Bridging resolution: Task definition, corpus resources and rule-based experiments

Ina Rösiger, Arndt Riester and Jonas Kuhn
Institute for Natural Language Processing
University of Stuttgart, Germany
{roesigia,arndt,jonas}@ims.uni-stuttgart.de

Abstract

Recent work on bridging resolution has so far been based on the corpus ISNotes (Markert et al., 2012), as this was the only corpus available with unrestricted bridging annotation. Hou et al.'s (2014) rule-based system currently achieves state-of-the-art performance on this corpus, as learning-based approaches suffer from the lack of available training data. Recently, a number of new corpora with bridging annotations have become available. To test the generalisability of the approach by Hou et al. (2014), we apply a slightly extended rule-based system to these corpora. Besides the expected out-of-domain effects, we also observe low performance on some of the in-domain corpora. Our analysis shows that this is the result of two very different phenomena being defined as bridging, which we call referential and lexical bridging. We also report that filtering out gold or predicted coreferent anaphors before applying the bridging resolution system helps improve bridging resolution.

1 Introduction

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

(1) Our correspondent in Egypt is reporting that **the opposition** is holding a rally against **the constitutional referendum**.[1]

One can think of bridging anaphors as expressions with an implicit argument, e.g. *the opposition (in Egypt)*. Compared to coreference resolution, which has become one of the standard NLP tasks, the progress in bridging resolution is much slower. The main issue for most researchers aiming to apply statistical algorithms to this task is the lack of training data, as well as the rather diverse bridging definitions and annotations. The resolution of bridging links is important because it can prove beneficial in tasks which use the concept of textual coherence, for example Barzilay and Lapata's (2008) entity grid or Hearst's (1994) text segmentation.

Note that while a benchmark dataset for bridging has not yet been established, most recent work is based on the ISNotes corpus (Markert et al., 2012), which contains Wall Street Journal articles. Full bridging resolution on this corpus has been investigated in Hou et al. (2014), following earlier experiments on the subtasks of bridging anaphor detection (Hou et al., 2013a) and antecedent selection (Hou et al., 2013b). Apart from this, there is some work on bridging detection as a subclass of information status classification, where bridging is typically a category with low annotator agreement and low detection accuracy (Markert et al., 2012; Rahman and Ng, 2012; Hou, 2016a). In the meantime, a few other corpora

[1] Anaphors are marked in bold face, their antecedents are underlined.

Proceedings of the 27th International Conference on Computational Linguistics, pages 3516–3528
Santa Fe, New Mexico, USA, August 20-26, 2018.

have been made available (some of the same domain as ISNotes, newspaper, some of other domains) which make it possible to assess how generalisable the rule-based approach is.

The focus of this paper lies on full bridging resolution but we also report numbers for anaphor detection. We start with a re-implementation of Hou et al. (2014),[2] which we extend with one new rule to find more general bridging cases, as we think that this contributes to the generalisability of the approach, before we apply the extended system to other corpora. We also include a couple of additional experiments, where we compare predicted and gold markables and investigate the effect of coreference information. We report that filtering out gold or even just predicted coreferent anaphors before bridging resolution significantly helps improve bridging resolution.

Besides the expected out-of-domain effects, we also observe low performance on some of the in-domain corpora. Our analysis shows that this is the result of two very different – though often co-occurring – phenomena being defined as bridging, namely *referential* and *lexical bridging*, which is why we have included a rather extensive review of bridging definitions in the next section.

2 Defining bridging

Bridging has been examined in many theoretical studies (Clark, 1975; Hawkins, 1978; Hobbs et al., 1993; Asher and Lascarides, 1998; Baumann and Riester, 2012) 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; Cahill and Riester, 2012; Markert et al., 2012; Hou et al., 2013b; Hou et al., 2013a; Hou, 2016b; 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. On the contrary, many different phenomena have been described as bridging. While some of the issues have been controversial for a long time, e.g. the question of definiteness, the importance of the distinction between *referential* and *lexical* bridging, inspired by the two-level *RefLex* annotation scheme by Baumann and Riester (2012), became evident in our experiments. The two terms describe two different phenomena which are currently both defined and annotated as bridging.

2.1 Referential bridging

Referential bridging describes bridging at the level of referring expressions, i.e. we are considering noun phrases that are truly anaphoric, in the sense that they need an antecedent in order to be interpretable, like in (2). As such, (referential) bridging anaphors are non-coreferent, 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. We claim that referential bridging anaphors can be seen as expressions which require for their interpretation the antecedent as an implicit argument, e.g. *the construction of the new townhall* in (2). When uttered out of context, their referent is unidentifiable. The two above-mentioned examples are captured by this linguistically motivated definition.

