Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Rule-based and Learning-based Approaches for Automatic Bridging Detection and Resolution in German

Janis Pagel

Master thesis

Prüfer:	Prof. Dr. Jonas Kuhn
	Prof. Dr. Uwe Reyle
Betreuer:	Dr. Arndt Riester
	Dipl.-Ling. Ina Rösiger

Beginn der Arbeit:	14.12.2017
Abgabe der Arbeit:	03.04.2018

Eigenständigkeitserklärung (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. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein.[1]

Stuttgart, 03.04.2018

Ort, Datum	Unterschrift
(Place, Date)	(Signature)

[1] Non-binding translation: This text is the result of my own work, and any material from published or unpublished work of others which is used either verbatim or indirectly in the text is credited to the author including details about the exact source in the text. This work has not been part of any other previous examination, neither completely nor in parts. It has neither completeley nor partially been published before. The submitted electronic version is identical to this print version.

Acknowledgements

I wish to thank my supervisors Arndt Riester and Ina Rösiger for guiding me through this thesis, listening to my ideas and giving useful feedback in many meetings. Additionally I wish to thank Ina for providing me the groundwork of the rule-based system and helping me with questions regarding the implementation.

Furthermore I would like to thank my examiner Jonas Kuhn for giving useful feedback in meetings and my thesis colloquium talk.

Another important thank goes to Prajit Dhar. Without his constant support, this thesis would not be what it is now. Also, his suggestions for the machine-learning part of the thesis and his expertise in R and regression have helped me a lot. Finally, he also provided me with a vector space built on SdeWaC from his master's thesis.

Last but not least, I am very grateful to my family. Ihr seid eine mentale und emotionale Stütze, die ich niemals missen möchte.

Abstract

The phenomenon of *bridging* describes types of non-coreferential entities, which stand in a prototypical or inferable relationship to a previously introduced discourse entity. The machine-aided resolution of such bridging relations tries to detect bridging anaphors and automatically link these anaphors to their antecedents. Research on automatic bridging resolution is rare and resources for training algorithms on the problem of bridging resolution are as well.

This thesis therefore introduces new data for bridging resolution in German, the GRAIN corpus, and evaluates the data with regard to the goodness of annotation quality and occurring types of bridging. To ensure the generalizability of the approach, the established corpus DIRNDL is additionally used.

In order to determine the difficulty of the task for the present data, an informed baseline is implemented and evaluated. Furthermore, a rule-based system based on Hou et al. (2014) is created in order to perform bridging resolution. To determine the possibilities of using learning-based models for resolving bridging relations, a gradient boosting model is trained on the same data as the rule-based system.

The rule-based system performs better than the baseline and achieves an F1-Score of 5.3% for DIRNDL and 4.0% for GRAIN. An analysis with oracle lists for the rule-based system shows that many rules do not have any access to the correct antecedent. The gradient boosting model is able to outperform the rule-based system for DIRNDL (F1 = 11.3%), but is not able to generalize on GRAIN. The differences can be explained by looking at the different structure of the corpora and their topic distribution. Furthermore, the results of the gradient boosting model suggest that more training data would greatly improve learning-based approaches for bridging resolution.

Contents

List of Figures

List of Tables

1 Introduction

Bridging might be re-phrased with the term *Inference* and generally describes any type of linguistic phenomenon, where users of a language are forced to use some kind of inference in order to resolve the referent or antecedent of a phrase. See Example (1) from (Clark, 1975; p. 171, Ex. 19).

(1) I walked into the room. The chandeliers sparkled brightly.

In (1), the phrase *the chandeliers* can only be understood, i.e. the referent can only be uniquely identified, by looking into the previous sentence and inducing that *the chandeliers* must be the chandeliers of the specific room mentioned before. By explicitly formulating this reading of Ex. (1), Clark states: "The room mentioned had chandeliers; they are the Antecedent for the chandeliers." (Clark, 1975; p. 171, Ex. 19', underlining by Clark).

Note however, that bridging is also often understood as being (additionally) located on a lexical level, e.g. as a meronymic/holonymic relationship, illustrated in Example (2), again from Clark (1975; p. 171, Ex. 13).

(2) I looked into the room. The ceiling was very high.

Clark distinguishes here a necessary part (*the ceiling* in (2)) and an inducible part (*the chandeliers* in (1)). It can be argued that this addresses the topic of bridging from the wrong angle, mixing up a lexical and a referential level of coreference and information status (Riester and Baumann, 2017; pp. 8–9).

Over the years, many researchers added to the catalogue of Clark (1975), each time introducing their own understanding of bridging (e.g. Prince, 1981; 1992; Poesio and Vieira, 1997; Strube, 1998; Riester and Baumann, 2017). At the same time, the definition of bridging was always strongly connected to attempts in automatically resolving bridging relations, i.e. either finding bridging anaphors, linking bridging antecedents to their anaphors or both. Early attempts in bridging resolution started in the 1990s with Poesio et al. (1997) and continue to this day.

Motivation for the thesis Bridging resolution is a fairly underrepresented branch of research. It is often a part of information status recognition, which itself is rarely done. Moreover, studies on bridging resolution for German data do not exist and most, if not all, bridging resolution research is for English. Furthermore, previous research on bridging resolution has shown very limited success for learning-based approaches, making rule-based systems the current state-of-the-art approaches. Lastly, the variability of different domains for training data is rather negligible.

Building on these observations, I will follow three main questions in this thesis:

Question 1 What kind of challenges does a bridging resolution system face?

Bridging is a very diverse and under-researched phenomenon. It is not entirely clear, what kind of features a system needs in order to successfully resolve bridging relations. This thesis wants to give a thorough analysis of problems that arise when performing bridging resolution, by not focusing on creating state-of-the-art approaches, but by clarifying: what features really help for bridging resolution, what are the limitations of current research in bridging resolution, and lastly what might be necessary for future research to better handle and understand bridging resolution.

Question 2 Are there special requirements for a bridging resolution system when dealing with German and non-standard data?

Bridging relies on basic inferring mechanisms that may be identical throughout languages. Therefore, it does not seem very likely that bridging resolution approaches proposed for English will differ significantly from German. On the other hand, since some bridging phenomena rely on lexical knowledge, it seems reasonable to investigate if German enforces different domains of lexical inference. When talking about different domains, the question of the type of a text also comes to mind. Hence it might also be interesting to investigate in what areas different text types influence the category and realization of bridging.

Question 3 Building on Hou et al. (2014) – can learning-based systems be successfully applied to bridging resolution?

As described in Section 3.4, Hou et al. (2014) report that no significant improvement could be achieved using a SVM-based machine-learning system over their rule-based system. The question arises if this is due to an insufficient amount of training data or because the chosen machine-learning models are not capable of modeling such a complicated problem. Since for this thesis, training data is even more sparse than for Hou et al. (2014), the focus will lie on exploring what is possible at the current stage of bridging resolution research, as far as using machine-learning.

Structure of the thesis Chapter 2 gives an overview of the various approaches towards bridging, their history and annotation guidelines covering different kinds of bridging phenomena. Chapter 3 covers research done in bridging resolution, coming from different points of view. In Chapter 4, the various resources used in this thesis are introduced and, in the case of GRAIN, further analyzed. Chapter 5 presents the two systems evaluated in the thesis, a rule-based system and a gradient boosting system. Chapter 6 gives details about the conducted experiments on these systems and the results. In Chapter 7, a broader discussion of the results is undertaken, interrelating the findings of Chapter 6. Suggestions for future work are looked upon in Chapter 8. Finally, a conclusion is drawn in Chapter 9, summarizing the contents of the thesis.

2 Theories of Bridging

Bridging is a field of research that has not obtained much interest so far. The fact that it was a sub-field of information status early on did not help the cause of studying bridging as a stand-alone phenomenon. The definitions for bridging are highly differing in the research literature and thus guidelines for the annotation of bridging also differ highly in their sets of labels and number of fine-grained cases. This chapter provides an overview of the history of the bridging term and of the different approaches to annotate and classify bridging occurrences.

2.1 A Short History of the Term “Bridging”

Clark (1975) first introduced the term *Bridging* for a form of inference, more specifically for a form of implicature in the sense of Grice (1975). He sets bridging in the context of a new-old information paradigm that he calls the “Given-New Contract” (p. 169) and that is defined by syntactic and intonational features. The Given-New Contract therefore serves as a tacit agreement between a listener and a speaker, in that the speaker tries to convey new information to a listener and assumes that the listener already possesses certain kind of information. Furthermore the listener assumes that the speaker is honestly interested in trying to convey their message and will mark utterance, from which the speaker assumes they are given information, as actually given. As soon as the listener is not able to find a plausible antecedent for this utterance in their own knowledge, they will will try to infer or “bridge” this antecedent from context. Clark (1975) then defines certain types of these bridging processes, namely coreference, set membership, metonymy and more abstract concepts like reason, cause, consequence or concurrence.

Not satisfied with a concept of "shared knowledge", Prince (1981) introduces the term of *Assumed Familiarity*, where a speaker only has an assumption about what a listener might know and forms appropriate linguistic formulations. In her taxonomy, Prince (1981) uses the term *inferrable* for a class of assumed familiarity that comes closest to what Clark (1975) calls "bridging". Prince (1981)'s inferrable entities are inferred by logical or plausible reasoning. Her example that *the driver* can be inferred from the phrase *a bus* suggests that her understanding of plausible reasoning includes context knowledge plus lexical and world knowledge. Additionally, she distinguishes *containing inferrables* from other inferrables; containing inferrables are those entities, where the antecedent of an inferrable is syntactically contained in the phrase of the inferrable itself. In Prince (1992), Prince returns to her definitions of Prince (1981) and uses the term *information status* the first time for her assumed familiarity categorization. Her usage of the terms *inferrable* and *containing inferrable* is still the same though, although she provides some more examples on how to understand them.

Poesio and Vieira (1997) describe the term of *associative anaphora* developed by Hawkins (1978), that is more or less identical with Prince (1981)'s *inferrable*.

Strube (1998) redefines the categories of Prince (1981) by merging them into more coarse-grained information status classes. Prince's *inferrable* and *containing inferrable* are hereby combined with *anchored brand-new* to form the new *mediated* label. This means, inferrable entities are now a combination of Clark (1975)'s bridging, Prince (1981)'s contained inferrable and new discourse entities, which are linked to other discourse entity. Looking into Prince (1981), it is not completely clear what she means by *anchored* and what distinguishes *anchored brand-new* from *containing inferrable*; Prince (1992) does not mention the *anchored* subcategory at all. As all examples in Prince (1981) involve phrases, where an indefinite expression contains a nested phrase with a personal pronoun, like *a guy I worked with*, this leads to the conclusion that the category *anchored brand-new* is created for entities of exactly this type.

The *mediated* category is henceforth used extensively to describe bridging in the scope of information status classification (e.g. Eckert and Strube, 2000; Nissim et al., 2004; Nissim, 2006; Rahman and Ng, 2012).

Only when bridging resolution came into the focus of attention, researchers started to use the term *bridging* again more frequently (e.g. Poesio et al., 1997).

2.2 Annotation of Bridging

This section will cover some of the proposed ways of annotating bridging anaphors that were made over time by the research community. All proposals cover bridging as a smaller part of information status and were used for the approaches of bridging resolution described in Chapter 3 (except for Grishina, 2016).

The *RefLex* scheme (Section 4.3) of Baumann and Riester (2012) was used for the information status annotations of the DIRNDL corpus (Section 4.4). Riester and Baumann (2017) is an improved version of the RefLex guidelines and was used for the annotations in GRAIN (Section 4.5).

2.2.1 Nissim et al. (2004)

Nissim et al. (2004) had a great influence on the annotation of information status and hence also on the annotation of the information status category *bridging.* They use the term *mediated* when talking about bridging, described in Section 2.1. The mediated category takes a place between the categories *new* and *old* and is defined as being newly introduced into the discourse, by also being inferrable from previous context or generally known. They define nine subcategories as follows:

1. **general** Used for generally known nouns, usually proper names
2. **bound** Used for syntactically bound pronouns, Nissim et al. (2004) give the following example: “[...] it’s hard to raise *one child* without **them** thinking they’re the pivot of the universe.” (p. 1024)

3. **poss** All possessive relations inside a phrase
4. **part** Meronymic relation, where the markable is part of a aforementioned object: "*home* ... **the door**" (Nissim et al. (2004), p. 1024)
5. **situation** Same as *part*, but for markables that are part of a situation, set up by previous context
6. **event** Used for markables where the antecedent is an event, e.g. a VP
7. **set** Used whenever the markable is part of a set or a subset or super-set of previous context
8. **func_value** Used for values of entities that represent numerical scales: "I had kind of gotten used to *centigrade temperature* you know – if it's between **zero** and **ten** it's cold." (Nissim et al. (2004), p. 1024)
9. **aggregation** Used for coordinated NPs

Markables labeled as *mediated* are not linked to their antecedents. Nissim et al. (2004) also perform a reliability study by annotating on the Switchboard corpus[1] using their guidelines. The mediated category generally achieves a reliable κ value of 0.8, while the subcategories *part*, *situation* and *event* get the lowest scores and *part* being the worst category in terms of agreement having a κ score of 0.59.

2.2.2 Markert et al. (2012)

Markert et al. (2012) create a corpus of information status annotations on news data from OntoNotes[2], called *ISNotes*[3]. They base their guidelines on Nissim et al. (2004) in that they also distinguish three information status categories: *old*, *new* and *mediated*. However, their subcategories of the *mediated* category look slightly different:

[1] https://catalog.ldc.upenn.edu/LDC97S62

[2] https://catalog.ldc.upenn.edu/LDC2013T19

[3] https://www.h-its.org/en/research/nlp/isnotes-corpus/

1. **comparative** sometimes also called *other-anaphora*, used when a contrast or similarity to a previous mention is established
2. **bridging** Not clearly defined, examples let assume that it is a category for all uncovered cases and for general inference
3. **knowledge** Generally known entities (cf. *general* in Nissim et al., 2004)
4. **synt** Anaphors which include their antecedent syntactically
5. **aggregate** Same as Nissim et al. (2004)'s *aggregation*
6. **func** Same as Nissim et al. (2004)'s *func_value*

Note that *knowledge* is based on Prince (1981)'s *unused* and *synt* on Prince (1981)'s *containing inferrable* category.

The authors also perform an inter-annotator agreement study with three annotators. They report reliable Cohen's κ values for each annotator pair and for each category, with the subcategory *bridging* having the lowest agreement between annotators (κ between 0.6 and 0.7).

2.2.3 Grishina (2016)

Grishina (2016) makes an effort to annotate bridging outside of the scope of information status.

She observes six different broader type of bridging relations:

1. **physical parts - whole** meronymy, e.g. *the telephone* ... **the dial pad**
2. **set-membership** subset or element of set, e.g. *these studies* ... **the main study**
3. **entity-attribute/function** attribute if an entity or function with respect to other entity, e.g. *Mrs. Humphries* ... **the monotonous voice** or *Kosovo region* ... **the government**

4. **event-attribute** bridge to an event, e.g. *the surgical intervention* ... **the operating room**
5. **location-attribute** geographical bridging, e.g. *Germany* ... **in the south**
6. **other** Everything else that also appears to be bridging

Grishina (2016) also executed an annotation study using these categories. The annotators were asked to extend existing annotations of coreference. Furthermore, the possible markables were pre-selected. The task then was to decide on the type of bridging. Grishina (2016) reports F1 scores, yielding 64% for anaphor recognition and 79% for antecedent selection. She also finds that only 17% of bridging anaphors start coreference chains, meaning that bridging anaphors are usually unique in their context. Since she does not prohibit bridging anaphors to be part of coreference chains, she is able to investigate the average length of coreference chains which contain bridging anaphors. Her findings show that half of all coreference chains contain bridging anaphors and these chains are much longer on average than chains without bridging anaphors. She also finds most bridging anaphors to be in close proximity to its antecedent. When trying to transfer the German annotation to Russian data, the lack of definiteness markers caused issues; in general, the transfer was successful though, also to English data.

