeC. Working papers of the SFB 732 Vol. 14, University of Stuttgart.
- Riester, A., Lorenz, D., and Seemann, N. (2010). A recursive annotation scheme for referential information status. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 717–722.
- Riester, A. and Piontek, J. (2015). Anarchy in the NP. When new nouns get deaccented and given nouns don't. *Lingua*, 165(B):230–253.
- Rodríguez, K. J., Delogu, F., Versley, Y., Stemle, E. W., and Poesio, M. (2010). Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*.
- Roitberg, A. and Nedoluzhko, A. (2016). Bridging Corpus for Russian in comparison with Czech. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016), The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 59–66.

- Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.
- Rosenberg, A., Cooper, E., Levitan, R., and Hirschberg, J. (2012). Cross-language prominence detection. In *Speech Prosody*.
- Rösiger, I. (2016). SciCorp: A Corpus of English Scientific Articles Annotated for Information Status Analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.
- Rösiger, I. (2018a). BASHI: A corpus of Wall Street Journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Rösiger, I. (2018b). Rule- and learning-based methods for bridging resolution in the arrau corpus. In *Proceedings of NAACL-HLT: Workshop on Computational Models of Reference, Anaphora and Coreference*, New Orleans, USA.
- Rösiger, I., Köper, M., Nguyen, K. A., and im Walde, S. S. (2018a). Integrating predictions from neural-network relation classifiers into coreference and bridging resolution. In *Proceedings of NAACL-HLT: Workshop on Computational Models of Reference, Anaphora and Coreference*, New Orleans, USA.
- Rösiger, I. and Kuhn, J. (2016). IMS HotCoref DE: a data-driven co-reference resolver for German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.
- Rösiger, I. and Riester, A. (2015). Using prosodic annotations to improve coreference resolution of spoken text. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 83–88, Beijing.
- Rösiger, I., Riester, A., and Kuhn, J. (2018b). Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3516–3528, Santa Fe, NM, US.
- Rösiger, I., Stehwien, S., Riester, A., and Vu, N. T. (2017). Improving coreference resolution with automatically predicted prosodic information. In *Proceedings of the First Workshop on Speech-Centric Natural Language Processing*, pages 78–83, Copenhagen. Association for Computational Linguistics.
- Rösiger, I. and Teufel, S. (2014). Resolving coreferent and associative noun phrases in scientific text. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
- Russell, B. (1905). On denoting. *Mind*, pages 479–493.

Bibliography

- Saeboe, K. J. (1996). Anaphoric presuppositions and zero anaphora. *Linguistics and Philosophy*, 19(2):187–209.
- Sasano, R. and Kurohashi, S. (2009). A probabilistic model for associative anaphora resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1455–1464, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Schäfer, R. (2015). Processing and Querying Large Web Corpora with the COW14 Architecture. In Baski, P., Biber, H., Breiteneder, E., Kupietz, M., Lungen, H., and Witt, A., editors, *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28 – 34.
- Schäfer, U., Spurk, C., and Steffen, J. (2012). A fully coreference-annotated corpus of scholarly papers from the ACL Anthology. In *Proceedings of the 24th International Conference on Computational Linguistics. International Conference on Computational Linguistics (COLING-2012), December 10-14, Mumbai, India*, pages 1059–1070.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of NeMLaP*, Manchester, UK.
- Schulte im Walde, S., Poesio, M., and Brew, C. (1998). Resolution of Inferential Descriptions in Lexical Clusters. In *Proceedings of the ECML Workshop 'Towards Adaptive NLP-driven Systems: Linguistic Information, Learning Methods and Applications'*, pages 41–52, Chemnitz, Germany.
- Schwarzschild, R. (1999). GIVENness, AvoidF, and Other Constraints on the Placement of Accent. *Natural Language Semantics*, 7(2):141–177.
- Schweitzer, K., Eckart, K., Gärtner, M., Falenska, A., Riester, A., Rösiger, I., Schweitzer, A., Stehwien, S., and Kuhn, J. (2018). German radio interviews: The GRAIN release of the SFB732 Silver Standard Collection. In *Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC 2018*.
- Shwartz, V. and Dagan, I. (2016). Path-based vs. distributional information in recognizing lexical semantic relations. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*.
- Siegel, S. and Castellan, N. J. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Berkeley, CA, 2nd edition.
- Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.
- Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *The Thirty-First AAAI Conference on Artificial Intelligence*