Referential bridging anaphors are typically short, definite expressions (*the construction, the door*), and several accounts explicitly restrict bridging to definites, e.g. Poesio and Vieira (1998), Nedoluzhko et al. (2009), Grishina (2016), Rösiger (2016) or Riester and Baumann (2017), while others also allow for indefinite bridging, e.g. Löbner (1998) or Markert et al. (2012), with the consequence that some studies have linked indefinites as bridging anaphors (e.g. in ISNotes and others). Although having held different views on this issue, we nowadays think that indefinite expressions can indeed – in some cases – be referential bridging anaphors, for example in (3) or (4), where the (partitive) expressions *one employee (of Starbucks)* or *leaves (of the old oak tree)* are introduced.

(3) Starbucks has a new take on the unicorn frappuccino. **One employee** accidentally leaked a picture of the secret new drink.

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

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

However, while short, definite expressions signal identifiability and are thus either anaphoric expressions or familiar items, it is much harder to decide which indefinite expressions are bridging anaphors, since indefinite expressions are prototypically used to introduce new discourse referents and principally do not need an antecedent/argument in order to be interpretable. This is, for example, also reflected in the higher inter-annotator-agreement for definite than for indefinite bridging anaphors (Rösiger, 2018a).

Thus, despite the interpretational uncertainty surrounding indefinites, we take linguistic anaphoricity/context-dependence to be the defining criterion for referential bridging. Semantic relations like meronymy will be addressed in the next section under the notion *lexical bridging*. In this connection, it is important to concede, however, that the reason why certain definite or indefinite expressions function as bridging anaphors (while others do not) is typically due to some kind of semantic proximity between antecedent and anaphor. However, the specific relation we are dealing with may be rather abstract, vague and difficult to define, as the examples in (1)-(3) show.

2.2 Lexical bridging

Baumann and Riester (2012) use the term *lexical accessibility* to describe lexical semantic relations, such as meronymy or hyponymy, at the word or concept level (e.g. *house – door*). It is important to bring to mind that lexical relations are defined as part of the intrinsic meaning of a pair of concepts, thus, abstracting away from specific discourse referents: it is the words *house* and *door* which stand in a meronymic relation, not two actual physical objects or their mental images, although typically the referents of a holonym-meronym combination will, at the same time, stand in a physical whole-part relation. Since this physical relation has often been taken as one of the defining criteria for bridging, e.g. by Gardent et al. (2003), Nissim et al. (2004), Nedoluzhko et al. (2009) or Grishina (2016), we suggest to use the term *lexical* (or *lexically induced*) *bridging* for this phenomenon.

The referents of the proper nouns *Europe* and *Spain* are in a whole-part relation,[3] and the referring expressions can thus be considered a case of lexical bridging. However, the expression *Spain* is not anaphoric, since its interpretation does not depend on the "antecedent" *Europe*. Whole-part is probably the prototypical pre-defined relation, and it is a straightforward concept to annotate in the case of nouns denoting physical objects. However, it is less applicable in connection with abstract nouns, which is why many additional relations have been suggested, including, for instance *thematic role in an event, attribute of an object* (like *price*), *professional function in an organisation* (like *president*), *kinship* (like *mother*), *possessed entity* and so on. And yet, few schemes get by without an "other" category for the many examples which cannot be naturally classified into one of the assumed classes.

It should be noted that lexical and referential bridging are two different concepts with completely different properties: one deals with the question of pragmatic anaphoricity (or grammatical saturation) of an expression, the other with lexical proximity between two words and the relation between entities in the real world, although the two types of bridging often co-occur within one and the same pair of expressions, such as in (5), where we have a relation of meronymy between the content words *sea urchin(s)* and *spine(s)*, but also an anaphoric relation between the referring expressions *most sea urchins* and *the spines*, i.e. a case of referential bridging.

(5) In most sea urchins, touch elicits a prompt reaction from **the spines**.

The second release of the ARRAU corpus (Uryupina et al., to appear, first released in Poesio and Artstein, 2008), as used in the first shared task on bridging resolution, for example, contains instances of both referential and lexical bridging, with the majority of the bridging links being purely lexical bridging pairs, i.e. most expressions labeled as bridging are actually not context-dependent.

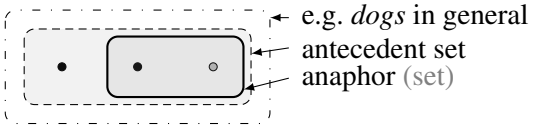
[3]Note that for proper nouns (names), like *Spain*, there is a one-to-one mapping between the word and its referent in the real world, which is not the case for common nouns, cf. Kripke (1972).

2.3 Subset relations and lexical givenness

Another relation often brought up in connection with (lexical) bridging is the *subset* or *element-of* relation, which is the most common relation in ARRAU. In principle, an expression referring to an element or a subset of a previously introduced group can be of the referential type of bridging, like in (6), where the anaphor is interpreted as *the small pug (from the prementioned group of dogs)*, but this is not always the case, as (7) shows.

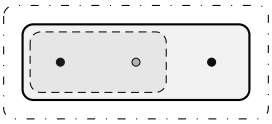
(6) I saw some dogs yesterday. **The small pug** was the cutest.

(7) Newsweek said it will introduce the Circulation Credit Plan, which awards space credits to advertisers on renewal advertising. The magazine will reward with page bonuses **advertisers who in 1990 meet or exceed their 1989 spending**, [. . .]