2.2.4 Riester and Baumann (2017)

In Riester and Baumann (2017)[4], bridging is defined exclusively on definite noun phrases. Furthermore, Riester and Baumann (2017) distinguish a referential and a lexical level of annotation. Meronymic bridging such as Nissim et al. (2004)'s *mediated/part* are therefore not called bridging, but located on the lexical level as *l-accessible-part.* Bridging is only located on the referential level and similar to Prince (1981) separated into *r-bridging* and *r-bridging-contained.* For bridging, a

[4] An updated version of Baumann and Riester (2012).

non-coreferential anaphor, with an antecedent to which it has a unique or prototypical connection, is annotated. This antecedent might also be a more abstract entity such as a clause or sentence. In their guidelines, every markable labeled as *bridging* must be linked to not more than one antecedent. A single antecedent might be an antecedent to multiple anaphors though. A bridging markable can furthermore be marked as +generic or +predicative. The definition of bridging in Riester and Baumann (2017) is more general than in the other guidelines, renouncing subcategorization of bridging relations. On the one hand, this makes the annotation process more vulnerable to inconsistencies, since annotators might disagree more when the specific types of bridging are not explicitly stated; on the other hand, it makes the annotation process more open and enables the annotators to consider a variety of bridging relations that might occur in real life data.

3 Related Work

This chapter aims to give an exhaustive overview of different approaches that have been proposed for automatic bridging resolution. In Section 3.2, only those studies are considered, which yield an insight in bridging resolution going beyond the mere aspect of information status classification.

3.1 Early Bridging Resolution

Poesio et al. (1997) present a system to classify definite descriptions with a focus on bridging relations. They define bridging as either coreference in cases where the head nouns of coreferent phrases are not identical or as a semantic relatedness of a phrase to previous context. While Poesio and Vieira (1997) focused on the former, Poesio et al. (1997) investigate the latter. They decide to use knowledge from WordNet in order to not be forced to restrict the domain of the data, i.e. allow for unrestricted text. They give six types of bridging relations, coming from Vieira and Teufel (1997): Synonymy/Hyponymy/Meronymy relations; reference to proper names; cases where the anaphor head is part of a compound antecedent; abstract anaphors other than pronouns; discourse topics, where the anaphor refers to the unmentioned topic or domain of a text and general inferences like reason or consequence. They make heavy use of WordNet in order to resolve these cases, but are only able to report satisfying results for proper names resolution. They conclude that WordNet is not suitable for identifying the type of bridging resolution due to the large number of false positives. They achieve promising results for abstract bridging anaphor-antecedent pairs by converting the verb of the abstract antecedent into a noun though. Overall they achieve a recall of 65% and a precision of 82%.

Markert et al. (2003) criticize the WordNet approach since it is to reliant on often unreliable knowledge bases, for not all expressions might be present in WordNet and certain information has to be extracted in cumbersome ways (cf. Poesio et al., 1997; for all their link types except synonyms and hyponyms). They run two experiments, one for other-anaphor[1] relations, where all phrases that contain the words *other* or *another* are resolved to an antecedent and one for meronymic bridging. For other-anaphors they make use of a simple structural pattern to extract possible candidates and rank antecedents using a mutual information scoring which is obtained using frequencies retrieved from Google search results. Additionally, they substitute possible antecedents with their named entity category in order to reduce data sparsity. Their scoring classifies 63 out of 120 anaphor-antecedent pairs correctly, compared to a WordNet based solution on the same dataset, which classifies 61 pairs correctly. For meronymic bridging using a different pattern, they also use raw counts from the Google API results and outperform a WordNet-based system by correctly classifying 7 out of 12 cases.

Poesio et al. (2004) were among the first to use machine learning approaches for the classification of mereological bridging relations. They combine salience and lexical information in a fully automatic approach. They combine a WordNet and a Google distance to form their lexical features and choose the utterance distance between anaphor and antecedent as well as focus position of the antecedent as salience features. Since positive examples of bridging are fewer, compared to members of the non-bridging class, they balance the dataset accordingly. The types of machine learning approaches they test include Naive Bayes and Multi-Layer Perceptron. They find the WordNet and Google distance features to perform almost the same, with the WordNet distance performing better when definite bridging anaphors are involved (compared to cases with indefinite bridging anaphors). The salience features perform noticeably better than the lexical distance features. Lastly, the authors train a model on the balanced dataset and test it on an unbalanced dataset. They notice that testing on this larger dataset generally improves the performance.

[1] Sometimes also called *comparative anaphors*.

3.2 Bridging Resolution as Part of Information Status Classification

Since bridging is related to the notion of new-old information in a discourse, many approaches of bridging resolution are actually general methods of information status resolution, where bridging resolution is usually a subtask equal to other new-old annotation tasks. Below, an overview of the literature of information status resolution is presented, as well as the role and interpretation, that bridging was assigned to in these studies.

Nissim (2006), based on the annotations of Nissim et al. (2004), which is described in Section 2.2, trains a model with handcrafted rules as a baseline as well as a decision tree model for three possible categories: *new*, *old* and *mediated*. The mediated label roughly corresponds to the definition of bridging in this thesis. It covers a broader range of cases though, since it is taken from Strube (1998), who created it as a merge of Prince's categories *inferable* and *anchored brand-new* (Prince, 1981; 1992). Hence *mediated* describes all phenomena where a discourse-new entity can be related back to previous mentioned context plus all entities that are generally known to the recipient (i.e. common knowledge). The handcrafted rules in their baseline model are of the nature: check for type of noun phrase, check for string match with previous mentions and check for the type of determiner. Depending on the value, different labels are assigned following a decision tree structure (cf. Nissim, 2006; p. 96). They implement similar features for their decision tree model, with additional features such as length of the NP, grammatical role and time of mention. The authors' decision tree model is able to outperform the baseline for the mediated category, particularly improving the recall for mediated entities. The precision is equally low for both the baseline and the decision tree model, fluctuating around a value of 60%. Overall, the decision tree model achieves an F1-Score of 76.6% for *mediated* on the evaluation set over an F1-Score of 54.5% for the baseline.

Especially interesting seems to be the observation, that the *mediated* category is equally falsely predicted by the model to be either *new* or *old* (cf. Nissim, 2006;

Table 6). This is in favor of an interpretation of bridging being in the middle of old and new information. In order to investigate this hypothesis further, Nissim (2006) conducts a second experiment, where they collapse the categories of *old* and *mediated* into one pairing and *new* and *mediated* into another pairing. It turns out that the combination of *mediated* and *new* as one single category gives better results than the other option, suggesting, that mediated entities are more closely related to new entities than to old ones (Nissim, 2006; p. 98). Furthermore, the learned decision tree ranks information about the type of the determiner higher than the type of the noun phrase itself, emphasizing the importance of definiteness for information status and bridging.

Based on Nissim (2006), **Rahman and Ng (2011)** train an SVM (Support Vector Machine), allowing them to add additional features not present in Nissim (2006). One of their new features captures one of the properties of the mediate category to be generally known by the listener. They create a list of unigrams from the training set and assume these to be generally known entities, if the unigram was part of a discourse entity in the training set. The second new feature is the use of a convolutional tree kernel inside the SVM, which enables them to use subtrees of parses as features and capture more of the syntactic context than Nissim (2006), who only used the grammatical role of the entity. They outperform their re-implementation of Nissim (2006) for the mediated category (72.1%) by scoring an F1-Score of 79.0% using both their new features. Additionally, using only the additional feature which generates a list of known entities together with their re-implementation increases the F1-Score by 3 percentage points over the baseline, resulting in an F1-Score of 75.1% for the mediated category. This demonstrates that even their frankly crude and simple implementation of world knowledge already yields some useful information for the resolution of the mediated category.

In **Rahman and Ng (2012)** the authors take a step further by inducing more fine-grained information status classes. They introduce subclasses for the *old* and *mediated* categories, resulting in a total amount of 16 classes. For *mediated* they introduce 9 subtypes, namely *general, bound, part, situation, event, set, poss, func_value* and *aggregation*. The *part*, *situation*, *event* and *set* subclasses are common bridging categories, while *general* covers the commonly known entities and *poss* and *bound* are dealing with possessive pronouns and bound intra-phrasal possessives (Every cat ate **its** dinner., Rahman and Ng (2012; p. 800)) respectively. The subclass *func_value* is used for cases like values of currencies and temperature and *aggregated* covers coordinations where at least one of the coordinated entities is not *new*.

The authors create a rule-based system as a baseline that requires the presence of coreference information and contains eight handcrafted rules for the *mediated* category and its subclasses. For details on these rules, see Rahman and Ng (2012). Using this baseline system, they report an average F1-Score of 46% for the *mediate* classes on gold coreferences. Interestingly, the precision values for the mediated categories *part*, *situation* and *event* result in 100%. The recall is comparatively low with an average recall of 20% for these three categories. This is not surprising, since these three rules rely on external data like WordNet and therefore it is harder to recognize these cases; if the external resource is not sufficient, but always accurate if the resource yields the needed data.

Secondly, the authors create a multi-class SVM in order to compare it to the rule-based system. The input features are unigrams, markables and also make use of the rules of the rule-based system, either identifying the most probable class for a markable in the training set or directly using the predictions of the rule-based system as a feature. This system outperforms the classification of *mediated* by achieving an average F1-Score of 77.7% for all 9 subclasses on gold coreference. Compared to the rule-based system, this in an improvement of 31.7 percentage points or 69%. Also, the low recall for *part*, *situation* and *event* is improved to an average recall of 62.6%, an improvement of 213%, by naturally dropping the precision from 100% to an average of 97.4%. This means, the learning-based system is able to extrapolate from missing data in the knowledge bases and learns to find similar cases to the ones already present in the external resources.

Cahill and Riester (2012) use a Conditional Random Field model (CRF) for information status classification. They train and test on the DIRNDL corpus (Eckart et al., 2012), where the category *bridging* is a single information status class and not further sub-classified. Different features of their CRF are of the category countable, like number of words in a phrase, boolean, like being a pronoun and descriptive, like determiner type. They notice a poorer performance for the *bridging* class, but do not report actual numbers for single label prediction. They assume this to be a result from the lack of world knowledge in their features and add two semantically informed features based on GermaNet[2]: First the distance of the head noun to its root in GermaNet and second a semantic relatedness measure, that measures the similarity of the head noun to nouns in present and preceding phrases. They report an improvement in accuracy of 3 to 4 percentage points for each of their sets of information status labels and for the dataset of Nissim (2006).

Markert et al. (2012) perform information status classification on written text, unlike Nissim (2006) or Rahman and Ng (2011), who use conversational dialog data. They also view fine-grained information status classification as a necessary prerequisite to successfully resolving general bridging phenomena. They create a new corpus based on OntoNotes, described in Section 2.2. They extend the feature set for a mention of Rahman and Ng (2011) with information about comparative markers, semantic classes and some information about previous mentions. Additionally, they use relations between mentions such as parent-child and precedence relations. They hypothesize that two *mediated* subcategories will benefit especially from the parent-child relationship, since they have a higher syntactic complexity and occur in their corpus quite frequently. For the features of the extended Rahman and Ng (2011) feature set they make use of a SVM and for the relational features they use Iterative Collective Classification. They report significant improvement over Nissim (2006) and Rahman and Ng (2011) for many *mediated* classes, except for *mediated/bridging*. The F1-Score ranges from 1.9 to 18.9%. They suggest to focus on bridging anaphor recognition in future research, instead of only bridging antecedent recognition.

[2] http://www.sfs.uni-tuebingen.de/GermaNet/

3.3 Bridging Resolution using Coreference Resolution

Rösiger and Teufel (2014) perform general coreference resolution and information status recognition on scientific text with a focus on bridging resolution. They distinguish between usual bridging links and self-contained bridging relations (*associative* and *associative (self-containing)* in the vocabulary of Rösiger and Teufel, 2014). Self-contained bridging is a phenomenon such as *The structure of the protein* (Rösiger and Teufel, 2014; p. 47), i.e. the bridging anaphor is syntactically bound to its antecedent. They evaluate on a 16 document corpus covering the topics of computational linguistics and genetics, containing 652 bridging links and 562 self-contained bridging entities. The authors modify a coreference resolution classifier in order to also capture bridging resolution by incorporating WordNet relations such as synonymy, meronymy, hyponymy and topic, whereas the generic coreference resolution system only includes synonymy. For comparison, they also evaluate the unchanged coreference resolution system on the task of bridging resolution, yielding an average CoNLL score[3] of 33.14, showing that a coreference resolution system is also able of capturing some aspects of the task of bridging resolution. Their final system described before significantly outperforms this baseline with an average CoNLL score of 34.88, where the feature of meronymy added the highest increase in performance. Note however that Hou et al. (2014) criticize this approach of using a coreference resolution system, arguing that bridging resolution is not a set problem, since an antecedent might have several links of unrelated bridging anaphors connected to it.

3.4 Unrestricted Bridging Resolution

Hou et al. (2013b) make use of a Markov Logic Network in order to apply local and global features to predict bridging anaphors and suggest the most likely antecedent. Compared to other approaches, they do not limit the type of bridging anaphor or antecedent (a restriction to definite NPs is common). They perform evaluation on their self-created dataset developed specifically for bridging analysis

[3] `https://ufal.mff.cuni.cz/conll2009-st/scorer.html`

(cf. Markert et al., 2012), which comes with 50 documents yielding 663 bridging anaphors and linked antecedents. However, they only perform antecedent resolution while assuming that the respective bridging anaphors are already given information. They report an accuracy of 41.32% for their best model.

Hou et al. (2014) is loosely based on Markert et al. (2012) and Hou et al. (2013a;b)[4]. They evaluate a rule-based system on the same dataset as Hou et al. (2013b), but different from Hou et al. (2013b), they are no longer restricted to antecedent resolution. The eight rules of this system cover semantic categories such as parts of buildings, relative persons, geo-political entities, percentage phrases etc. as well as certain argument-taking NPs, to predict potential bridging anaphors and antecedents. A post-processing step chooses the best anaphor and antecedent, in case different rules output contradicting results, based on a rank of rules developed on a subset of 10 documents. Evaluation on the remaining 40 documents gives an F1-Score of 18.6%, with a recall of 42.9% and a lower precision of 11.9% for the system. The authors also compare their rule-based system to a learning-based approach that incorporates the information of the rule-based system with additional features. This system is not able to significantly outperform the rule-based system, scoring an F1-Score of 18.7% and suggest that the amount of training data is not sufficient for a learning-based approach.

[4] Hou et al. (2013a) perform general IS resolution though.

4 Resources

Given below, the resources used to run the experiments for bridging resolution are described. The main resources for running the experiments are DIRNDL (4.4) and GRAIN (4.5), which hold the annotations for bridging relations. GermaNet (4.1) and SdeWaC (4.2) serve as additional resources for the systems in Chapter 5. The RefLex scheme (4.3) was used as the guidelines for GRAIN and for DIRNDL in a previous version.

4.1 GermaNet

GermaNet[1] (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) is a lexical-semantic net for German, similar to the English *WordNet* (Miller, 1995; Fellbaum, 1998). It organizes words into *synsets*. Each synset represents a set of words that are said to be synonyms. Between these synsets, GermaNet then defines relationships such as hyponymy or meronymy. Currently (effective May 2017, version 12.0), GermanNet consists of 120,032 synsets, 154,814 lexical units in these synsets and 133,652 relations between synsets. GermaNet offers synsets for nouns, adjectives and verbs. Other important properties of GermaNet are that it distinguishes readings of words, decoded in unique identifiers and splits compounds into their constituents (Henrich and Hinrichs, 2011). Throughout all experiments described in this thesis, GermaNet, version 11.0 (release May 2016) was used. In this version, GermaNet consists of 110,167 synsets, 142,814 lexical units and 123,678 relations.

[1] `http://www.sfs.uni-tuebingen.de/GermaNet/`

4.2 SdeWaC

The *Stuttgart German Web as Corpus* (SdeWaC) corpus[2] (Faaß and Eckart, 2013) is a web corpus with a focus on sentences, by only including parseable sentences. This means that a dependency-based parser was applied to all sentences and only those sentences were kept where the parser could provide a full parse on sentence level. SdeWaC is therefore a cleaner subset of deWaC (Baroni and Kilgarriff, 2006); it consists of 44,084,442 sentences, 846,159,403 tokens and 1,094,902 types.

