Rule- and Learning-based Methods for Bridging Resolution in the ARRAU Corpus

Ina Rösiger
Institute for Natural Language Processing
University of Stuttgart, Germany
roesigia@ims.uni-stuttgart.de

Abstract

We present two systems for bridging resolution, which we submitted to the CRAC shared task on bridging anaphora resolution in the ARRAU corpus (track 2): a rule-based approach following Hou et al. (2014) and a learning-based approach. The re-implementation of Hou et al. (2014) achieves very poor performance when being applied to ARRAU. We found that the reason for this lies in the different bridging annotations: whereas the rule-based system suggests many referential bridging pairs, ARRAU contains mostly lexical bridging. We describe the differences between these two types of bridging and adapt the rule-based approach to be able to handle lexical bridging. The modified rule-based approach achieves reasonable performance on all (sub-)tasks and outperforms a simple learning-based approach.

1 Introduction

Bridging (Clark, 1975) is an anaphoric phenomenon where the interpretation of a bridging anaphor, sometimes also called associative anaphor (Hawkins, 1978), is based on the non-identical associated antecedent.

The related NLP task of bridging resolution is about linking these anaphoric noun phrases and their antecedents, where both do not refer to the same referent, but are related in a way that is not explicitly stated. Bridging anaphors are thus discourse-new, but dependent on previous context.

(1) The 2018 Winter Olympics was a major multi-sport event held in February 2018 in Pyeongchang County, South Korea. **Ticket prices** were announced in April 2016 ...

Full bridging resolution combines two subtasks: (i) detecting bridging anaphors (anaphor recognition) and (ii) finding an antecedent for given bridging anaphors (anaphor resolution).

Recently, there has only been few work on these tasks (Hou et al., 2014, 2013b,a; Markert et al., 2012; Rahman and Ng, 2012), which is partly due to the lack of annotated data, which makes the application of statistical methods difficult. Most recent work has focused on the corpus ISNotes (Markert et al., 2012), on which Hou et al. (2014)'s rule-based system currently achieves state-of-the-art results.

The first shared task on bridging resolution, co-located with the workshop on computational models of reference, anaphora and coreference (CRAC), deals with the task of bridging anaphora resolution in the RST domain of the ARRAU corpus (Poesio and Artstein, 2008). The dataset used in the shared task is part of the second release of the ARRAU corpus (Uryupina et al., to appear).

This paper presents a rule-based and a learning-based system, as submitted to the shared task. We start with a re-implementation of Hou et al. (2014)'s rule-based approach, which we then apply to the ARRAU corpus. Although the approach was designed for the same domain (news), we find that the performance is very poor. Our analysis shows that this is due to two different phenomena being defined as bridging, namely referential and lexical bridging. We present the differences between lexical and referential bridging and adapt the rule-based model so that it can also handle lexical bridging. We also compare the rule-based approach with a learning-based model which has access to the same information than the rule-based system. We report the results achieved for bridging anaphora detection, bridging anaphora resolution as well as full bridging resolution on all three domains of the ARRAU corpus: the RST domain (news), TRAINS (dialogue) and PEAR (narratives). For the shared task's main focus, bridg-

Proceedings of the Workshop on Computational Models of Reference, Anaphora and Coreference, pages 23–33
New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

ing anaphora resolution in the RST domain of ARRAU, i.e. finding an antecedent for a given bridging anaphor, we achieve an accuracy of 39.8%. Surprisingly, although the rules were designed for the RST domain of ARRAU, they perform even better on the two other domains. The rule-based system outperformed the learning-based one in every setting.

2 The ARRAU corpus

The second release of the ARRAU corpus, first published in Poesio and Artstein (2008), was used as the data basis for the shared task. It is a multi-domain corpus that aims at "providing much needed data for the next generation of coreference/anaphora resolution systems" (Uryupina et al., to appear). The current version of the dataset contains 350K tokens and 5512 bridging anaphors. The shared task data comprises text from three domains: RST (newspaper), TRAINS (dialogues) and the PEAR stories (narrative text). Following earlier attempts on the reliable annotation of bridging (Poesio, 2004), where it became evident that better annotation quality could be achieved by limiting the annotation to the three relations `subset, element` and `poss`, most of the bridging relations in ARRAU are of these types, as shown in Table 2. Additionally, comparative anaphora are included and marked as `other`, and bridging cases which do not fit the pre-defined relations, but are obvious cases of bridging, are marked with the relation `undersp-rel`.