- (AAAI-17).
- Stehwien, S. and Vu, N. T. (2017). Prosodic event detection using convolutional neural networks with context information. In *Proceedings of Interspeech*.
- Steinberger, J., Poesio, M., Kabadjov, M. A., and Ježek, K. (2007). Two uses of anaphora resolution in summarization. *Information Processing & Management*, 43(6):1663–1680.
- Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.
- Strawson, P. F. (1950). On referring. *Mind*, 59(235):320–344.
- Strube, M. and Müller, C. (2003). A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 168–175.
- Terken, J. and Hirschberg, J. (1994). Deaccentuation of words representing ‘given’ information: Effects of persistence of grammatical function and surface position. *Language and Speech*, 37(2):125–145.
- Tetreault, J. and Allen, J. (2004). Dialogue structure and pronoun resolution. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium*.
- Tetreault, J. R. (1999). Analysis of syntax-based pronoun resolution methods. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL ’99, pages 602–605, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Tuggener, D. and Klenner, M. (2014). A hybrid entity-mention pronoun resolution model for german using markov logic networks. In *Proceedings of Die Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*, pages 21–29.
- Umbach, C. (2002). (De)accenting definite descriptions. *Theoretical Linguistics*, 2/3:251–280.
- Uryupina, O., Artstein, R., Bristot, A., Cavicchio, F., Delogu, F., Rodriguez, K., and Poesio, M. (2018). Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Journal of Natural Language Engineering*.
- Vieira, R. and Poesio, M. (2000). An empirically based system for processing definite descriptions.
- Vieira, R. and Teufel, S. (1997). Towards resolution of bridging descriptions. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 522–524. Association for Computational Linguistics.

Bibliography

- Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding, MUC6 '95*, pages 45–52, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Voorhees, E. M. et al. (1999). The TREC-8 Question Answering Track Report. In *Text Retrieval Conference (TREC)*, volume 99, pages 77–82.
- Wallin, A. and Nugues, P. (2017). Coreference resolution for Swedish and German using distant supervision. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 46–55.
- Watson, R., Preiss, J., and Briscoe, T. (2003). The contribution of domain-independent robust pronominal anaphora resolution to open-domain question-answering. In *Proceedings of the Symposium on Reference Resolution and its Applications to Question Answering and Summarization. Venice, Italy June*, pages 23–25.
- Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Pradhan, S., Ramshaw, L., and Xue, N. (2011). Ontonotes: A large training corpus for enhanced processing. pages 54–63.
- Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., et al. (2013). Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.
- Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, 3(1):1–191.
- Zeldes, A. (2017). The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation (LREC)*, 51(3):581–612.
- Zhang, R., Nogueira dos Santos, C., Yasunaga, M., Xiang, B., and Radev, D. (2018). Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 102–107. Association for Computational Linguistics.
- Zhekova, D. and Kübler, S. (2010). Ubiu: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99, Uppsala, Sweden. Association for Computational Linguistics.
- Ziering, P. (2011). Feature Engineering for Coreference Resolution in German: Improving the link feature set of SUCRE for German by using a more linguistic background. Diploma thesis, Institute for Natural Language Processing, University of Stuttgart.
- Zikánová, Š., Hajičová, E., Hladká, B., Jínová, P., Mírovský, J., Nedoluzhko, A., Poláková, L., Rysová, K., Rysová, M., and Václ, J. (2015). *Discourse and Coherence*.

From the Sentence Structure to Relations in Text, volume 14 of *Studies in Computational and Theoretical Linguistics*. Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Praha, Czechia.