4.3 RefLex

The RefLex scheme (Riester and Baumann, 2017)[3] comprises guidelines for the annotation of information status. It is based on the assumption that information status properties are different for referring and non-referring expressions. While referring expression are given information if the same *entity* already occurred in the discourse, non-referring expressions are given if the same *expression* occurred in the discourse (Riester and Baumann, 2017; p. 3). From this assumption, Riester and Baumann (2017) develop the distinction between a referential level (r-level) and a lexical level (l-level). The labels for the r-level and the l-level are presented in Table 4.1 and 4.2, respectively. On the r-level, all noun phrases and prepositional phrases receive a label; only relative pronouns are not labeled, since their reference can be identified syntactically. If an anaphor refers to a markable other then a noun phrase or prepositional phrase, this antecedent is labeled as an abstract antecedent with a special label. Abstract antecedents are usually verb phrases, sub-clauses or (multiple) sentences. Sometimes, a markable is interrupted by an intervening element. In this case, the parts of the discontinuous markable are linked to each other with a special link. In another case, when an anaphor refers back to multiple antecedents, the special link *aggregated* is used to distinguish these cases from pure coreferential ones.

[2] `http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/sdewac.en.html`

[3] Riester and Baumann (2017) were already discussed in Section 2.2.4, but only under the aspect of bridging annotation. This section describes the full annotation guidelines in the context of information status annotation.

Label	**Description**
r-given-sit	Symbolically deictic, referent not part of text, e.g. first person pronoun, time reference, locations
r-environment	gestural deictic, referent is identified by pointing or gazing
r-given	coreferent expressions
r-given-displaced	coreferent expressions, whose antecedent is more than five clauses away
r-cataphor	coreferent to an antecedent that follows the expression (i.e. postcedent)
r-bridging	see Section 2.2.4
r-bridging-contained	see Section 2.2.4
r-unused-unknown	expression, which is identifiable by its own, not generally known, typically phrases with embedded phrases
r-unused-known	generally known entities, typically proper names, e.g. *the Pope* or *Germany*
r-new	discourse-new, non-unique and indefinite expressions
r-expletive	non-referring, expletive expressions, e.g. *it* in *it starts raining*
r-idiom	non-referring, idiomatic expressions, e.g. *for Example*
+generic	optional attribute for generic expressions
+predicative	optional attribute for the predicate in a predicative formulation

Table 4.1: Overview of the r-level labels in RefLex.

Label	Description
l-given-same	repetition of identical content word
l-given-syn	mention of synonym to previous expression
l-given-super	mention of hypernym to previous expression
l-given-whole	mention of holonym to previous expression
l-accessible-sub	mention of hyponym to previous expression
l-accessible-part	mention of meronym to previous expression
l-accessible-stem	repetition of stem or compound part
l-new	newly introduced lexical material

Table 4.2: Overview of the l-level labels in RefLex.

4.4 DIRNDL

The Discourse Information Radio News Database for Linguistic Analysis (Eckart et al., 2012; Björkelund et al., 2014), or DIRNDL in short, is a corpus developed at the "Institut für Maschinelle Sprachverarbeitung, IMS" (Institute for Natural Language Processing) at the University of Stuttgart in Germany. It combines different layers of automatic and manual annotation and most interestingly brings together prosodic information with information status annotations.

The bridging annotations follow Baumann and Riester (2012), which replaces a previous annotation scheme using Riester et al. (2010).

In total, 655 bridging pairs are annotated in DIRNDL, making it a fruitful resource for bridging resolution in German.

4.5 GRAIN

The German Radio Interviews Corpus or short GRAIN (Eckart and Gärtner, 2016; Schweitzer et al., 2018) is developed, like the DIRNDL corpus, at the IMS and under development. It was initiated with the goal of creating a silver-standard corpus, meaning that several types of non-standard data are annotated multiply, both manually and automatically, so that “a level of annotation quality between a manually created gold standard and the unchecked output of automatic processing” (Eckart and Gärtner (2016), p. 91) is achieved. The different annotations of the same linguistic layer can then be mapped and compared against each other.

GRAIN consists of 23 collected broadcast interviews conducted at the German broadcasting station SWR (Südwestrundfunk). The data comes from a program of the SWR called “Interview der Woche”[4] (transl. Interview of the week) covering the years 2014 and 2015. The SWR offers transcriptions for downloading, which were additionally modified. These modifications covered insertions of words not included in the SWR transcripts, but which were uttered by the speaker such as repairs, slips of the tongue or missing words. Additionally, overlaps of speaker utterances were included into the transcript and thus making the transcripts more applicable for research of speech-related topics.

The dataset also comes with audio data of the interviews and thus, experiments on the interaction of information structure, information status and prosody can be investigated.

4.5.1 Quantitative Analysis of Information Status Annotations

In order to get a better understanding of the information status annotations in GRAIN, a quantitative analysis of the annotations was performed. The results are presented in Figure 4.1.

Common categories like *r-given* and *r-new* dominate the amount of annotated categories. *R-bridging* lies, quantitatively spoken, in the middle of the categories, with 274 bridging pairs and 297 markables being annotated as bridging. The difference stems from the fact that not for every bridging anaphor a reasonable antecedent could be found. This is considerably less than for DIRNDL, making learning-based approaches complicated, since they rely on a sufficient amount of data in order to optimize their performance.

[4] `https://www.swr.de/swr2/programm/sendungen/interview-der-woche/startseite-mit-vorschau/-/id=659202/did=13778120/nid=659202/r2sjey/index.html`

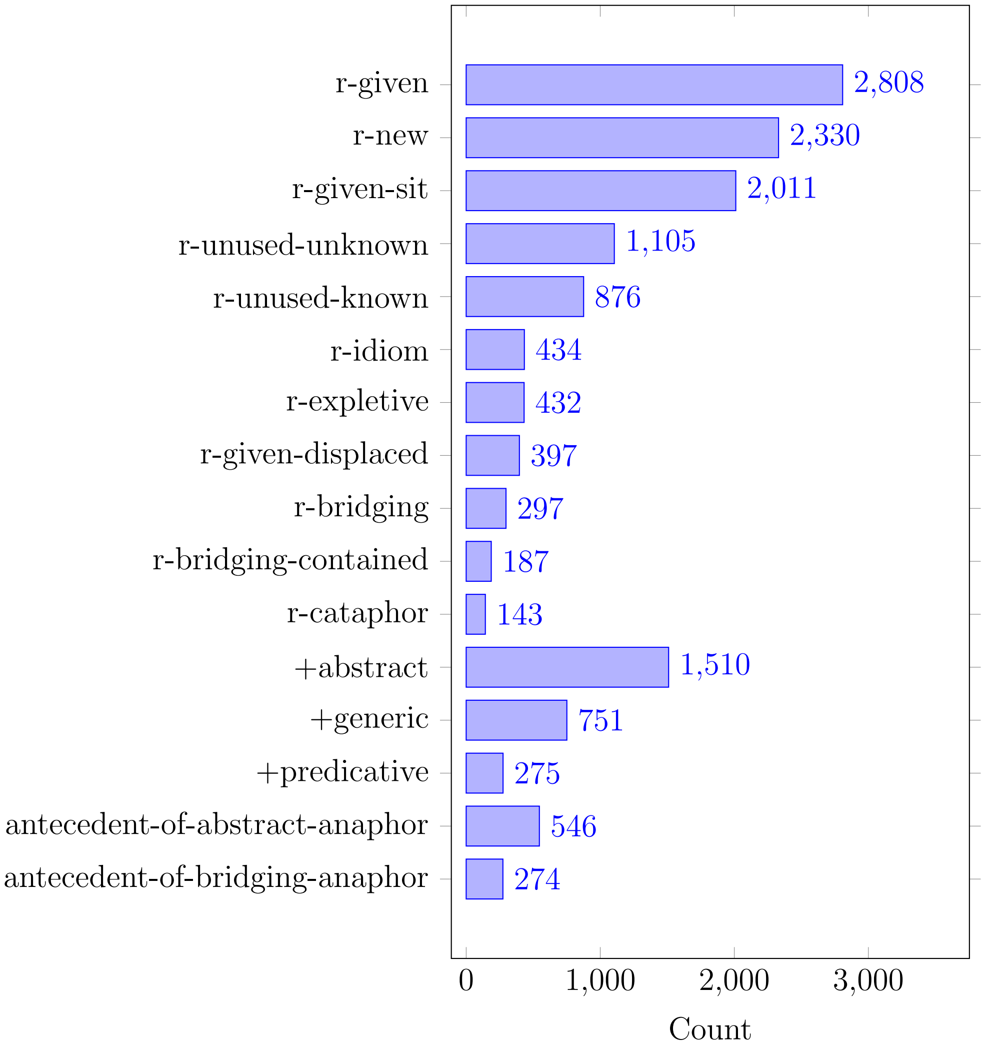

Figure 4.1: Distribution of information status categories, attributes and antecedents in GRAIN.

An important constraint of Hou et al. (2014) is to only search for antecedents not more than a certain number of sentences away. For GRAIN, the median for the distance between anaphor and antecedent is 2.4 and the average is 4.9. The difference is explained by some outliers, with the maximum distance being 48 sentences. Since in every document there is at least one antecedent more than 5 sentences away from the anaphor, it seems to make sense to implement an algorithm that tries to search for an antecedent in a smaller distance first and then goes back as much as necessary in order to find an antecedent.

Several antecedents function as an antecedent for multiple bridging anaphors. In order to check if it is sensible to implement a feature that prefers antecedents with multiple anaphors over other antecedents, the ratio of all *multi-antecedents* to the number of all antecedents was computed, which turns out to be 29:220. This is approximately a ratio of 1:7, that means that every seventh antecedent is a *multi-antecedent*. Therefore it indeed seems sensible to let a model take advantage of this information.

4.5.2 Inter-Annotator Agreement Analysis

In order to get an understanding of the goodness of the annotations in GRAIN, an inter-annotator agreement study was performed, using Cohen's κ (Cohen, 1960) and Fleiss' κ (Fleiss, 1971). Fleiss' κ is a useful measure when comparing more than two annotators for an annotation task, since Cohen's κ is not able to perform such agreement comparisons. Cohen's κ is a widely used measure for annotator agreement and hence, all experiments were additionally performed using Cohen's κ, where applicable. Since the values of Cohen's κ only differed by an amount of 0.01 points compared to Fleiss's κ, all values are reported using Fleiss's κ only.

Table 4.4 shows the number of documents annotated by all the annotators involved, the total number of markables contained in these documents, the κ value and the Z-test value for different pairings of annotators (A, B, C, D and E). All values were significant with $p = 0$. The table shows that the information status annotations are very reliable, ranging from $\kappa = 0.6$ to 0.8.

Annotators	**#Documents**	**#Markables**	κ	**z**
A+B	5	1808	0.822	81.0
A+C	5	1917	0.788	79.1
A+D	5	1608	0.734	66.5
A+E	4	1181	0.654	51.6
B+C	6	2282	0.759	84.2
B+D	3	775	0.696	42.6
B+E	3	1042	0.635	44.5
C+D	2	510	0.651	32.5
C+E	1	243	0.712	23.5
D+E	6	2038	0.733	74.6
A+B+C	3	825	0.783	83.5
A+B+D	2	383	0.711	49.0
A+B+E	1	173	0.756	34.1
A+C+D	2	417	0.707	50.3
A+C+E	1	196	0.753	35.0
A+D+E	2	235	0.775	41.1
B+C+D	2	390	0.688	48.6
B+C+E	1	176	0.737	33.1
A+B+C+D	2	352	0.729	67.5
A+B+C+E	1	153	0.777	44.7
B+C+D+E	1	171	0.742	46.6
A+B+C+D+E	1	149	0.770	56.1

Table 4.4: Number of shared documents, number of markables with same span, Fleiss' κ and Z-Test values for all annotator pairs over all categories. $p = 0$ for all κ and Z values.

In Table 4.5, a detailed analysis of the single information status categories is given. The entry $\{A, B, C, D, E\}^2$ is calculated by taking all annotations of all annotators and concatenating them. It therefore serves as an approximation to an average value of agreement for the whole corpus over all annotators. Its value lies at $\kappa = 0.7$ and shows that the overall agreement is very reliable. The annotation of the category *r-bridging* reveals to be one of the lowest agreements, which is not surprising, since the annotation of bridging usually obtains the lowest values of agreement in information status annotation (see e.g. Poesio and Vieira, 1997; Markert et al., 2012). This means however that the bridging annotations are potentially more unreliable than for the other categories and hence a system might give lower performance, because the cases of bridging were not always correctly classified.

Annotators	κ	**z**
$\{A, B, C, D, E\}^2$	0.738	195.000
Category		
antecedent-of-abstract-anaphor	0.941	108.893
r-bridging	0.356	41.238
r-bridging-contained	0.286	33.114
r-cataphor	0.568	65.779
r-expletive	0.833	96.472
r-given	0.785	90.912
r-given-displaced	0.489	56.664
r-given-sit	0.906	104.892
r-idiom	0.672	77.745
r-new	0.804	93.127
r-unused-known	0.591	68.376
r-unused-unknown	0.479	55.476

Table 4.5: Fleiss' κ, Z-Test values and p values for all combined annotator pairs and all categories. $p = 0$ for all κ and Z values. Number of subjects: 13,404.

In order to investigate how reliable the bridging annotations actually are, the κ

values for single annotator pairs are calculated and presented in Table 4.6. It shows that the poor results for bridging agreement come from specific annotator pairs. The highest values of 0.5 and 0.6 are comparable to the results of Markert et al. (2012), though, who achieve agreement for their bridging category of $\kappa = 0.6$ to 0.7.

Annotators	κ	**z**	**p**
A+B	0.602	25.615	0.000
A+C	0.524	22.952	0.000
A+D	0.225	9.039	0.000
A+E	0.213	7.306	0.000
B+C	0.458	21.867	0.000
B+E	−0.010	−0.329	0.742
D+E	0.269	12.142	0.000

Table 4.6: Fleiss' κ, Z-Test values and p values for the r-bridging category for all annotator pairs with more then 1,000 matching markable spans.

Another interesting aspect of agreement evaluation is to investigate, how good the agreement is for annotating an antecedent for a specific anaphor? In order to do so, two separate approaches are taken.

First, for each anaphor that two annotators annotate as *r-bridging*, the κ value for choosing the same antecedent for these bridging anaphors is calculated. The results are reported in Table 4.7. Three pairs of annotators seem of interest, since they share the highest number of mutually annotated bridging anaphors: A+B, A+C and B+C. The κ values for these pairs seem very promising, considering that antecedent selection is a difficult task to agree on. They range from $\kappa = 0.635$ to $\kappa = 0.864$.

Another way of measuring the agreement of links is using the Jaccard index and obtain the similarity in sets of anaphor-antecedent pairs. This method has the advantage that it is easier to interpret it, compared to the κ measure. The Jaccard

Annotators	#Links	κ	z	p
$\{A, B, C, D, E\}^2$	84	0.637	47.2	0.000
A+B	23	0.635	13.9	0.000
A+C	20	0.635	13.7	0.000
A+D	7	0.364	2.89	0.004
A+E	4	0.111	0.528	0.597
B+C	23	0.864	19.3	0.000
B+D	1	−1.000	−1.000	0.317
D+E	6	0.25	1.76	0.078

Table 4.7: Number, Fleiss' κ, Z-Test values and p values for bridging links. Coreference between antecedents was not considered.

index J of two sets of anaphor-antecedent pairs A and B is defined as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.1}$$

Using the validated version of an interview as gold standard and measuring the agreement with one of the five annotators results in values given in Table 4.8. As can be seen, the results are once again mixed, ranging from a good agreement of 0.36 to a rather poor agreement of 0.11. This shows that finding a reasonable antecedent seems to be a rather difficult task in GRAIN, posing challenges for a bridging resolution system.

4.5.3 Types of Bridging

Lastly, an investigation of the different types of bridging occurring in GRAIN is executed. Such an investigation is of special interest, since bridging is a highly diverse phenomenon and a system for bridging resolution will greatly benefit from information about what types of bridging it is dealing with.

Annotator	*J*
A	0.28
B	0.17
C	0.36
D	0.11
E	0.11

Table 4.8: Jaccard index for comparing a validated annotation with an annotator annotation.

- **Omitted Possessives**:

  The RefLex category *r-bridging-contained* from Riester and Baumann (2017) covers bridging anaphors which syntactically contain their antecedent. Often, the antecedent is realized as a personal pronoun (e.g. The townhall was constructed last week. Let us visit ***its* opening ceremony**). In cases were no personal pronoun is used, usual bridging applies (*The townhall* was constructed last week. Let us visit **the opening ceremony**). Such cases can also be found in GRAIN:

  (3) *Sigmar Gabriel* mit einer Frau **an der Seite**, die nicht Angela Merkel heißt, würde sicher auch gut tun.