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

Table 1: Number of bridging anaphors in the single domains of the ARRAU corpus

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

Table 2: Bridging relations in ARRAU

3 Bridging definition

Bridging has been studied in many theoretical studies (Clark, 1975; Hawkins, 1978; Hobbs et al., 1993; Asher and Lascarides, 1998) as well as in corpus and computational studies (Fraurud, 1990; Poesio et al., 1997; Vieira and Teufel, 1997; Poesio and Vieira, 1998; Poesio et al., 2004; Nissim et al., 2004; Nedoluzhko et al., 2009; Lassalle and Denis, 2011; Baumann and Riester, 2012; Cahill and Riester, 2012; Markert et al., 2012; Hou et al., 2013b,a; Hou, 2016; Zikánová et al., 2015; Grishina, 2016; Roitberg and Nedoluzhko, 2016; Riester and Baumann, 2017). Unlike in work on coreference resolution, these studies do not follow an agreed upon definition of bridging. Many issues have been controversial for a long time, for example whether definiteness should be a requirement for bridging anaphors, or the restriction to certain pre-defined relations. In this paper, we do not want to go deeper into the definition of bridging, but we would like to discuss one additional aspect that will be relevant in our discussion of the experiments with the ARRAU corpus: the distinction between referential and lexical bridging, inspired by, though different from, the two-level *RefLex* scheme by Baumann and Riester (2012).

Referential vs. lexical bridging We propose the terms *referential* and *lexical bridging* to distinguish two different phenomena which are currently both defined – and annotated – as bridging. **Referential bridging** describes bridging on the level of referring expressions, i.e. we are considering bridging anaphors that are truly anaphoric, in the sense that they need an antecedent in order to be interpretable, as in (2). As such, (referential) bridging anaphors are context-dependent expressions.

(2) The city is planning a new townhall and **the construction** will start next week.

Referential bridging is often a subclass of (referential) information status annotation. The corpus ISNotes (Markert et al., 2012) is one example of a corpus which solely includes referential bridging.

Lexical bridging (called *lexical accessibility* in Baumann and Riester (2012)), on the other hand, describes lexical semantic relations, such as meronymy, at the word/concept level (*house – door*), rather than at the level of referring expressions (*a house – the door*). It is important to realise that lexical relations are defined as part of the intrinsic meaning of a pair of concepts, thus, abstracting away from specific discourse referents: it is the word *door* which is a meronym of *house*, not some actual physical object or its mental image. The proper nouns *Europe* and *Spain* are in a meronymic relation and can thus be considered a case of lexical bridging. However, *Spain*, is not anaphoric, as its interpretation does not depend on the antecedent *Europe*. Lexical bridging is often annotated when certain pre-defined relations are defined as bridging.

It should be noted that lexical and referential bridging are two different phenomena with completely different properties, although they often co-occur in one and the same expression, such as in (3), where we have a relation of meronymy between the content word *house* and *door*, but also an anaphoric referring expression *the door* on the referential level.

(3) a house ... **the door**.

The second release of the ARRAU corpus contains instances of both referential and lexical bridging, with the majority of the bridging links being purely lexical bridging pairs, i.e. most expressions labeled as bridging are actually not context-dependent. This is probably because the focus of the annotation was set on the pre-defined relations.

Another relation often brought up in connection with bridging is the *subset* or *element-of* relation, which is the most common relation in ARRAU. In principle, an expression referring to an element or a subset of a previously introduced group can be of the referential type of bridging, e.g. in (4).

(4) I saw some dogs yesterday. **The small pug** was the cutest.

It should be noted, however, that at the lexical level, subset/element-of pairs have more in common with coreference pairs, e.g. (5), since the lexical relation between their head nouns tends to be hypernymy, synonymy or plain word repetition, i.e. relations which are summarised as *lexical givenness* in Baumann and Riester (2012).

(5) I saw a dog yesterday. **The small pug** was very cute.

Finally, the subset relations identified as bridging in ARRAU often comprise cases such as in (6), where *supercomputers priced between [...]* is a subset of *supercomputers*. Even if this is justifiable at the lexical level (the concept *supercomputer* is lexically given), we should note that there is, once more, no referential bridging involved here, since the expression denoting the subset can be interpreted independently of the context.

(6) Cray Computer also will face intense competition, not only from Cray Research, which has about 60 % of the world-wide supercomputer market and which is expected to roll out the C-90 machine, a direct competitor of the Cray-3, in 1991. The new company said it believes there are fewer than 100 potential customers for **supercomputers priced between 15 million and 30 million [...]**.

Distinguishing referential and lexical cases in ARRAU automatically is non-trivial, although our assumption is that many referential cases of bridging are probably included in `undersp-rel`.

4 Data preparation

The ARRAU corpus was published in the MMAX format, an XML-based format of different annotation layers. We converted the data into our own, CoNLL-12-style format and used the following annotation layers to extract information:
the word level, to obtain the words, document names and word number, the sentence level, to obtain sentence numbers, the part-of-speech level to extract POS tags and the phrase level to extract bridging anaphors, their antecedent, the bridging relation, coreference information, as well as the following attributes of the markables: gender, number, person, category, genericity, grammatical function and head word.

A couple of non-trivial issues came up during the preparation of the data: anaphors with multiple antecedents, antecedents spanning more than one sentence, empty antecedents and discontinuous markables, such as in

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

After filtering out these cases, the corpus statistics have changed, which are given in Table 3.

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

Table 3: Number of bridging anaphors in the shared task after filtering out problematic cases.

5 Evaluation scenarios and metrics

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

5.1 Tasks/evaluation scenarios

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

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

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

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

5.2 Evaluation metrics

We report our results in the form of the widely known metrics of precision, recall and F1 measure.

Internal scorer We take coreference chains into account during the evaluation, i.e. the predicted antecedent does not have to be the exact gold antecedent to be considered correct, as long as they are in the same coreference chain.

For bridging anaphora resolution (all), i.e. when anaphors are given and one antecedent has to be determined for all anaphors, precision and recall are the same, so in this case we report accuracy.

Official scorer Recently, the official scorer for the evaluation of the shared task has become available, which differs from our internal evaluation in the handling of some of the special cases (cf. Section 4 and Table 3). As we ignored these special cases, the official scores will most likely be lower than our own scores, in most of the cases.

In Section 9, we will report the performance using our own and the official scorer script (in the entity setting, which also takes coreference into account).

5.3 Data splits

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

6 Applying Hou et al. (2014)'s rule-based system to ARRAU

As a starting point, we adopt the approach by Hou et al. (2014) and re-implement a rule-based system for full bridging resolution[1]. The system contains eight rules. The input to the rules are the gold markables. Before applying the rules, we filter out coreferent anaphors, as this increases precision, even with predicted coreference. Each rule then proposes bridging pairs, independently of the other rules. For a more detailed description, please refer to the original paper.

Our re-implementation achieves comparable results to the original version on ISNotes (Markert et al., 2012), the corpus on which the rules were designed, with an F1 score of 17.8 for full bridging resolution (Hou et al. (2014) report an F1 score of 18.4, but on a different, unknown test-development split).

When applying the re-implementation to the complete RST dataset, the performance drops to an F1 score of 0.3 for the task of full bridging resolution, although both datasets are of the same domain (WSJ articles). We investigated the reasons for this huge difference and analysed the rules and their predicted bridging pairs. Table 4 shows the rules and their performance on the RST dataset.

We soon realised that the annotations differ quite a lot with respect to the understanding of the category bridging, as described in the section about referential and lexical bridging. We noticed that besides predicting wrong pairs, the original system would suggest bridging pairs which are fine from a referential point of view on bridging, but are not annotated in the corpus, such as in

[1] The system will be made available: `https://github.com/InaRoesiger/BridgingSystem`

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

(9) I heard from **friends** yesterday that ...

Additionally, it would miss a lot of lexical bridging pairs, as these often involve mentions with matching heads, which are filtered out in the preprocessing step of the system because they tend to signal coreferent anaphors, such as in

(10) Her husband and older son [...] run a software company. Certainly life for her has changed considerably since the days in Kiev, when she lived with her parents, her husband and **her two sons** in a 2 1/2-room apartment. (*relation: element-inverse*).

This is why the performance is so poor: a lot of referential bridging pairs which are not annotated were predicted, while the system missed almost all cases of lexical bridging.

In the remainder of this section, we give examples of some of the correct and incorrect pairs (according to the gold standard in ARRAU), as proposed by the respective rules. Note that some of the incorrect cases (according to the gold standard) might actually be good bridging pairs.