- **Prototypical**:

  Riester and Baumann (2017) describe the bridging anaphor as being set in a context, "in which [it] plays a unique and perhaps even prototypical role" (p. 8). Such prototypical relationships can also be found in GRAIN:

  (4) a. Die werden durchgescrollt und *das Kästchen* gesucht, wo man **den Haken** dran machen kann.

  b. Aber das ist letztlich Aufgabe *des Generalbundesanwalts*, welchen Weg er jetzt beschreiten will **in der weiteren Ermittlungsarbeit**.

c. Aber jetzt zum Beispiel am Bürokratiewahnsinn *in den Heimen*, der **den Pflegekräften** die Zeit für **die Patienten** nimmt, ändert sich ja dadurch erstmal nichts.

- **Comparative bridging**:

  Comparative or "other-bridging" has been investigated in Markert et al. (2003). Again, we find such instances in GRAIN:

  (5) Und ich weiß, dass es *Meinungen, wie ich sie vertrete*, in allen Bundestagsfraktionen gibt. Und es gibt auch in allen Bundestagsfraktionen **die andere Auffassung**.

- **Building-part**:

  Hou et al. (2014) obtain good results by implementing a rule for finding meronymic building-part bridging. Following is an example from GRAIN for building-part bridging:

  (6) [...], da ist dieser Geist *eines Hauses* wichtiger als die Frage, ob **die Fassade** zuletzt neu gestrichen wurde.

- **Professional Role**:

  Hou et al. (2014) also propose this category and suggest to find pairs of professional roles as an anaphor and organisations as the antecedents. Unfortunately, only one such construct can be found in GRAIN, where the antecedent is not even an organisation:

  (7) In Deutschland und auch weltweit wird über den Absturz *der Germanwingsmaschine* in den Alpen diskutiert. **Der Co-Pilot** soll dieses Flugzeug absichtlich gegen das Bergmassiv gesteuert haben.

- **Country-part**:

  Since the discourse topic of GRAIN is of political nature, often a part of a country is being issued. The same phenomenon can be observed for DIRNDL. Following an example from GRAIN:

(8) Es gibt ja die These, die nicht wenige vertreten, weil *in Jordanien* **die Bevölkerung** über 80 Prozent palästinensisch ist, sollte daraus ein solches Gebilde bestehen.

- **Politics-part**:

  As said before, both DIRNDL and GRAIN deal with a political domain. Hence, many bridging anaphors are related to politics:

  (9) Es ist überhaupt keine, wenn Sie so wollen, Linie *in dieser Regierung* mehr. [...] Netanjahu könnte das schaffen, indem er Lieberman **aus der Koalition** entlässt und eine andere Partei aufnimmt, die für Verhandlungen eintritt.

- **Purely Contextual**:

  Contextual bridging is a problem for bridging resolution systems, since the amount of inference in order to resolve the bridging pairs can be immense, even for human annotators. Following an especially complicated example with crossing antecedent links and an abstract antecedent:

  (10) Wir hatten hier ein sehr starkes Unternehmernetzwerk, das sich insbesondere *in Sachen Eigenstrom*$_1$ stark gemacht hat. Es gab eine eigene „Mainzer Erklärung“ mit Mittelständlern, die bereits Eigenstromanlagen betreiben und auch in der Zukunft in solche investieren wollen. Und die haben natürlich gesagt, sie wollen nicht darauf verzichten, sie wollen weiter *da rein investieren*$_2$, und sie wollen es auch entsprechend verrechnen können. Und nun sind natürlich dafür **die Bedingungen**$_2$ äußerst verschlechtert worden. Das heißt, die Wirtschaft hat an dieser Stelle stark protestiert, mit mir zusammen. Wir haben das vorgetragen, wir haben ein bisschen was erreicht, aber nicht alles. Und vor allen Dingen **die Situation**$_1$ ist unsicher.

- **World Knowledge / General Education**:

Also this type of bridging is difficult to resolve, since world knowledge is required in order to link an anaphor to its antecedent. For GRAIN and DIRNDL, which contain rather contemporary news and interviews, the type of world knowledge is often also quite up-to-date and specific:

(11) a. Da ging es unter anderem auch natürlich mit dem EU-Wettbewerbskommissar Alumnia über mögliche Beihilfeverfahren der EU *beim Nürburgring*. [...] Dann kam **die Insolvenz**.

b. [...], dass ich nicht nach Sotschi fahren konnte, obwohl ich als Sportlerin da wirklich sehr, sehr gerne jetzt auch in der neuen Rolle hingefahren wäre, um **die Sportler** zu unterstützen.

Table 4.9 shows an overview of the various types found in GRAIN. The types *prototypical* and *context* are clearly prevailing. In addition to these types, other types occur, such as *building-part* and *professional role* (proposed in Hou et al., 2014) and *country-part* (proposed in Rösiger, 2018, in preparation). Also some instances of comparative bridging can also be found. However, these additional types are not dominant.

Type	**Count**
Building-part	3
Professional role	1
Country-part	19
Prototype	92
World-Knowledge	23
Context	101
Comparative	8

Table 4.9: Types of bridging in GRAIN and their counts.

5 Systems

5.1 Rule-based System

The foundation of the rule-based system is based on Hou et al. (2014). They define eight rules, all covering different types of bridging. For a given markable, each rule predicts if it is a bridging anaphor or not, and if so, returns an anaphor-antecedent pair. In case the given markable does not comply with the constraints of the rule for a bridging anaphor or in case no suitable antecedent could be found for an anaphor candidate, the rule returns nothing.

The re-implementation of Hou et al. (2014) for German data was provided by Ina Rösiger and is described in Rösiger (2018, in preparation). The additional rules 9 and 10 also come from Rösiger (2018, in preparation), in addition to the calculation for the argument-taking ratio and the semantic connectivity, described in the next two sections. The system of Rösiger (2018, in preparation) was further modified and extended with the rules 11–13.

5.1.1 Argument-Taking Ratio

Several rules use an *argument-taking ratio*, as described in Hou et al. (2014). The general idea is that bridging anaphors expect their antecedent as an elliptic argument and the introduced inference can be resolved through this implicit argument-ship (Löbner, 1985; p. 304). The argument-taking ratio is computed using SdeWaC (4.2) and three regular expressions involving Part-of-Speech tags:

1. $\mathrm{N}_{\mathrm{target}}$ P (D) (ADJ)* N

2. N_{target} D (ADJ)* N
3. POSS N_{target}

Expression 1 searches for all occurrences of a noun followed by a PP, e.g. *Bücher in der Bibliothek* (books in the library). Expression 2 handles nouns modified by genitives, e.g. *Dach des Hauses* (roof of the house; note that genitive can be expressed using an article in German, in contrast to the preposition *of* in English). Expression 3 covers possessive modifications, e.g. *sein Sohn* (his son). The count of the target noun's head in these patterns is then divided by the total count of this noun's head in SdeWaC and the result is set as its argument-taking ratio $ATR(h)$:

$$ATR(h) = \frac{C_{hp}}{C_h}, \tag{5.1}$$

where h is the head of the target noun, C_{hp} is the count of h occurring in a modifier pattern and C_h is the total count of h. Thus, ATR ranges from 0 to 1.

In order to reduce sparsity issues, the compound splitter Compost (Cap, 2014) is used and the argument-taking ratio is only computed for the head of the compound. Doing so, for DIRNDL, 97.9% and for GRAIN, 98.3% of all nouns receive an argument-taking ratio, respectively.

5.1.2 Semantic Connectivity

Semantic connectivity is another used metric in the rule-based system. While argument-taking ratio aims to better predict bridging anaphors, semantic connectivity is a measure of how connected a potential antecedent is to a given anaphor. Similar to the argument-taking ratio, a pattern N P (D) (ADJ)* N is applied on SdeWac, but this time the log-transformed count (using Dunning, 1993) for both noun heads occurring together in this construction is being stored:

$$SC(h, m) = H(C_{hm}) \tag{5.2}$$

where the semantic connectivity is $SC(h, m)$ of a head h and a modifier m. C_{hm} is the count of h occurring together with m in the pattern described above and H is a function to compute the Dunning log-likelihood of the count, as described in Dunning (1993). Compost is used to perform the compound splitting.

For semantic connectivity, using a compound splitter also brings disadvantages. Sometimes, the relevant semantic information might be located in the modifier of the compound. For example, to connect the anaphor *mit möglichen Ermittlungsarbeiten* (with possible investigative work) to the antecedent *ein Generalbundesanwalt* (German title, comparable to a Chief Federal Prosecutor in the United States), it would be beneficial to take the modifier *Ermittlung* (investigation) rather than the head *Arbeit* (work) to correctly connect it to *Anwalt* (prosecutor). Since it is not straightforward to identify these cases, the head of the compound is used for all instances and possible unfavorable head comparisons are condoned. After applying the compound splitter, 45.9% noun-pairs in DIRNDL and 24.0% noun-pairs in GRAIN are assigned a semantic connectivity score.

5.1.3 Rules

In the following, the eight rules used in Hou et al. (2014) are described in detail. Several changes were applied in order to make the rules applicable for German bridging resolution. In Table 5.1, an overview of these rules can be found.

Rule 1: Building-parts This rule uses 1,149 nouns describing building parts, extracted from GermaNet, in order to predict bridging anaphors. The approach for extracting the nouns was as follows: The terms *Gebäude* (building) and *Haus* (house) were taken as the root nouns. Then, all meronymical relations to these terms were retrieved. Finally, the building-part nouns were all the meronyms and all hyponyms of these meronyms. This is of course different from Hou et al. (2014), who retrieve English nouns. Only if a noun is on this building-parts list and only if it has no further nominal premodifications, is it considered a bridging anaphor. Additionally to Hou et al. (2014), also all markables with PP post-modifications are filtered out. An antecedent candidate is then the markable with the highest semantic connectivity to the proposed anaphor. This emphasizes the intuition that the antecedent for a building-part anaphor should be in the immediate context.

Rule 2: Relative person NPs From GermaNet, a list of 275 relative persons is extracted. This time, only hyponyms are considered, coming from the roots *Verwandter* (relative) and *nichtehelicher Sexualpartner* (extramarital sexual partner). Again, only markables whose heads are on that list are being considered, while pre-modified and post-modified markables are filtered out. Additionally, only markables with an argument-taking ratio higher than a certain threshold are taken into account. This helps in filtering out relative person nouns which are used mostly generically (e.g. *children* as in “Children are loud”). The proposed antecedent is then the closest person NP[1] or pronoun.

Rule 3: GPE job title NPs From GermaNet, a list of 439 Geo-Political professional roles is extracted, using all hyponyms of the terms *Regierungsbeamter* (government official) and *Repräsentant* (representative). If a markable is on this list and if it is not post-modified and not modified by a country or organization, it is considered a bridging anaphor. The reason for excluding country and organization modifications is to filter out markables of the information status type *r-bridging-contained*, and to find bridging anaphors whose antecedent occurs separated. For Geo-Political Entity (GPE) job titles, the antecedent is usually a country or organization. The antecedent is the preceding markable with the highest count in the document, as a measure of salience.

Rule 4: role NPs A list of general role nouns is extracted from GermaNet, using all hyponyms of the roots *Berufstätiger* (employed person), *Vorgesetzter* (superior) and *professioneller Mensch* (professional person), resulting in 6,990 nouns. Filtering out markables follows the same steps as for rule 3, but additionally, proper names are also filtered out. A markable is then chosen as antecedent, if it is an organization.

[1] Hou et al. (2014) suggest to choose the antecedent from the first preceding non-relative person NP. This is however confusing, since it does not seem necessary to exclude nouns on the relative person list from being an antecedent. Therefore, it was not implemented.

In case several such markables exists, the most salient is chosen as antecedent, again using its frequency in the document as a measure of salience.

Rule 5: percentage NPs All markables which contain the term *Prozent* (percent) or the symbol % and appear in the subject position are set to be a bridging anaphor. The antecedent should modify another percentage expression with the term *von* (of) or as a genitive modifier. Hence such modified percentage expressions are not bridging anaphors, but part of the antecedent and therefore, all percentage expressions which are modified in such a way are also excluded from the set of potential bridging anaphors.

Rule 6: other set member NPs Since all bridging anaphors in RefLex are definite expressions, this rule was not implemented. In Hou et al. (2014), it covers so called "other"-bridging or comparative bridging: a number expression or indefinite pronoun refers to a member of a set, introduced by the antecedent. Hou et al. (2014) choose the closest plural subject NP from the previous two sentences to be the antecedent, or the closest plural NP, in case no subject is plural.

Rule 7: argument-taking NPs I Rule 7 is a more general and complex rule than the ones before. Hou et al. (2014) base it on the assumption that argument-taking nouns behave in a similar manner as arguments of verbs (Laparra and Rigau, 2013), namely that arguments of different instances of the same noun (or verb) are often identical. A markable is chosen as a bridging anaphor if is not pre- or post-modified and if its argument-taking ratio is above a certain threshold. Then, all occurrences of the head of this anaphor appearing in patterns as described in 5.1.1 are taken and the most recent one is chosen as the antecedent.

Rule 8: argument-taking NPs II Hou et al. (2014) follow the observation from Prince (1992) that bridging anaphors frequently appear in the subject position. A markable is chosen as a bridging anaphor, if its argument-taking ratio is above a certain threshold, if it is not modified by another noun and if it appears in the subject position. The antecedent is then the closest markable with the highest semantic connectivity.

Rule 1	building part NPs *The house* ... **The basement**
Rule 2	relative person NPs *She* ... **The husband**
Rule 3	GPE job title NPs *Japan* ... **The prime minister**
Rule 4	role NPs *University of Stuttgart* ... **The professor**
Rule 5	percentage NPs *22% of the firms* ... **Seventeen percent**
Rule 6	other set member NPs *Several problems* ... **One**
Rule 7	argument-taking NPs I → Argument taking ratio > threshold
Rule 8	argument-taking NPs II → Argument taking ratio > threshold and in subject position

Table 5.1: Overview of the rules used in Hou et al. (2014).

Five additional rules are added to the system in order to better adapt to the data of DIRNDL and GRAIN. Rule 9 and 10 are taken from Rösiger (2018, in preparation). Table 5.2 provides a short overview.

Rule 9: country-part NPs In DIRNDL and GRAIN, many bridging anaphors are parts of countries, due to the news and interview domain of the two corpora. A list of 188 country-parts is extracted from GermaNet, using the hyponyms of all meronyms of the term *Land* (country). A markable is considered to be a bridging anaphor, if it occurs on this list and not pre- or post-modified and not demonstrative. For the antecedent detection, a list of 518 words marked as countries is retrieved from

GermaNet, taking all hyponyms of the term *Staat* (state). The most recent markable on this countries list is then taken as the antecedent. Additionally, only markables with word length 1 or 2 are taken into consideration, in order to either return single nouns or preposition-noun combinations (*Deutschland* or *in Deutschland*). If a country occurs as the first mention in the document, it is taken as the antecedent, being the most salient entity.

Rule 10: argument-taking NPs III This rule is a re-implementation of Hou et al. (2014)'s rule 8, by removing the constraint that a potential anaphor has to be the subject of a sentence. In contrast to finding the antecedent in rule 8, rule 10 chooses a markable as the antecedent, if its semantic connectivity is above a certain threshold. This way, rule 10 is more flexible than rule 8, since there is also the option to not pick an antecedent, if no suitable candidate can be found.

Rule 11: politics NPs Because of the domain, many bridging anaphors in DIRNDL and GRAIN are related to political issues. In an attempt to retrieve these anaphors, a list of 1,269 terms is taken from GermaNet, choosing all hyponyms of the terms *staatliche Institution* (public institution) and *Politik* (politics). If not modified, the antecedent is chosen to be the markable with the highest semantic connectivity.

Rule 12: exclude r-unused-known Rule 12 aims to improve the recognition of bridging anaphors. Since the pre-processing step already removes all indefinite markables, and hence markables of the information status category *r-new* and all coreferent markables and therefore markables of the information status category *r-given*, there are only two more categories, which are usually short and unmodified: *r-bridging* and *r-unused-known*. *r-unused-known* contains all entities which are definite and generally known to the annotator. In order to filter out these markables, it is assumed that markables of the category *r-unused-known* appear more frequently in a document, e.g. talking about Germany several times during a news report or radio interview. Bridging anaphors on the other hand are unique in the context of their antecedent and will most likely only appear once in a document. Hence, all markables, which appear more than once in a document are excluded from being a bridging anaphor. The usual filtering out of pre- and post-modified markables is also applied. The antecedent is then chosen as in rule 10.