Rule 1: Building parts

(11) Once inside, she spends nearly four hours measuring and diagramming each room in the 80 year-old house [...] She snaps photos of **the buckled floors** ... (correct)

(12) And now Kellogg is indefinitely suspending work on what was to be a 1 billion cereal plant. The company said it was delaying **construction** ... (wrong)

Rule 2: Relatives

(13) I heard from **friends** that state farms are subsidized, ... (wrong)

Rule 3: GPE jobs

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

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

Rule 4: Professional roles

(15) Meanwhile the National Association of Purchasing Management said its latest survey indicated ...[]. **The purchasing managers**, however, also said that orders turned up in October ... (correct)

(16) A series of explosions tore through the huge Phillips Petroleum Co.$_{\text{pred}}$ plastics plant near here$_{\text{gold}}$, injuring more than a hundred and [...]. There were no immediate reports of deaths, but **officials** said a number of workers ...
(different antecedent)

Rule 5: Percentage expressions

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

Rule 6: Set members

(18) Back in 1964, the FBI had five black agents. **Three** were chauffeurs for ... (correct)

(19) ... a substantial number of people will be involved. **Some** will likely be offered severance package ... (wrong)

Rule 7: Argument-taking I

(20) In ending Hungary's part of the project, **Parliament** authorized ... (wrong)

(21) Sales of information-processing products$_{\text{pred}}$ increased and accounted for 46% of total sales$_{\text{gold}}$. In audio equipment, **sales** rose 13 % to ... (different antecedent)

Rule 8: Argument-taking II

(22) As aftershocks shook the San Francisco Bay Area, rescuers searched through rubble for survivors of Tuesday's temblor, and **residents** picked their way through ... (correct)

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

Table 4: Applying Hou et al. (2014) on the RST part of the ARRAU corpus
.

(23) Lonnie Thompson, a research scientist at Ohio State$_{\text{pred}}$ $_{\text{gold}}$ who dug for and analyzed the ice samples. To compare temperatures over the past 10,000 years, **researchers** analyzed ...
(different antecedent)

7 Rules for lexical and referential bridging in ARRAU

As the rule-based system is very modular, it is easy to design new rules that can also handle lexical bridging. We add a number of rather specific rules, which are meant to increase precision, but also include more general rules to increase recall. We also leave in three rules of the original rule-based system: building parts (Rule 1), percentage expressions (Rule 5) as well as set members (Rule 6).

Comparative anaphora While comparative anaphors are a different information status class in ISNotes, the ARRAU corpus contains comparative anaphors as a subclass of bridging anaphors, which are labeled as `other`. For a markable to be considered a comparative anaphor, it must contain a comparative marker[2], e.g. *two additional rules, the other country*, etc.

We then search for the closest markable which is of the same category than the anaphor and whose head matches its head in the last seven sentences. If this search is not successful, we search for an antecedent of the same category than the anaphor in the same and previous sentence. If this fails too, we search for a markable with the same head or a WordNet synonym appearing before the anaphor.

(24) the issue ... **other issues in memory**

We exclude a couple of very general terms, such as *things, matters* as potential anaphors, as they are typically used non-anaphorically, such as in *Another thing is that ...*[3].

Subset/Element-of bridging This is a rather general rule to capture mostly lexical bridging cases of the relations `subset/element`.

As the anaphor is typically more specific than the antecedent (except for cases of the relation `subset-inverse/element-inverse`), it must be modified by either an adjective, a noun or a relative clause. We then search for the closest antecedent of the same category with matching heads in the last three sentences.

(25) computers ... **personal computers**

If this fails, we check whether the head of the anaphor is a country. If so, we look for the closest antecedent with *country* or *nation* as its head in the same sentences or the previous five sentences. This is rather specific, but helps to find many pairs in the news domain.

(26) countries... **Malaysia**

If this also fails, we take the closest WordNet synonym of the same category within the last three sentences as the antecedent. Again, we use our small list of general terms to exclude rather frequent general expressions, which are typically not of the category bridging.

Time subset For this rule, we list a number of time expressions, such as *1920s, 80s, etc.*. The anaphor must have time annotated as its category and must be one of the above mentioned time expressions. We then search for the closest antecedent of the same category in the last seven sentences for which the decade numbers match.

[2] other, another, similar, such, related, different, same, extra, further, comparable, additional

[3] The full list is: *thing, matter, year, week, month.*

(27) 1920s ... **1929**.