Rule 13: cosine similarity Next to semantic connectivity, the cosine similarity between vector representations of words also might offer certain semantic relationships between anaphor and antecedent. Therefore, rule 13 uses the cosine similarity[2] of vectors trained on SdeWaC. For this purpose, the vector space described and implemented in Dhar (2018; p. 25 ff.) was used. In order to build the vectors, Dhar (2018) used a co-occurrence window of 21, i.e. 10 context words to the left and to the right of the target word were considered. The vectors contain the PPMI-transformed counts of the context words. Additionally, all vectors were reduced to 300 dimensions using singular-value decomposition (SVD). The markable with the highest cosine similarity is chosen to be the antecedent. Apart from using the cosine similarity, instead of semantic connectivity, rule 13 is identical to rule 10.

Hyper-Parameters Several rules require hyper-parameters that define certain constraints of a rule. These hyper-parameters are the maximal sentence distance that a rule is able to look back (maxSentDist), the minimal argument-taking ratio that an anaphor should have (argTakingRatioThreshold) and the minimal semantic connectivity that an anaphor-antecedent pair should have (semConThreshold). Note that several rules make use of semantic connectivity, but only rule 10 and 12 make use of a threshold; the other rules always pick the antecedent with the highest semantic connectivity. Rule 3 and 9 require no hyper-parameters. The parameters can be set by hand, but are estimated on a held-out development dataset. Table 5.3 gives an overview about which rules require which hyper-parameter.

Sibling anaphor treatment Hou (2016) found syntactically related bridging anaphors to often share the same antecedent. She calls these anaphors *sibling anaphors*. In DIRNDL and GRAIN, there are also several bridging anaphors with identical antecedents (a ratio of ≈ 1:3 for DIRNDL and GRAIN). In order to integrate this

[2] For more information on the general use of cosine similarity and vector spaces as used in distributional semantics, see e.g. Sahlgren (2006); Padó and Lapata (2007)

Rule 9	country-part NPs *Germany* ... **The border**
Rule 10	argument-taking NPs III → Argument-taking ratio > threshold and high semantic connectivity to antecedent
Rule 11	politics NPs *Bei der Doppelspitze* ... **Der Grünen** *Die ersten Prognosen* ... **der Landtagswahlen, nächstes Jahr** *die Fraktionsspitze* ... **der Links-Fraktion**
Rule 12	exclude *r-unused-known* Exclude markables of category *r-unused-known* by only considering markables of count 1 in the document
Rule 13	cosine similarity Use cosine similarity instead of semantic connectivity

Table 5.2: Overview of the new rules.

Rule	Hyper-Parameter
1	maxSentDist
2	maxSentDist, argTakingRatioThreshold
3	–
4	maxSentDist
5	maxSentDist
6	maxSentDist
7	maxSentDist, argTakingRatioThreshold
8	maxSentDist, argTakingRatioThreshold
9	–
10	maxSentDist, semConThreshold
11	maxSentDist
12	maxSentDist, semConThreshold
13	maxSentDist

Table 5.3: The hyper-parameters of the rule-based system.

information into the rule-based system, every rule will prefer antecedents that were already predicted by the same rule previously. If a rule predicts a markable to be an antecedent for a specific bridging anaphor which was already predicted to be an antecedent for a previous anaphor, this markable is chosen as the antecedent. In case of multiple such markables, the most recent one is predicted to be the antecedent.

Post-processing In order to avoid conflicts between rules, e.g. if multiple rules choose different antecedents for the same anaphor, the precision[3] for each single rule is evaluated and the rules are ordered according to their precision. A rule with a higher precision receives a higher evaluation precedence over other rules when applied to the same markable.

5.2 Learning-based System: Gradient Boosting Model

Gradient Boosting is a machine-learning method that uses an ensemble of decision trees in order to combine weaker models into a full system. The idea is that each of the weaker models potentially captures an aspect that the other weaker models do not and thereby closes gaps of the prediction power. Gradient boosting makes use of the gradient descent algorithm (see e.g. Zinkevich et al., 2010), in order to find an optimal solution for the ensemble. By using a loss function based on the valuation of a certain parameterization of the model, one can find the gradient of this loss function by computing the first derivative. Logically, the global minimum of the loss function is the optimal solution of the problem, i.e. the parameterization of the model, where the error is minimal. As long as the gradient points downwards, the gradient descent algorithm continues the search in the same direction. Once the gradient points upwards, the algorithm changes its direction, since the minimum must be in the past. The step size that the algorithm goes forwards or backwards is called learning rate. The local minimum is found, if the first derivative of the function equals 0. The algorithm guaranties to find a local minimum for an appropriate learning rate. There are certain techniques for also finding the global minimum of a

[3] The definition for precision is given in Section 6.2.

loss function with a higher probability and much faster, such as mini-batch gradient descent or Adagrad, described in Ruder (2016).

Since explaining gradient boosting in detail would be beyond the scope of this thesis, the interested reader is pointed to specialized literature going into the depth of gradient boosting, for example Friedman (1999a;b).

5.2.1 Features

All information available to the rule-based system is used as features. These features are either based only on an anaphor, only on an antecedent or on a pair of anaphor and antecedent. Table 5.4 shows the features, on what type of markable they are based and their value type.

5.2.2 Hyper-Parameters

Gradient boosting requires certain hyper-parameters, which have to be set manually. The respective hyper-parameters are:

1. Shrinkage or Learning Rate ν: Since gradient boosting makes use of gradient descent, the step size in finding the local minimum needs to be specified. A higher value means taking bigger steps on the graph of the cost function, while a smaller value follows the graph more smoothly. This also means that smaller values of shrinkage are more accurate, but also result in a longer training time. ν might range between 0 and 1 and is commonly set to 0.1 (Friedman, 1999b; p. 368).
2. Number of trees M: Number of the components, i.e. the weaker decision trees. M can also be seen as the number of iterations of the model. Friedman (1999a) suggests to set M to a higher value and decrease it while optimizing the shrinkage. He finds values between 100 and 500 to be efficient.
3. Tree Depth or Interaction Depth J: The depth of the tree determines, how many interacting features are used. A value of $J = 1$ will result in an additive model. Friedman (1999a) advises to determine J on a held-out dataset.

Feature	**Value**
For anaphor	
AnaDocFreq	Integer
AnaLength	Integer
AnaWordCount	Integer
AnaModByPP	Boolean
ArgumentTakingRatio	Decimal
AnaIsSubject	Boolean
isPolitics	Boolean
isProfRole	Boolean
isCountryPart	Boolean
isBuildingPart	Boolean
For antecedent	
AnteLength	Integer
AnteWordCount	Integer
isCountry	Boolean
AnteIsSubject	Boolean
AnteDocFreq	Integer
isORG	Boolean
For anaphor-antecedent pair	
SemanticConnectivity	Decimal
SentDist	Integer
CosineSimilarity	Decimal

Table 5.4: Features implemented for the gradient boosting model.

5.2.3 Relative Variable Importance

Methods like gradient boosting allow for interpreting the influence of features (or variables), by approximating the change in the performance of the system by leaving out a single feature (compare Friedman, 1999a; pp. 1213 ff.). For a single decision tree and a certain variable j, the (squared) relative variable importance $\hat{I}_j^2(T)$ is given by

$$\hat{I}_j^2(T) = \sum_{t=1}^{J-1} \hat{\imath}_t^2 1(v_t = j), \tag{5.3}$$

where T is a tree with depth J, t is a non-terminal node, v_t is the splitting variable at t and $\hat{\imath}_t^2$ is the observed improvement of the squared error at t, caused by splitting the variable. For evaluating an ensemble of trees $\{T_m\}_1^M$, the average relative variable importance $\hat{I}_j^2$ for a variable j is given by

$$\hat{I}_j^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{I}_j^2(T_m), \tag{5.4}$$

where M is the total number of trees.

The most important variable is set to 100 and the remaining variables receive a scaled, relative value according to their $\hat{I}^2$ value.

6 Experiments

In this chapter, three experiments are conducted in order to perform bridging resolution. The experimental setup as well as the results of the experiments are reported.

6.1 Data

The data used for the experiments are the information status annotations from the corpora DIRNDL (4.4) and GRAIN (4.5). The specific setup is described in detail for the respective experiment.

6.2 Evaluation Metrics

Results are reported using *precision* (P), *recall* (R) and *F1-Score* (F1), which are predominantly used in information retrieval (Manning et al., 2008), but are also applied in various machine learning evaluations in general. These measures are calculated using the different possible outcomes of both the annotated gold standard and the system's prediction for bridging anaphors and antecedents. These outcomes are given an overview in Table 6.1 and are called *true positive* (TP), *false positive* (FP), *true negative* (TN) and *false negative* (FN).

TP is the count of all instances, where the system predicts an item to be a bridging pair and it actually is one. Equivalently, TN is the count of those pairs, where both the system and the annotation say that the pair is not a bridging pair. FP, or Type I errors, occur when the pair is actually not a bridging pair, but the system predicts it to be one. Similarly, FN, or Type II errors, occur when the instance is an actual

Is bridging pair, according to:		
Gold	**System**	**Name**
True	True	True positive
False	True	False positive
False	False	True negative
True	False	False negative

Table 6.1: Overview of the four possible outcomes of gold-pred evaluation.

bridging pair, but the system predicts it to be none. Given that, precision, recall and F-Score can be defined as follows:

$$P = \frac{TP}{TP + FP} \tag{6.1}$$

$$R = \frac{TP}{TP + FN} \tag{6.2}$$

$$F = (1 + \beta^2)\frac{P \times R}{(\beta^2 \times P) + R} \tag{6.3}$$

$$F1 = 2\frac{P \times R}{P + R} \tag{6.4}$$

where β in Equation 6.3 is a positive, real value and describes the weighting of P and R. $\beta > 1$ means that R is weighted higher, while $\beta < 1$ assigns more importance to P. Equation 6.4 shows the simplified formula for $\beta = 1$, which is then also called F1-Score. F1 is used for all further evaluations, meaning that P and R are weighted equally. As shown in Equation 6.1, P can be understood as a measure of how well the system is able to correctly classify instances as bridging pairs, while R, as shown in Equation 6.2, is a measure of how many bridging pairs the system is able to return. F1 is the harmonic mean of these two measures. All values of precision, recall and F1 are given in percentage.

6.3 Experiment 1: Rule-based System

The first experiment involves the rule-based system, described in Section 5.1.

6.3.1 Experimental Setup

As possible bridging anaphor and antecedent candidates, all gold-annotated information status markables are extracted. For pre-processing, several markables are then excluded from being a potential bridging anaphor, if they adhere to certain conditions:

1. The markable is indefinite (always information status category *r-new*)
2. The markable is a pronoun (always *r-given* or *r-cataphor*)
3. The markable is a proper name (proper names are of information category *r-unused-unknown,*, *r-unused-known* or *r-given*)
4. The markable contains an embedded NP or PP (if bridging, these markables usually classify as *r-bridging-contained*)
5. Coreferent markables, except for the first occurrence in a coreference chain (are of category *r-given*, only the first mention of a coreference chain might be *r-bridging*)

For the last point, the coreference information is used as being annotated in the gold standard of DIRNDL and GRAIN.

For the antecedent recognition, all extracted markables are suitable to be a bridging antecedent and are therefore kept unaltered, when used as candidates for antecedents.

The hyper-parameters are trained on a combination of training and development set (which is henceforth called development set), since the rule-based system does not need training data other than for determining the hyper-parameters. For this purpose, the official train-development-test split of DIRNDL is adopted[1]. For GRAIN, no such split exists yet. Therefore, a 60-20-20 split is being proposed, visualized in Table 6.2. The concrete hyper-parameters, which are evaluated on a combination of the train and the development set for DIRNDL and GRAIN, are presented in Table 6.3. Some rules do not fire at all on the development set and their hyper-parameter

[1] Downloadable from `http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/dirndl.en.html`

is therefore given as *NaN*. If a rule does not make use of a certain hyper-parameter, this is indicated with a dash (–).

Set	**Documents**
Train	20140524, 20140719, 20141129, 20141206, 20150221, 20150228, 20150425, 20150613, 20150808, 20150815, 20150912, 20150919, 20151010
Development	20140614, 20140802, 20150124, 20150404, 20151024
Test	20140517, 20140927, 20141011, 20150110, 20150620

Table 6.2: Train-development-test split, proposed for GRAIN.

Pilot experiments showed that it does not make much sense to optimize the hyper-parameter *semConThreshold* on a held-out set, since the threshold is very vocabulary dependent and the system greatly overfits on the development set. Therefore, a fixed value of 15 is set for semConThreshold as a heuristic.

6.3.2 Baseline

In order to judge the difficulty of the task, an informed baseline is implemented, covering certain aspects of bridging anaphors and antecedents. The baseline receives the same input markables as the rule-based system, but it consists of only one rule. A markable is a bridging anaphor, if it is not modified by any PP, adjective or demonstrative pronoun. Then, the antecedent is chosen to be the subject of the previous sentence. The baseline reflects the common ground that bridging anaphors are usually short, unmodified NPs and that their antecedents usually appear in the previous sentence (cf. Hou, 2016). The results of the baseline on the test sets are shown in Table 6.4. The baseline achieves good results on anaphor recognition, suggesting that many bridging anaphors are indeed unmodified NPs. The high recall is expected since the baseline suggests many candidates to be an anaphor, independent of other properties of the candidate. Naturally, the system underperforms using such an approach, since not all unmodified NPs are necessarily bridging anaphors. The poor performance on the full prediction task is not surprising: even though the

Rule	maxSentDist	argTakingRatio-Threshold	semConThreshold
1	2	–	–
2	NaN	NaN	–
3	–	–	–
4	1	–	–
5	NaN	–	–
6	NaN	–	–
7	NaN	NaN	–
8	1	0.45	–
9	–	–	–
10	2	–	15
11	1	–	–
12	2	–	15
13	1	–	–

(a) DIRNDL

Rule	maxSentDist	argTakingRatio-Threshold	semConThreshold
1	1	–	–
2	NaN	NaN	–
3	–	–	–
4	NaN	–	–
5	NaN	–	–
6	NaN	–	–
7	NaN	NaN	–
8	2	0.45	–
9	–	–	–
10	0	–	15
11	NaN	–	–
12	0	–	15
13	1	–	–

(b) GRAIN

Table 6.3: Hyper-parameter setting on DIRNDL (a) and GRAIN (b).

antecedent often occurs in close proximity to its anaphor, it is not necessarily the subject in the previous sentence.

Data	P	R	F1
DIRNDL	12.6	65.1	21.1
GRAIN	15.8	69.8	25.9

(a) Anaphor Recognition

Data	P	R	F1
DIRNDL	0.5	2.3	0.8
GRAIN	0.4	1.6	0.6

(b) Bridging Resolution

Table 6.4: Results of the baseline on DIRNDL and GRAIN for anaphor recognition (a) and full bridging resolution (b).

6.3.3 Results

The results for the precision of single rules is shown in Table 6.5 for the development set of DIRNDL and GRAIN, both on the full prediction task, i.e. finding anaphor-antecedent pairs (bridging resolution), and additionally also on finding only the correct bridging anaphor (bridging anaphor recognition). If a rule did not fire at all, the precision value is given as NaN. Note, that P can only be maximally as high as PAna. If both values are the same, it means that for the predicted anaphors, all antecedents could be predicted correctly.

It becomes clear that many rules from Hou et al. (2014) do not fire at all. These rules are the numbers 3, 5, 6 and 7. Except for rule 7, these rules are all lexically based. But also the other lexically based rules do not retrieve much candidates, in particular rule 1, 2, 4, 9 and 11. The more general rules retrieve many more candidates, but intuitively, the precision goes down, since a lot more false positives are being retrieved as well. Rule 9 performs the best overall, for both DIRNDL and GRAIN. As a final observation, performance of the rules is always higher on DIRNDL, as compared to GRAIN.

Based on the precision for a rule on the full prediction task, the rule ordering for DIRNDL and GRAIN is given in Table 6.6, as used for the post-processing step of the system. Taking DIRNDL as an example, rule 9 and rule 8 both suggest an

Rule	Fire Rate	PAna	P
1	0	NaN	NaN
2	0	NaN	NaN
3	0	NaN	NaN
4	4	100.0	0.0
5	0	NaN	NaN
6	0	NaN	NaN
7	0	NaN	NaN
8	47	48.9	17.0
9	34	79.4	64.7
10	113	44.2	17.7
11	20	50.0	20.0
12	113	44.2	17.7
13	871	27.4	4.9

(a) DIRNDL

Rule	Fire Rate	PAna	P
1	6	16.6	16.6
2	1	0.0	0.0
3	0	NaN	NaN
4	2	0.0	0.0
5	0	NaN	NaN
6	0	NaN	NaN
7	0	NaN	NaN
8	26	38.5	7.7
9	32	46.9	40.6
10	37	18.9	8.1
11	14	7.1	0.0
12	34	17.6	8.8
13	719	20.6	2.5

(b) GRAIN

Table 6.5: Fire Rate and Precision for every rule on anaphor recognition (PAna) and the full prediction task (P) on DIRNDL (a) and GRAIN (b).

antecedent for the same anaphor candidate. However, only rule 9 will be evaluated for this pair, as its overall precision is higher. This can of course also mean, that rule 9 could predict a wrong antecedent, where rule 8 would have suggested the correct antecedent. The rule precision is used for the complete system evaluation on the test set. Note, that the rule precedence is almost identical for DIRNDL and GRAIN, meaning that the rules perform the same, relatively speaking.

Ordering	**Rule**
1	9
2	11
3	12
4	10
5	8
6	13
7	NaN
8	NaN
9	NaN
10	NaN
11	NaN
12	NaN
13	NaN

(a) DIRNDL

Ordering	**Rule**
1	9
2	1
3	12
4	10
5	8
6	13
7	NaN
8	NaN
9	NaN
10	NaN
11	NaN
12	NaN
13	NaN

(b) GRAIN

Table 6.6: The precedence of rules for the post-processing step, based on their precision for bridging resolution, on DIRNDL (a) and GRAIN (b).

Table 6.7 shows the performance of the full system for anaphor recognition and bridging resolution on the test set of DIRNDL and GRAIN.

The performance on the test set is rather poor, scoring an F1-Score of 5.3% on DIRNDL and 4.0% on GRAIN, respectively. Furthermore, the performance for anaphor recognition is even lower than the results of the baseline. On the other hand, the system is able to improve precision for anaphor recognition, by sacrificing recall.

Data	P	R	F1
DIRNDL	15.8	14.0	14.8
GRAIN	20.1	18.0	19.0

(a) Anaphor Recognition

Data	P	R	F1
DIRNDL	5.6	4.9	5.3
GRAIN	4.2	3.7	4.0

(b) Bridging Resolution

Table 6.7: Results of the rule-based system on DIRNDL and GRAIN for anaphor recognition (a) and full bridging resolution (b).

Evaluation on the development set The results on the test set are somehow unsatisfactory. In order to evaluate, how much of the results is explainable due to a skewed distribution of the development set and the test set, the results for evaluating on the development set are additionally reported, in Table 6.8. As can be seen, the results are much higher than for the evaluation on the test set. This suggests that the rule-based system is generally able to capture a good amount of bridging relations when optimized, but is not able to adapt to new data. This is especially true for DIRNDL, as performance increases drastically from an F1-Score of 5.3% to a score of 14.8%. Possible reasons for this result are discussed in Chapter 7.

Data	P	R	F1
DIRNDL	33.6	84.5	48.1
GRAIN	21.6	89.1	34.8

(a) Anaphor Recognition

Data	P	R	F1
DIRNDL	10.4	25.8	14.8
GRAIN	4.6	19.0	7.4

(b) Bridging Resolution

Table 6.8: Results of the rule-based system for DIRNDL and GRAIN for anaphor recognition (a) and full bridging resolution (b) on the development set.

Omission of gold coreference information Rösiger (2018, in preparation) implemented a previous version of the rule-based system, optimized on DIRNDL. She also reports evaluation values for using no coreference information, increasing the size of markables, the system has to choose from. Since coreferent markables cannot be bridging by definition and the search space for the system increases, the performance decreases, as anticipated, displayed in Table 6.9.

Method	**P**	**R**	**F1**
No Coref	10.7	11.6	11.1
Gold Coref	14.9	11.6	13.0

Table 6.9: Results from Rösiger (2018, in preparation), comparing results on DIRNDL with and without coreference information.

6.4 Experiment 2: Oracle Lists

From Section 6.3, it became clear that for both DIRNDL and GRAIN, finding the correct antecedent is much harder than finding bridging anaphors. In order to investigate if the rules are generally able to capture the antecedents, or if most antecedents lie outside the scope of the rules, oracle lists are implemented.

An oracle list is a new output of each rule, and instead of suggesting a single anaphor-antecedent pair for a markable, a rule now outputs an anaphor and a list of suitable antecedents for this anaphor. The list is ordered, with the best fitting antecedent on top, followed by other possible antecedents in decreasing order of probability. The system can then be evaluated given a certain length of oracle lists. An evaluation given the oracle list of 5 for example means that the rule output is counted as a success (i.e. a true positive), if the correct antecedent is within the five top suggested antecedents in the oracle list. The evaluation is therefore not an evaluation of the systems actual performance, but of the system's potential and can be seen as an outlook on what the system would be able to predict if the actual antecedent was ranked highest, hence the name *oracle* list. Because the rules do not output probabilities, for each rule, an individual way of implementing has to be found.

Table 6.10 gives an overview of how each rule implements its ranking. For all rules, if the antecedent was already predicted to be the antecedent for another anaphor, it is put on top of the ranking. Depending on the rule, the ranking can either be ascending or descending.

6.4.1 Experimental Setup

The experimental setup is the same as for the rule-based system without oracle lists in 6.3.1. As before, the rules are evaluated on the development set (being again a combination of the training and the development set) and the whole system's performance on the test set, using the rule precedence, which was determined on the development set.

6.4.2 Results

Figure 6.1 shows the precision for the individual rules on DIRNDL and GRAIN, evaluating differing length of oracle lists.

It becomes obvious that the rules in general can benefit from the oracle list evaluation, meaning that the rules have a general scope over the correct antecedent, but sometimes favor a wrong candidate. Rule 8 in DIRNDL is just an example, increasing the precision from around 17% when the oracle list has length 1 (equivalent to not using oracle lists at all), up to a precision of roughly 40% for an oracle list length of 4. Also, rule 8 in GRAIN and rule 4 in DIRNDL benefit a lot. Apart from these positive results, Figure 6.1 also shows that the rules do not benefit to the same amount and some rules that do not benefit at all from using oracle lists. However, a lack in increase of precision is not necessarily bad: looking at rule 9 in DIRNDL, the rule does not benefit from an increasing oracle list length. This is actually positive, since the precision is, with over 60%, already quite high, meaning that rule 9 is able to capture the correct antecedents already when using the regular system without oracle lists. This is different for rule 2 and 4 in GRAIN: Also this rule does not change its precision, but it stays at a constant value of 0%. This means that none of the correct antecedents of rule 2 and 4 are in their scope, no matter

Rule	Ranking Method
1	rank according to semantic connectivity, markable with highest value on top
2	rank according to proximity, closest relative person NP or pronoun on top
3	rank according to document frequency, markable with highest document frequency on top
4	Same as rule 3, but only consider organizations
5	rank according to proximity, closest percentage expression on top
6	not implemented
7	rank according to proximity of modified head, closest markable on top
8	rank according to semantic connectivity, markable with highest value on top
9	rank according to proximity, closest country NP on top
10	rank according to semantic connectivity, markable with highest value on top
11	rank according to semantic connectivity, markable with highest value on top
12	rank according to semantic connectivity, markable with highest value on top
13	rank according to cosine similarity, markable with highest value on top

Table 6.10: Ranking method of an oracle list for each rule.

how long the oracle list gets. None of the rules reach a precision identical to its anaphor recognition precision, which suggests that for no rule all antecedents are in its scope, emphasizing that bridging resolution is rather complex on DIRNDL and GRAIN. Taking all rules together, the system does not benefit much from oracle lists, meaning that the power of the rule-based system is very limited in terms of finding the correct antecedent.

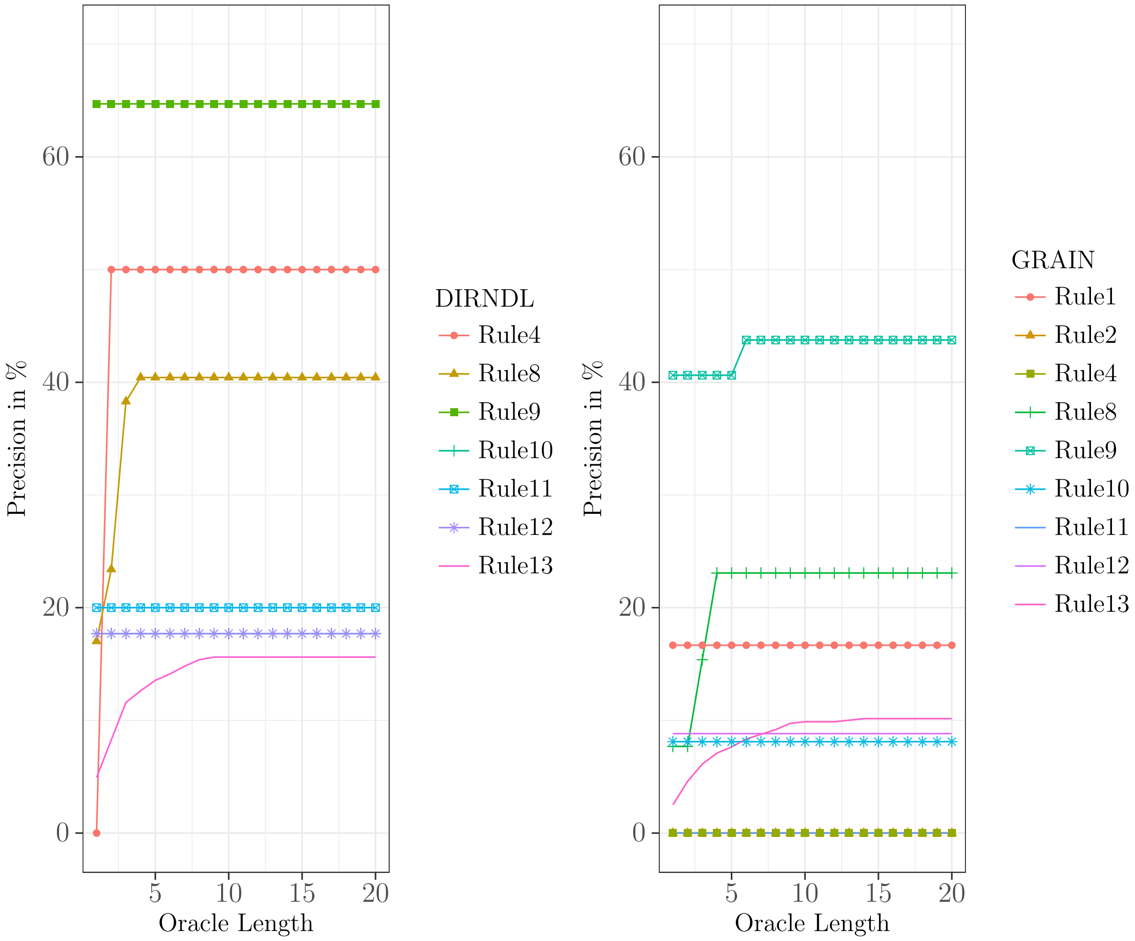

Figure 6.1: Performance of rules on DIRNDL and GRAIN, using different lengths of oracle lists.

In Figure 6.2, the full performance of the system on DIRNDL and GRAIN with respect to different oracle lengths is displayed.

It can be seen that the usage of oracle lists is beneficial for bridging resolution, especially for smaller oracle list lengths. After the length of 7 for DIRNDL and 19 for GRAIN, no further correct antecedent can be found in the oracle lists[2]. This suggests that the problem of finding the correct antecedent is not only that the rules are not able to predict the correct antecedent, but also that certain antecedents are not even in the scope of a specific rule. Recall that for bridging resolution, precision and recall can maximally be as high as the values for anaphor recognition; if the values would be actually the same, the system would have correctly classified all antecedents for all found anaphors. Even with the full use of oracle lists, the system is still far away from the maximum values, i.e. the anaphor recognition values, which is 14.8% F1 for DIRNDL and 19.0% F1 for GRAIN, compared to 9.0% on DIRNDL and 10.0% on GRAIN for the oracle list evaluation on the full task. Reasons for this discrepancy are explored in Chapter 7.

6.5 Experiment 3: Gradient Boosting Model

Lastly, the gradient boosting model from Section 5.2 is evaluated.

6.5.1 Experimental Setup

For implementing the gradient boosting model, the *gbm* package[3] of the statistical computing environment R[4] was used. The features, as being described in 5.2, are extracted from DIRNDL and GRAIN. The train-development-test split is as described in 6.3.1, but the train and the development set are not merged this time. The hyper-parameters (see 5.2) are trained on the development set and are listed in Table 6.11 for DIRNDL and GRAIN. The model parameters are trained on the train set and the system's performance evaluated on the test set.

[2] This was verified for an oracle list length up to 100.

[3] `https://cran.r-project.org/web/packages/gbm/gbm.pdf`

[4] `https://www.r-project.org/`

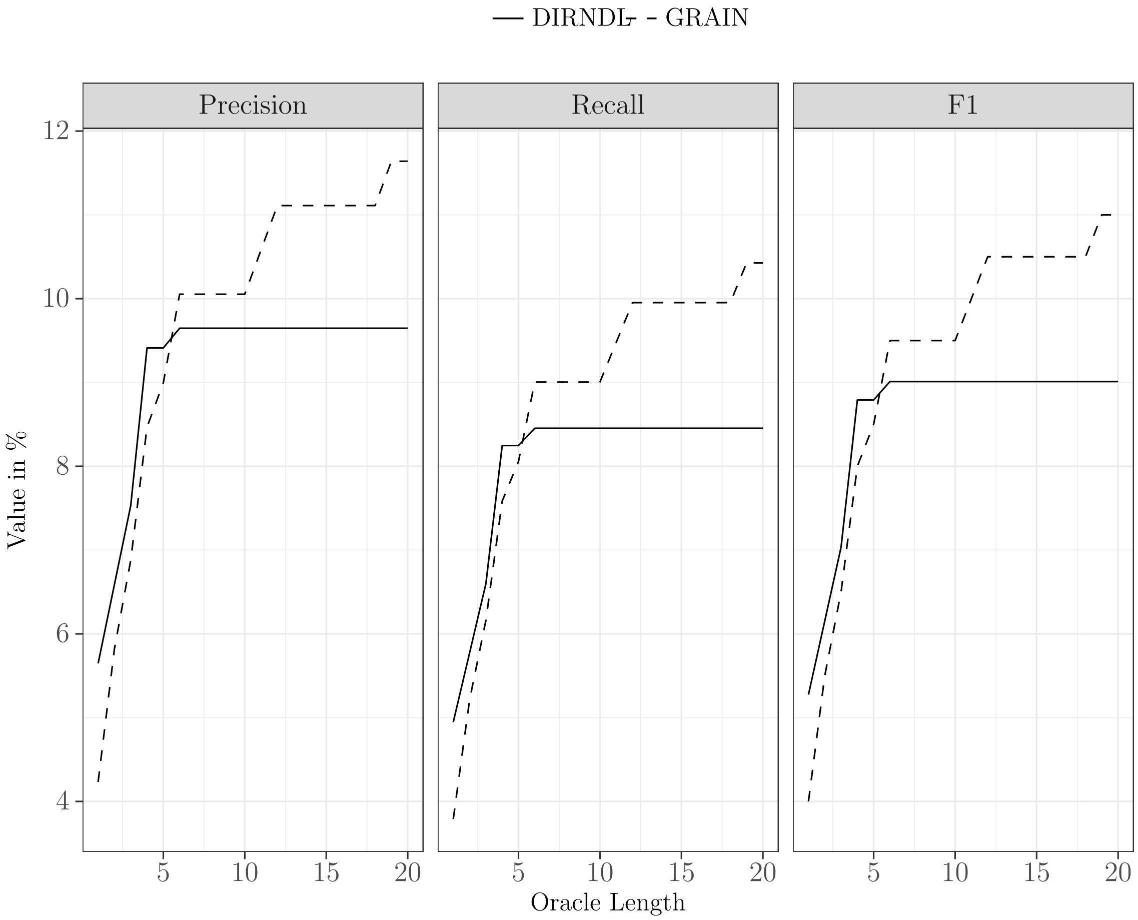

Figure 6.2: Performance of the rule-based system on DIRNDL and GRAIN, using different lengths of oracle lists.