(28) the 1950s ... **the early 1950s**

One anaphora We search for expressions where *one* is followed by a common noun. We then remember the common noun part of the expression, and search for the closest plural entity of the same category whose common noun part matches the common noun part of the anaphor. Taking into account all words with a common noun tag turned out to work better than just comparing the heads of the phrases.

(29) board members ... **one board member**

If this rule does not apply, we look for anaphor candidates of the pattern *one of the N* and again search for the closest plural entity for which the common noun parts of the expressions match.

(30) the letters ... **one of the letters**

As in a few of the other rules, we exclude a couple of very general terms as they typically do not refer back to something that has been introduced before.

Locations In the RST data, a lot of cities or areas are linked to their state/country. We can find these bridging pairs with the WordNet relation partHolonym. To be considered an anaphor, the markable must be of the category space or organization whose size is three words or less (as to exclude modification). We then search for the closest antecedent of the same category that is in a WordNet partHolonym relation with the anaphor.

(31) California ... **Los Angeles**

(32) Lebanon ... **Beirut**

Same heads This rule is very similar to the subset/element-of rule, but is designed to find more cases that have not yet been proposed by the subset/element-of rule. For a markable to be considered an anaphor, it must be a singular, short NP (containing four words or less). We then search for the closest plural expression of the same category whose head matches the head of the anaphor or that is in a WordNet synonym relation with the anaphor's head, in the last five sentences.

(33) Democrats ... **a democrat**

If this fails, we look for singular markables with a maximum size of three words which contain an adjective as anaphor candidates, and then search for a plural antecedent of the same category whose head matches the head of the anaphor or that is in a WordNet synonymy relation with the anaphor's head, in the last seven sentences.

(34) the elderly ... **the young elderly**

(35) market conditions ... **current market conditions**

If this also fails, we look for `inverse` relations, i.e. a plural anaphor and a singular antecedent of the same category and matching heads/WN synonym in the last seven sentences.

(36) an automatic call processor that ... **Automatic call processors**

Persons In this rather specific rule, we search for expressions containing an apposition which refer to a person, e.g. *David Baker, vice president*. For this, the anaphor candidate must match such a pattern and be of the category person. As the antecedent, we choose the closest plural person NP whose head matches the head of the apposition.

(37) Specialists ... **John Williams, a specialist**

The rest This rule is also very specific and aims to resolve occurrences of *the rest*, which, in many cases, is annotated as a bridging anaphor. We thus search for occurrences of *the rest* and propose as an antecedent a number expression within the last three sentences.

(38) 90 % of the funds ... **The rest**

Proposing antecedents for all remaining anaphors For the task of bridging anaphora resolution, i.e. choosing an antecedent for a given anaphor, we need to force the system to propose an antecedent for every bridging anaphor.

This is why we include a couple of rules, which are applied in the order presented here and which propose an antecedent for every anaphor which has not yet been proposed as an anaphor by the other rules.

1. Pronoun anaphors
   The anaphor must be a pronoun of the category person. As the antecedent, we choose the closest plural person NP in the last two sentences.

(39) At a recent meeting of manufacturing executives, everybody I talked with was very positive, he says. Most say **they** plan to ...

This is in a way a strange annotation, as pronouns should in theory alway be coreferent anaphors, not bridging anaphors. An alternative annotation would be to link *they* back to *most*, and *most* as a bridging anaphor to *manufacturing executives*.

2. WordNet synonyms in the last three sentences

(40) The purchasing managers ... **250 purchasing executives**

3. Cosine similarity greater than 0.5 in the last seven sentences
This is meant to find more general related cases of bridging. For the cosine similarity, we take the word2vec pre-trained vectors (Mikolov et al., 2013).

(41) "Wa" is Japanese for team spirit and Japanese ballplayers have miles and miles of it. **A player's commitment** to practice ...

4. The anaphor is a person and the antecedent is the closest organisation in the last two sentences.

5. First word head match
We choose the closest antecedent within the last two sentences, where the anaphor and the antecedent both start with a proper noun.

6. Same category in the last three sentences, choose closest

(42) ... that have funneled money into his campaign. After **his decisive primary victory over Mayor Edward I. Koch**

7. Global headmatch/WordNet synonyms: global in this case means that we search for an antecedent in the whole document, without a distance restriction.

8. Global same category

9. Choose closest NP as a fallback plan.

8 A learning-based method

To compare the performance of the rule-based system with a learning-based method, we set up an SVM classifier[4], which we provide with the same information than the rule-based system.

The classifier follows a pair-based approach similar to Soon et al. (2001), where the instances to be classified are pairs of markables. For training, we pair every gold bridging anaphor with its gold antecedent as a positive instance. As a negative instance, we pair every gold bridging anaphor with a markable that occurs in between the gold anaphor and gold antecedent[5]. During testing, we pair every markable except the first one in the document with all preceding markables. As the classifier can classify more than one antecedent-anaphor-pair as bridging for one anaphor, we choose the closest antecedent (closest-first decoding).

As the architecture of the machine learning is not designed to predict at least one antecedent for every given bridging anaphor, we cannot report results for bridging anaphora resolution (all). However, we report results for partial bridging anaphora resolution, where, during training, we pair the gold bridging anaphors with all preceding markables, instead of pairing all markables with all preceding markables.

We define the following features[6]:

Markable features words in the markable, gold head form, predicted head form, noun type (proper, pronoun, nominal), category, determiner (def, indef, demonstr, bare), number, gender, person, nested markable?, grammatical role, genericity, partial previous mention?, full previous mention?, modified by a comparative marker?, modified by an adjective?, modified by one?, modified by a number?, lengths in words.

Pair features distance in sentences, distance in words, head match?, modifier match?, WordNet synonym?, WordNet hyponym?, wordNet

[4] Using Weka's SMO classifier with a string to vector filter

[5] This is a common technique in coreference resolution, to reduce the number of negative instances and help the imbalance issue.

[6] Features marked with a ? are boolean features.

	anaphor recognition			anaphora-res.-all			anaphora-res.-partial			full bridging resolution		
	P	R	F1	P	R	F1	P	R	F1			
RST												
Rule (internal)	29.2	32.5	30.7	39.8	39.8	39.8	63.6	22.0	32.7	18.5	20.6	19.5
Rule (official)	-	-	-	36.5	35.7	36.1	58.4	20.6	30.5	16.8	13.2	14.8
ML (internal)	-	-	-	-	-	-	47.0	22.8	30.7	17.7	20.3	18.6
ML (official)	-	-	-	-	-	-	51.7	16.2	24.7	12.6	15.0	13.7
PEAR												
Rule (internal)	75.0	16.0	26.4	28.2	28.2	28.2	69.2	13.7	22.9	57.1	12.2	20.1
Rule (official)	-	-	-	30.5	28.2	29.3	62.5	11.3	19.1	53.1	4.8	8.8
ML (internal)	-	-	-	-	-	-	26.6	5.7	9.4	5.47	12.5	7.61
ML (official)	-	-	-	-	-	-	37.5	4.2	7.6	23.6	7.3	11.2
TRAINS												
Rule (internal)	39.3	21.8	24.2	48.9	48.9	48.9	66.7	36.0	46.8	27.1	21.8	24.2
Rule (official)	-	-	-	47.5	47.3	47.4	64.4	36.0	46.2	28.4	11.3	16.2
ML (internal)	-	-	-	-	-	-	56.6	23.6	33.3	10.3	14.6	12.1
ML (official)	-	-	-	-	-	-	63.2	12.8	21.3	19.0	11.0	13.9

Table 5: Performance of the different systems on the tests sets of ARRAU, using gold markables (and gold bridging anaphors in the anaphora resolution settings). We report performance using the official and our own internal scorer.

	Anaphor recognition			Full bridging resolution		
Rule	Correct pairs	Wrong pairs	Precision	Correct pairs	Wrong pairs	Precision
1: Building parts	0	0	-	0	0	-
2: Percentage	1	0	1.0	1	0	1.0
3: Set members	1	1	0.50	0	2	0.0
4: Comp anaphora	44	16	0.73	26	34	0.43
5: Subset/element	57	247	0.19	34	270	0.11
6: Time subset	3	6	0.33	3	6	0.33
7: One anaphora	0	0	-	0	0	-
8: Locations	25	11	0.69	22	14	0.61
9: Head matching	72	236	0.23	42	266	0.14
10: The rest	1	1	0.50	0	2	0.0
11: Person	10	1	0.91	8	3	0.73

Table 6: Performance of the single rules for full bridging resolution on the test set of RST, using gold markables

meronym?, WordNet partHolonym?, semantic connectivity score, highest semantic connectivity score in document?, cosine similarity.