Hyper-Parameter	**Value**	
	DIRNDL	**GRAIN**
Shrinkage ν	0.005	0.001
#Trees M	6000	250
Tree Depth J	15	2

Table 6.11: Hyper-parameters for the gradient boosting model for DIRNDL and GRAIN.

Cross validation 10-fold cross validation is applied for training. k-fold cross validation (see also Kohavi, 1995) is a technique, where the dataset is divided into k equal subsets. Iteratively, each of the subsets is taken as the test set and the remaining subsets as training data, until all subsets have been the test set exactly once. For evaluation, the mean of the results is computed and set as the overall result. This way, it is less likely that outlier documents influence the result, in case they were set as the test set.

Sampling Unbalanced datasets are a problem in machine learning. For example, in binary classification, the negative class might occur much more frequently than the desired class. This is also the case for DIRNDL and GRAIN, where the amount of markables not being bridging greatly exceeds the number of markables being bridging. A system then tends to over-fitting, meaning that it will most likely always assign all instances to the negative class.

Therefore, sampling is applied in order to adjust the training samples. In general, two methods are available: up and down sampling. While up sampling increases the number of positive instances artificially, down sampling removes negative instances until a reasonable ratio is reached. The sampling technique *SMOTE* (Synthetic Minority Over-sampling Technique, Chawla et al., 2002), which combines both methods, is chosen over other sampling techniques, since it yielded the best results.

6.5.2 Results

Table 6.12 shows the results of the gradient boosting system on DIRNDL and GRAIN.

For anaphor recognition, the gradient boosting system is able to capture many instances, leading to a good F1-Score of 29.0% for DIRNDL and 39.6% for GRAIN. Also, for the full prediction task, the gradient boosting system is able to outperform the rule-based system on DIRNDL, scoring an F1-Score of 11.3%. For GRAIN, the system is not able to generalize on the test set, predicting no bridging-antecedent pair correctly. This is probably due to a small amount of positive test subjects for

Data	P	R	F1
DIRNDL	25.3	33.6	29.0
GRAIN	35.1	45.5	39.6

(a) Anaphor Recognition

Data	P	R	F1
DIRNDL	7.0	28.1	11.3
GRAIN	0.0	0.0	NaN

(b) Bridging Resolution

Table 6.12: Results of the gradient boosting model on DIRNDL and GRAIN for anaphor recognition (a) and full bridging resolution (b).

the test set on GRAIN (90 bridging pairs). When evaluated on the training set, the system was able to score an F1-Score of 9.0% for GRAIN on the full task, showing that the model is at least able to find some generalizations on the data it was trained on[5].

Table 6.13 displays the results of calculating the relative variable importance as described in Section 5.2.3 for the features from Section 5.2.1.

As expected, the features for GRAIN were not deemed very useful by the model, since it was not able to gather any generalizations. The most useful features, according to the model, were still the sentence distance between anaphor and antecedent, the semantic connectivity and information about countries, which already proved quite useful for the rule-based system. For DIRNDL, the analysis offers more interesting and diverse results. The most valuable feature is the information, if a potential anaphor is a professional role. The next top-five ranked features are AnaLength, SemanticConnectivity, CosineSimilarity, ArgumentTakingRatio and SentDist. SemanticConnectivity and SentDist were also highest ranked on GRAIN. All these features seem plausible, since they are not lexically dependent and are based on non-discrete scales, giving the stochastic model the ability to make full use of them.

Figure 6.3 shows the most strongest features for bridging anaphor prediction on GRAIN: AnaArgumentTakingRatio and AnaLength. It becomes clear, why they are such strong predictors, since they are able to cluster instances of the positive class in a range of the anaphor length being smaller than 25 and the argument-taking ratio

[5] For comparison: When the system is evaluated for DIRNDL on the training set, it scores an F1-Score of 54.3% for the full task

Feature	VI
isProfRole	100.0000
AnaLength	60.7928
SemanticConnectivity	55.3577
CosineSimilarity	53.0699
ArgumentTakingRatio	42.9556
SentDist	39.6681
AnteLength	31.7330
AnteIsSubject	27.8216
isCountry	23.0130
AnaWordCount	16.4514
AnaIsSubject	15.3698
AnteWordCount	11.2507
isPolitics	10.3412
isCountryPart	2.6284
isORG	2.5526
AnteDocFreq	0.7929
AnaDocFreq	0.0002
AnaModByPP	0.0000
isBuildingPart	0.000

(a) DIRNDL

Feature	VI
SentDist	100.000
SemanticConnectivity	19.215
isCountrytrue	10.140
isCountryParttrue	3.145
AnaModByPPtrue	0.000
CosineSimilarity	0.000
AnteLength	0.000
isPoliticstrue	0.000
AnteWordCount	0.000
AnteIsSubjecttrue	0.000
isProfRoletrue	0.000
AnaDocFreq	0.000
ArgumentTakingRatio	0.000
isORGtrue	0.000
AnaIsSubjecttrue	0.000
AnteDocFreq	0.000
isBuildingParttrue	0.000
AnaWordCount	0.000
AnaLength	0.000

(b) GRAIN

Table 6.13: Relative variable importance, estimated with the gradient boosting classifier on DIRNDL (a) and GRAIN (b).

being between 0.25 and 0.6. On the other hand, these two features are clearly not sufficient predictors, since a lot of negative samples are still in this range too, and a considerable amount of instances of the positive and the negative class do overlap.

Figure 6.4 shows two features used for the full prediction task on DIRNDL and dependent on both the anaphor and the antecedent: SemanticConnectivity and the SentDist. Not surprisingly, all pairs have a sentence distance smaller than 5. The semantic connectivity range is harder to interpret. The main range for being a bridging pair seems to be between 0 and 20. Intuitively, one would assume the semantic connectivity to be higher for bridging pairs, since it is supposed to capture the quintessence of being a bridging pair: occurring together in a N PREP N pattern and therefore being in a prototypical relationship. Therefore, it can be assumed that easier the calculation of the semantic connectivity is suboptimal or that many bridging pairs simply do not fall under a prototypicality assumption (compare again Table 4.9 in Section 4.5.3).

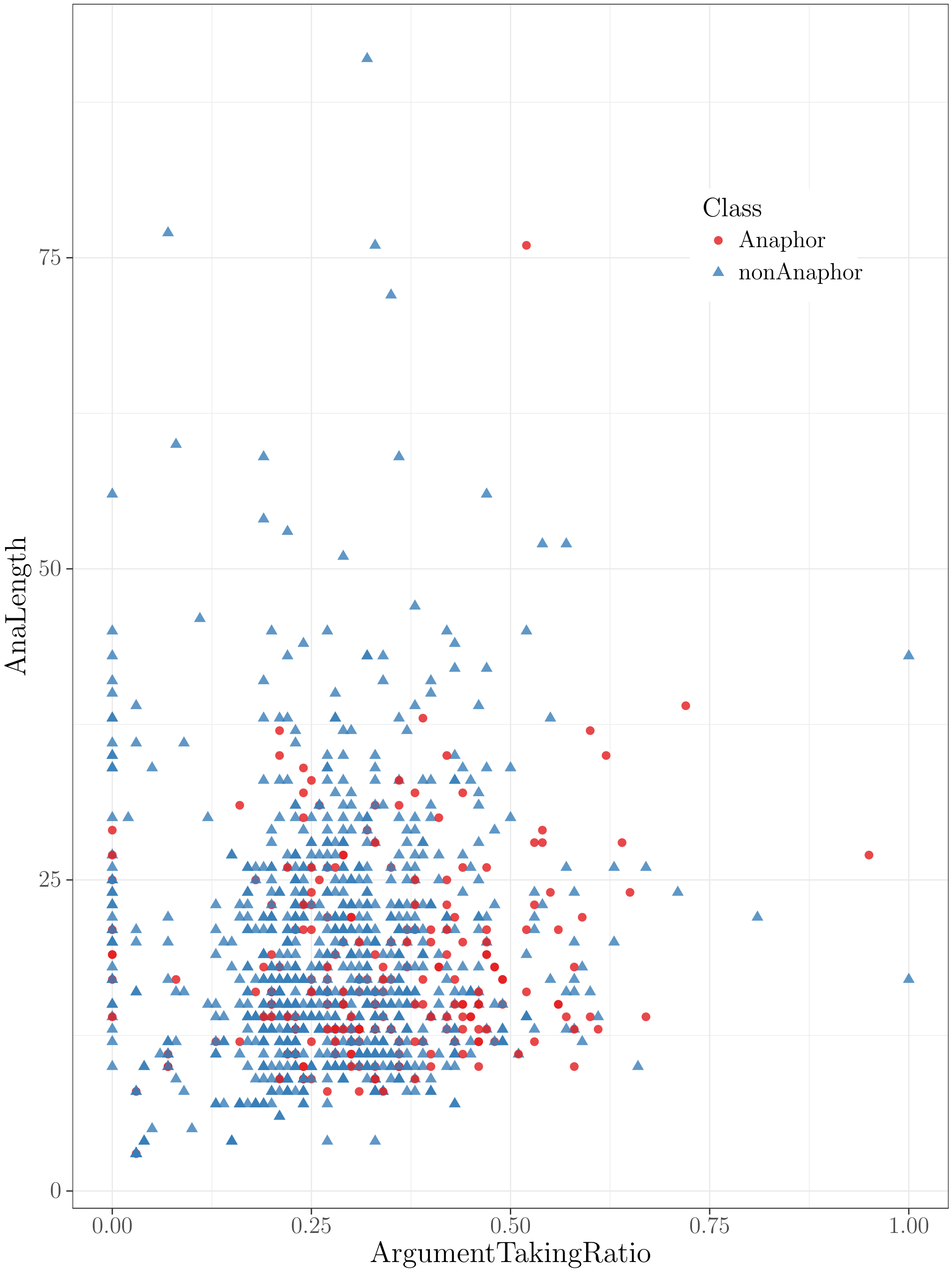

Figure 6.3: Argument-taking ratio and anaphor length as predictors for the class *bridging anaphor* in GRAIN.

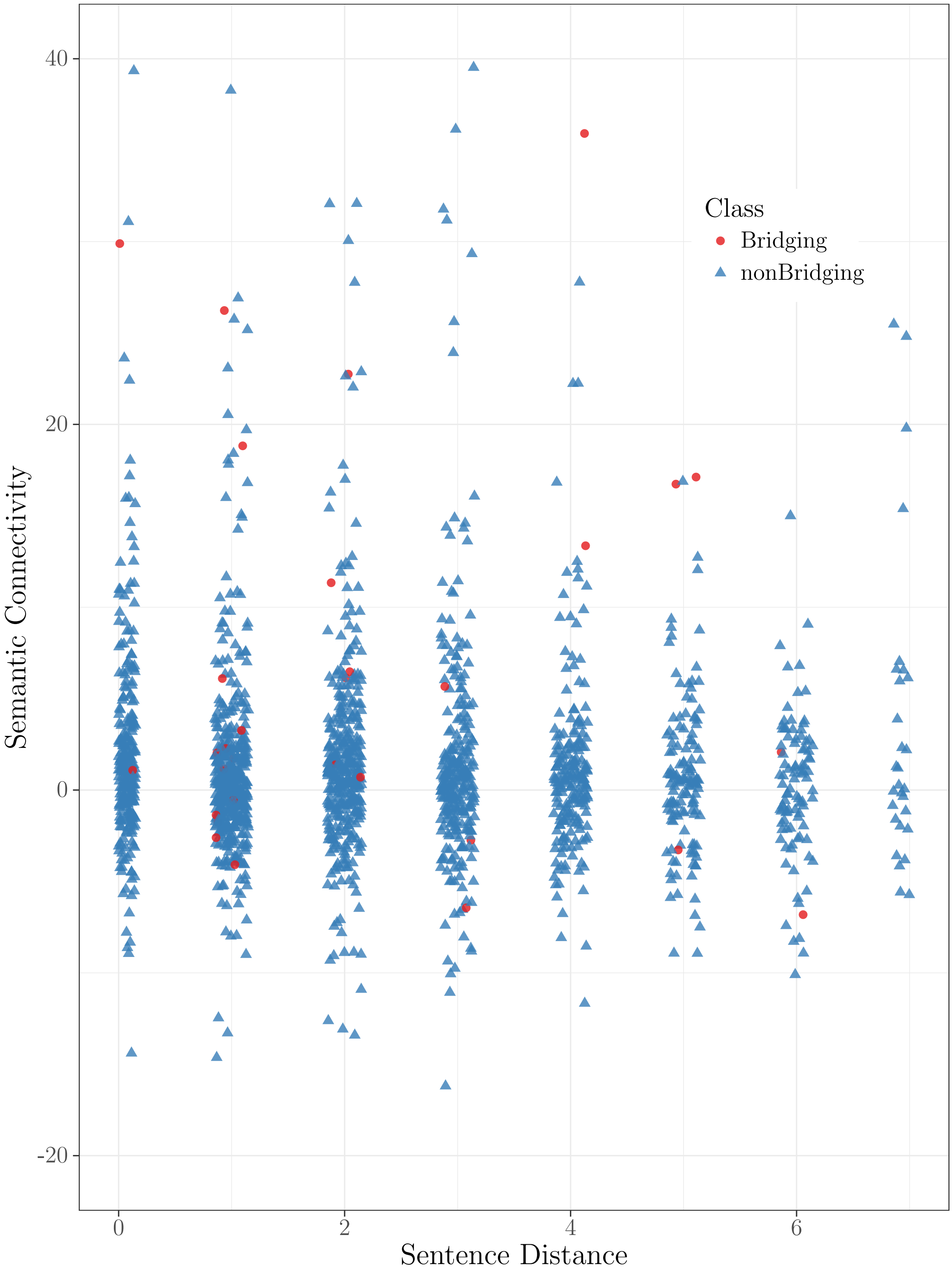

Figure 6.4: Semantic connectivity and sentence distance as predictors for the class *bridging pair* in DIRNDL. The data points have been "jittered" for visibility reasons, i.e. a small, random noise has been added alongside the x-axis.

7 Discussion

Some minor discussion of the results was already provided in Chapter 6, whenever it seemed convenient. This chapter aims to give a more holistic discussion, interrelating the findings of the thesis.

Interpreting the results of the first experiment, one has to conclude that building a rule-based system for German bridging resolution was only partially successful. Many rules from Hou et al. (2014) do not fire, meaning that they might be too domain-specific or that DIRNDL and GRAIN are specific domains that need special treatment. Indeed, six out of eight rules in Hou et al. (2014) are lexical-based. The other, more general rules are not able to cover all the variety of bridging types that are present in DIRNDL or GRAIN. Especially when evaluating on the test set, the results are rather disillusioning. Evaluation on the development set has shown that the system is generally capable of finding bridging pairs, but is not able to transfer its knowledge onto new data. This might be because topic shifts in the test set of DIRNDL and GRAIN is generally a corpus of very diverge topics. The new rules which were added to the system, going over the suggestions of Hou et al. (2014), were able to account for some of the bridging instances, but could not present a high precision. Only rule 9 by Rösiger (2018, in preparation) was really able to score high precision values. Unfortunately, this rule is highly domain-dependent and might not function on other corpora. Therefore, concentrating on more general, less domain and lexical based rules seems to be the right direction, but more research in order to improve and add to these rules needs to be conducted. The difference in the performance on DIRNDL and GRAIN can be explained by the difference between the corpora themselves: DIRNDL covers short broadcast news, for which it might be easier in general to retrieve bridging pairs, since the search space is limited. GRAIN on the other hand consists of longer, 10 minute interviews and the search space for potential bridging antecedents can be potentially very large. Another reason for the consistently poorer performance on GRAIN might be a lack in consistent bridging annotations, since the analysis in Section 4.5.2 has shown that agreement among the annotators was only limited.

Another difference in performance is observable between anaphor recognition and bridging resolution. As shown by all experiments, anaphor recognition is always more straight-forward and the model gives higher scores to it than to bridging resolution. This is not surprising, since bridging resolution can only be performed after anaphor recognition has already been done. In order to investigate, how difficult bridging resolution really is for the rule-based system, the oracle list based experiment 2 was conducted. It can be shown that re-ranking the antecedent candidates might let the system benefit. However, the power of re-ranking is limited, since the system is still not able to find all antecedents, independent of the length of the oracle lists. This is quite alarming and calls for new rules for the system or improved computation of the features it uses. In any way, it was once again confirmed, that bridging resolution is a very challenging task and by no means straight-forward.