9 Final performance

Table 5 shows the results of the modified rule-based approach and the learning-based approach for all tasks. It can be seen that the rule-based approach outperforms the learning-based one in every setting[7]. Surprisingly, in spite of the fact that the rules were designed on the training/dev sets of the RST domain, the performance for the PEAR and TRAINS domain is even better in most settings. However, these datasets are small, which is why this result should be taken with a grain of salt. Table 6 shows the rules and their performance in the final system for full bridging resolution. Some rules are included which do not predict any pairs because they predicted pairs in the training/dev setting (on which the system was designed).

10 Conclusion

We have presented two systems for full bridging resolution and bridging anaphora resolution. We started with a re-implementation of the state-of-the-art rule-based method by Hou et al. (2014), which did not achieve satisfactory performance when being applied to the ARRAU corpus. We found that the reasons for this lie in the different bridging annotations. Whereas the rule-based system suggests many referential bridging pairs, ARRAU contains mostly lexical bridging. The adapted rule-based approach achieves reasonable performance on all (sub-)tasks and outperforms a simple learning-based method.

Acknowledgments

I would like to thank Arndt Riester for his valuable comments as well as the anonymous reviewers for their insightful remarks. This work was funded by the Collaborative Research Center SFB 732, Project A6.

[7] We compute significance using the Wilcoxon signed rank test (Siegel and Castellan, 1988) at the 0.01 level.

References

Nicholas Asher and Alex Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1):83–113.

Stefan Baumann and Arndt Riester. 2012. Referential and Lexical Givenness: semantic, prosodic and cognitive aspects. In Gorka Elordieta and Pilar Prieto, editors, *Prosody and Meaning*, number 25 in Interface Explorations. Mouton de Gruyter, Berlin.

Aoife Cahill and Arndt Riester. 2012. Automatically acquiring fine-grained information status distinctions in German. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 232–236. Association for Computational Linguistics.

Herbert H. Clark. 1975. Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 169–174. Association for Computational Linguistics.

Kari Fraurud. 1990. Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7(4):395–433.

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

John A Hawkins. 1978. Definiteness and indefiniteness: A study in reference and grammaticality prediction. atlantic highlands.

Jerry R Hobbs, Mark E Stickel, Douglas E Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

Yufang Hou. 2016. *Unrestricted Bridging Resolution*. Ph.D. thesis.

Yufang Hou, Katja Markert, and Michael Strube. 2013a. Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 814–820, Seattle, USA.

Yufang Hou, Katja Markert, and Michael Strube. 2013b. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–917, Atlanta, USA.

Yufang Hou, Katja Markert, and Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2082–2093, Seattle, USA.

Emmanuel Lassalle and Pascal Denis. 2011. Leveraging different meronym discovery methods for bridging resolution in french. *Anaphora Processing and Applications*, pages 35–46.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Anna Nedoluzhko, Jiří Mírovský, Radek Ocelák, and Jiří Pergler. 2009. Extended coreferential relations and bridging anaphora in the prague dependency treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India*, pages 1–16.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. *Proceedings of The fourth international conference on Language Resources and Evaluation*.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the gnome corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, DiscAnnotation '04, pages 72–79, Stroudsburg, PA, USA. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2008. Anaphoric Annotation in the ARRAU Corpus. In *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

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

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 1–6. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Learning the fine-grained information status of discourse entities. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 798–807, Stroudsburg, PA, USA. Association for Computational Linguistics.

Arndt Riester and Stefan Baumann. 2017. The RefLex Scheme – Annotation guidelines. SinSpeC. Working papers of the SFB 732 Vol. 14, University of Stuttgart.

Anna Roitberg and Anna Nedoluzhko. 2016. Bridging corpus for russian in comparison with czech. In *CORBON@ HLT-NAACL*, pages 59–66.

Sidney Siegel and N. John Jr. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition. McGraw-Hill, Berkeley, CA.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa Rodriguez, and Massimo Poesio. to appear. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Journal of Natural Language Engineering*.

Renata Vieira and Simone Teufel. 1997. Towards resolution of bridging descriptions. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 522–524. Association for Computational Linguistics.

Šárka Zikánová, Eva Hajicová, Barbora Hladká, Pavlína Jínová, Jirí Mírovskỳ, Anja Nedoluzhko, Lucie Poláková, Katerina Rysová, Magdaléna Rysová, and Jan Václ. 2015. Discourse and coherence. *From the Sentence Structure to Relations in Text. Institute of Formal and Applied Linguistics*.