Experiment 3 on gradient boosting has shown that it is possible to successfully use learning-based methods for bridging resolution. Especially for DIRNDL, the gradient boosting was able to outperform the rule-based system on the test set. For GRAIN, the situation is different, revealing the limitations of the gradient boosting system, which is insufficient amount of training data. The variable importance analysis has shown that the more general features are most useful for the system, emphasizing again the need to shift from too lexically focused approaches of bridging resolution.

One problem of the rule-based system seems to be that it only covers certain types of bridging. By using GermaNet's hyponymy information, it was attempted to cover as many bridging related content words as possible, in order to be as domain-independent as possible. Since the rules, which use the information of GermaNet, still do not fire or only retrieve a very limited amount of candidates, this was not successful. Indeed, it might be beneficial to diverge from having a view on bridging that focuses too much on the lexical aspect of bridging resolution. As seen in Chapter 3, almost exclusively, bridging resolution research has focused on resolving lexical bridging relations. As seen from GRAIN, this type of bridging might only constitute a small amount of the bridging relations that can be found in a corpus.

Therefore, this thesis aimed to investigate, to what extent other types of bridging can be discovered. It was shown that especially the prototypical relations can be resolved to a certain extent. The rules 8, 10, 12 and 13 make use of features that are more general than lexical relations and concentrate on anaphor-antecedent combinations that prototypically occur together. These feature are in particular the argument-taking ratio, the semantic connectivity and the cosine similarity of the pairs. The rule-based system is able to make use of these features to a certain extent, but it is not able to reliably find bridging-pairs and returns many false positives. This might be due to the fact that the features were trained on SdeWaC and are not transferable to all domains and that the features not solely cover what it means to be a bridging relation. Many nouns might occur in the proposed patterns of the Sections 5.1.1 and 5.1.2, but in a concrete context, they might just not be used in a bridged way. This observation leads to a main point that many attempts in bridging resolution are unable to solve: handling the context-dependent nature of bridging. That this is a major problem became obvious in many findings of the thesis: low recall for finding bridging pairs, often picking the wrong pairs and rules not having any scope over the correct antecedent. Especially this last point emphasizes that hand-crafted rules would eventually fail in finding context-dependent relationships, since it is not predictable a priori, what kind of relationships will occur in a text.

Learning-based systems seem to be an obvious answer to this problem, since they are able to learn from given context and can work successfully on many different domains, provided that they are given sufficient and domain-crossing training data. Unfortunately, this is the crux of bridging resolution, as was laid out by this thesis: bridging resolution systems lacks sufficient amount of training data. Consequently, the gradient boosting system was able to show some promising generalizations for DIRNDL, but was not able to find these generalizations for GRAIN. GRAIN is a corpus of very diverse topics, since every interview covers another interviewee and another topic of conversation, and even inside a specific interview, the topic is often varied. DIRNDL on the other hand is more heterogeneous, as it covers broadcast news, which often present very similar and repetitive topics. The paramount claim of this thesis is therefore emphasizing the need to increase efforts in gathering more data on bridging resolution, in order to successfully perform bridging resolution in the future. Some thoughts building up on that are presented in Chapter 8.

8 Future Work

This chapter suggests different fields of possible future work to be considered.

Improve connectivity measures The variable importance analysis in Section 6.5.2 has shown that the argument-taking ratio as well as the semantic connectivity have a great information value for both DIRNDL and GRAIN. One problem that occurred during testing the systems was that the head of the phrase could not always be correctly classified. This naturally has a huge impact on the computation of the correct connectivity score and will further impact the detection of the correct antecedent for an anaphor. Using a dependency-parser and retrieving the head information from this layer might improve the system's performance. Currently, the heads are retrieved only on a rule-based basis. Furthermore, one could also develop experiments on finding more suitable patterns when searching for potential anaphor-antecedent co-occurrences.

Implement rules for world-knowledge and context dependent bridging The features used in this thesis focus mainly on the type of bridging that was called *prototypical* in Section 4.5.3. Prototypical bridging relations are more straightforward to approach, since they are not text-dependent, but lexically defined, and hence it is also more straightforward to engineer features, such as the semantic connectivity. Unfortunately, as also seen from Section 4.5.3, bridging is a rather diverse phenomenon and the prototypical bridging constitutes only roughly a third of the types in GRAIN. This calls for a more general treatment of bridging, that not only sees bridging as a lexical phenomenon, but also as context-dependent. It therefore could be helpful to perform a full-fledged information status classification first in order to detect bridging anaphors and afterwards find ways to detect suitable antecedents in a more general way. As for now, finding an antecedent in the rule-based system very much depends on the type of bridging anaphor, while this might be actually to limited. Also world-knowledge is not treated in this approach and needs appropriate resources to cover it.

Implement rules for bridging anaphors with abstract antecedents Other types of bridging not covered in the presented approach are bridging anaphors with abstract antecedents, i.e. antecedents that are verb phrases, clauses or sentences. While this phrases are technically part of the potential search space in both rule-based and learning-based systems, these cases might need special treatment.

Use re-ranker for the rule-based system Experiment 2 on the oracle list evaluation has shown that it might be beneficial to rank potential antecedents and re-rank promising candidates. A global and probabilistic re-ranker with access to more complicated features might be able to push the correct antecedent on the top of the ranking, in order to give the rule-based system a last performance boost.

Use information structure for improved salience modeling Intuitively, salience should play a major role in detecting the correct antecedent for an anaphor. Yet, the modeling of salience in the presented approach is only very limited. GRAIN currently obtains detailed annotation for information structure, covering information about focus and background. Additional studies could be carried out, investigating, if bridging antecedents are more likely to appear in the focus of a discourse. If there exists a correlation, information structure might be a new, powerful feature for inputting salience information into the systems.

Find better features Seemingly trivial, machine learning approaches rely heavily on engineering suiting and informative features for a problem. The learning-based results showed that the model is not able to use the given features in order to successfully predict bridging pairs on a large scale. New features not yet thought of might be needed to capture properties of the admittedly complex bridging relations, especially features of semantic nature. Furthermore, also certain syntactic properties might be common for bridging anaphors and their antecedents, yet the exact nature of the underlying syntactic properties is by no means clear at this point.

Use deep learning A lot of problems in bridging resolution stem from a lack of available data. Also Hou et al. (2014) made this observation. The kind of problems that arise are many-fold. Bridging is a discourse-dependent and inference-dependent phenomenon. With limited amount of training data, a model is not able to draw the appropriate generalizations. If one assumes that bridging obtains its meaning in context, than a model can only learn from enough amount of such context. Deep learning approaches have shown to handle logical inference-based problems quite well (see Bowman et al., 2014). It is likely that neural networks will also be able to capture the more subtle inferences that are drawn in bridging. Although, as long as the amount of training data is as it is currently, deep learning-based experiments on bridging resolution are far from being feasible.

More research on bridging and inference as a linguistic phenomenon As seen from Chapter 2, even though sharing a common ground, the specific theories of bridging in the research community can differ a lot. One part of the problem seems to be that bridging is such a diverse phenomenon that it seems somewhat arbitrary what to consider as bridging and what not. Also research about properties inherent to bridging anaphors and antecedents would let research on bridging resolution benefit a lot. Further linguistically motivated endeavors towards an understanding of bridging seem inevitable.

9 Conclusion

Bridging resolution proved to be a challenging task, requiring many different types of bridging relations to be resolved. This stems from several factors. Annotation of bridging is a challenging task to begin with, as can be seen from previous studies as well as an inter-annotator agreement study, carried out in this thesis for GRAIN. Furthermore, research on bridging is not consistent and it seems already difficult to decide, what all constitutes as bridging in the first place. Results of an extended rule-based system, which is based on Hou et al. (2014), therefore only showed limited success in resolving bridging relations. An oracle analysis showed, that the rule-based system often has no scope over the correct antecedent, calling for an improvement of the used rules and implementation of new rules. Promising results were obtained using a gradient-boosting system, showing that the future of bridging resolution might lie in using powerful methods, such as neural networks, given that a sufficient amount of training data is present. This is currently not the case.

Concluding the thesis, the questions of Chapter 1 are looked upon again and concisely answered.

On question 1 What kind of challenges does a bridging resolution system face?

The systems face a lack of training data and features for consistently covering a large variety of bridging phenomena.

On question 2 Are there special requirements for a bridging resolution system when dealing with German and non-standard data?

This seems to be the case, as using the systems on two different type of non-standard corpora only achieved limited results.

On question 3 Building on Hou et al. (2014) – can learning-based systems be successfully applied to bridging resolution?

This could be confirmed for DIRNDL, suggesting that more amount of training data might additionally push the use of learning-based methods forward for bridging resolution.

Bibliography

Baroni, Marco and Adam Kilgarriff. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2006, pages 87–90. 2006.

Baumann, Stefan and Arndt Riester. Referential and lexical givenness: Semantic, prosodic and cognitive aspects. In Elordieta, Gorka and Pilar Prieto, editors, *Prosody and Meaning*, volume 25 of *Interface of Explorations*, pages 119–162. Mouton de Gruyter, Berlin, 2012.

Björkelund, Anders, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer. 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, 2014.

Bowman, Samuel R., Christopher Potts, and Christopher D. Manning. Recursive neural networks for learning logical semantics. *CoRR*, abs/1406.1827, 2014. URL `http://arxiv.org/abs/1406.1827`.

Cahill, Aoife and Arndt Riester. 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*, SIGDIAL 2012, pages 232–236, 2012.

Cap, Fabienne. *Morphological processing of compounds for statistical machine translation.* PhD thesis, University of Stuttgart, 2014.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer.

SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Clark, Herbert H. Bridging. In *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, TINLAP 1975, pages 169–174, Stroudsburg, PA, USA, 1975. Association for Computational Linguistics.

Cohen, Jacob. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Dhar, Prajit. Modeling the influence of type and thematic fit in logical metonymy. Master's thesis, University of Stuttgart, 2018.

Dunning, Ted. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, March 1993. URL `http://dl.acm.org/citation.cfm?id=972450.972454`.

Eckart, Kerstin and Markus Gärtner. Creating silver standard annotations for a corpus of non-standard data. In Dipper, Stefanie, Friedrich Neubarth, and Heike Zinsmeister, editors, *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, volume 16 of *BLA: Bochumer Linguistische Arbeitsberichte*, pages 90–96, Bochum, Germany, 2016. URL `https://www.linguistics.rub.de/konvens16/pub/12_konvensproc.pdf`.

Eckart, Kerstin, Arndt Riester, and Katrin Schweitzer. A discourse information radio news database for linguistic analysis. In Chiarcos, Christian, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 65–75. Springer, Heidelberg, 2012.

Eckert, Miriam and Michael Strube. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89, 2000.

Faaß, Gertrud and Kerstin Eckart. SdeWaC - A corpus of parsable sentences from the web. In *Proceedings of the Language Processing and Knowledge in the Web - 25th International Conference*, GSCL 2013, pages 61–68, Darmstadt, Germany, 2013. URL `https://doi.org/10.1007/978-3-642-40722-2_6`.

Fellbaum, Christiane, editor. *WordNet: An Electronic Lexical Database.* MIT, Cambridge, MA, 1998.

Fleiss, Joseph L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Friedman, Jerome H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 1999a.

Friedman, Jerome H. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 1999b.

Grice, Herbert P. Logic and conversation. In Cole, Peter and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York, 1975.

Grishina, Yulia. Experiments on bridging across languages and genres. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, NAACL 2016, pages 7–15, 2016.

Hamp, Birgit and Helmut Feldweg. GermaNet - a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, 1997.

Hawkins, John A. *Definiteness and Indefiniteness.* Croom Helm, London, 1978.

Henrich, Verena and Erhard Hinrichs. GernEdiT – The GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, LREC 2010, pages 2228–2235, 2010.

Henrich, Verena and Erhard Hinrichs. Determining immediate constituents of compounds in GermaNet. In *Proceedings of Recent Advances in Natural Language Processing*, RANLP 2011, pages 420–426, 2011.

Hou, Yufang. *Unrestricted Bridging Resolution.* PhD thesis, Heidelberg University, 2016.

Hou, Yufang, Katja Markert, and Michael Strube. 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*, EMNLP 2013, pages 814–820, 2013a.

Hou, Yufang, Katja Markert, and Michael Strube. 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*, NAACL 2013, pages 907–917, 2013b.

Hou, Yufang, Katja Markert, and Michael Strube. 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*, EMNLP 2014, pages 2082–2093, 2014.

Kohavi, Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, volume 2 of *IJCAI 1995*, pages 1137–1143, 1995. URL `http://dl.acm.org/citation.cfm?id=1643031.1643047`.

Laparra, Egoitz and German Rigau. ImpAr: A deterministic algorithm for implicit semantic role labelling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL 2013, pages 1180–1189, Sofia, Bulgaria, 2013.

Löbner, Sebastian. Definites. *Journal of Semantics*, 4:279–326, 1985.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, New York, USA, 2008.

Markert, Katja, Malvina Nissim, and Natalia N. Modjeska. Using the web for nominal anaphora resolution. In *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, EACL 2003, pages 39–46, 2003.

Markert, Katja, Yufang Hou, and Michael Strube. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the*

Association for Computational Linguistics: Long Papers - Volume 1, ACL 2012, pages 795–804, 2012.

Miller, George A. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Nissim, Malvina. Learning information status of discourse entities. In *Proceedings of the 2006 Conference on Emprical Methods in Natural Language Processing*, EMNLP 2006, pages 94–102, 2006.

Nissim, Malvina, Shipra Dingare, Jean Carletta, and Mark Steedman. An annotation scheme for information status in dialogue. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC 2004, pages 1023–1026, 2004.

Padó, Sebastian and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

Poesio, Massimo and Renata Vieira. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, 1997.

Poesio, Massimo, Renata Vieira, and Simone Teufel. Resolving bridging references in unrestricted text. In *Proceedings of the ACL Workshop on Operational Factors in Robust Anaphora Resolution*, pages 1–6, 1997.

Poesio, Massimo, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, ACL 2004, pages 143–150, 2004.

Prince, Ellen F. Toward a taxonomy of given-new information. In Cole, P., editor, *Radical Pragmatics*, volume 14, pages 223–255. Academic Press, New York, 1981.

Prince, Ellen F. The ZPG letter: Subjects, definiteness, and information-status. In Mann, W. and S. Thompson, editors, *Discourse Description*, pages 295–325. John Benjamins, Amsterdam, 1992.

Rahman, Altaf and Vincent Ng. Learning the information status of noun phrases in spoken dialogues. In *Proceedings of the 2011 Conference on Emprical Methods in Natural Language Processing*, EMNLP 2011, pages 1069–1080, 2011.

Rahman, Altaf and Vincent Ng. 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 2012, pages 798–807, 2012.

Riester, Arndt and Stefan Baumann. The RefLex Scheme – Annotation guidelines. SinSpeC. Working papers of the SFB 732 Vol. 14, University of Stuttgart, 2017. URL `https://elib.uni-stuttgart.de/bitstream/11682/9028/1/RefLex-SinSpec14.pdf`.

Riester, Arndt, David Lorenz, and Nina Seemann. A recursive annotation scheme for referential information status. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC 2010, pages 717–722, 2010.

Rösiger, Ina. *Knowledge sources for coreference and bridging resolution (Working title)*. PhD thesis, University of Stuttgart, 2018, in preparation.

Rösiger, Ina and Simone Teufel. Resolving coreferent and associative noun phrases in scientific text. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2014, pages 45–55. Association for Computational Linguistics, 2014. URL `http://aclanthology.coli.uni-saarland.de/pdf/E/E14/E14-3006.pdf`.

Ruder, Sebastian. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016. URL `http://arxiv.org/abs/1609.04747`.

Sahlgren, Magnus. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University, 2006.

Schweitzer, Katrin, Kerstin Eckart, Markus Gärtner, Agnieszka Faleńska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwien, and Jonas Kuhn. 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, 2018.

Strube, Michael. Never look back: An alternative to centering. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, COLING-ACL 1998, pages 1251–1257, 1998.

Vieira, Renata and Simone Teufel. Towards resolution of bridging descriptions. In *Proceedings of the 35th Joint Meeting Of The Association For Computational Linguistics (Student Session)*, ACL 1997, pages 522–524, 1997.

Zinkevich, Martin A., Alex Smola, Markus Weimer, and Lihong Li. Parallelized stochastic gradient descent. In *Proceedings of the 24th Annual Conference on Neural Information Processing Sytems*, NIPS 2010, pages 2595–2603, 2010.