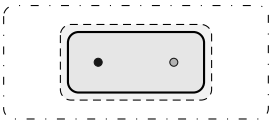
The subset relation can sometimes be reversed, as shown in (8).

(8) I saw a small pug yesterday. I like **many dogs**.



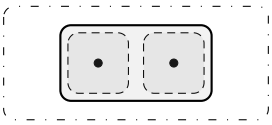
It should be noted, however, that subset/element-of pairs also have a lot in common with coreference pairs, since the lexical relation between their head nouns tends to be hypernymy, synonymy or plain word repetition (lexical relations which are summarised as *lexical givenness* in Baumann and Riester, 2012) or hyponymy (i.e. *lexical accessibility*). Note that, although the antecedent and anaphor expressions in (9) stand in a hypernym-hyponym relation (or reverse), their respective referent is the same. Hence, these cases do not exemplify bridging but coreference.

(9) a. I saw a dog yesterday. **The small pug** was very cute.
b. I saw small pugs yesterday. **The dogs** were very cute.



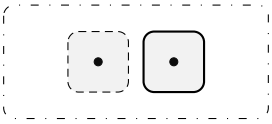
Note that *element-of* bridging is also conceptually very close to the phenomenon of aggregation/summation, in which the group entity follows a list of elements, and which also counts as a case of coreference.

(10) I saw a pug and a Yorkshire terrier. **The dogs** were very cute.



A final case, which is treated as a special class of information status in Markert et al. (2012) and annotated as a subclass of bridging in ARRAU, are so-called *comparative* or *other*-anaphors. The head noun of the anaphor must be lexically given (Riester and Piontek, 2015, 242f.) and the two expressions are marked as two contrastive elements from the same alternative set (Rooth, 1992). Comparative anaphors can be considered cases of referential bridging where the implicit argument is the implicit or explicit alternative set, i.e. *another dog (from a specific or unspecific set dogs)*.

(11) I saw a small pug two days ago and **another dog** yesterday.



2.4 Near-identity

While many approaches distinguish only between coreferent anaphors, which refer to the same referent as their antecedent, and bridging anaphors, which refer to a different referent, Recasens and Hovy (2010) and Recasens et al. (2012) have introduced a third concept, the concept of *near-identity* which has been picked up by others (e.g. Grishina, 2016). Near-identity is defined to hold between an anaphor and an antecedent whose referents are almost identical, but differ in one of four respects: name metonomy, meronymy, class or spatio-temporal functions.

(12) On homecoming night Postville feels like Hometown, USA, but a look around this town of 2,000 shows its become a miniature Ellis Island ... For those who prefer **the old Postville**, Mayor John Hyman has a simple answer.

We believe that the introduction of this additional category in between coreference and bridging introduces more uncertainty and, therefore, potentially makes the annotation process more difficult. Ex. (12), for instance, is structurally analogous to comparative anaphors.

3 Available corpus resources

Table 1 presents English corpora containing bridging annotations as well as their number of bridging pairs, limitations on the bridging anaphor, the type of bridging and the respective domain. Each corpus is quickly summarised below.

Corpus	# of pairs	Anaphor	Type	Domain
ISNotes	663	all	referential	news
ARRAU	5,512	all	mostly lexical, some referential	news, (narrative, dialogue)
BASHI	459	all	referential	news
SCiCorp	1366	only definite	referential	scientific text
CorefPro	188	only definite	referential	news, narrative, medicine
GUM	(growing)	all	referential, some lexical	dialogue, narrative, informative, instructional, ...

Table 1: Overview of available English corpora containing bridging annotations

There exist a few older corpora with bridging annotations which are not listed here because their annotations deviate from the ones presented here in one or several aspects. For example, in earlier experiments, "different-head-coreference", compare Ex. (9), was considered bridging (e.g. in Poesio and Vieira, 1998). As the bridging corpora are already rather diverse, we limit ourselves to the ones presented in Table 1.

ISNotes The ISNotes corpus (Markert et al., 2012), a corpus of newspaper text, contains bridging as a subclass of information status annotation, with 633 annotated bridging pairs. It contains definite and indefinite bridging anaphors, but no comparative anaphors, as these cases were considered a different information status category. To the best of our knowledge, the ISNotes corpus has been the only corpus on which recent results have been published, cf. Hou et al. (2013b; 2014).

ARRAU Recently, the first shared task on bridging resolution was announced. As a data basis, the second release of the ARRAU corpus was used, which contains 5,512 bridging pairs in three different domains: news text, dialogue and narrative text. This is, as far as we know, currently the largest corpus resource containing bridging pairs. However, only a small subset of the annotated pairs contains truly anaphoric bridging anaphors (cf. Section 2 for the distinction between referential and lexical bridging).

BASHI The BASHI corpus (Rösiger, 2018a) is a corpus of news text, which contains 459 bridging pairs, of which 114 are labeled as indefinite bridging anaphors and 70 as comparative anaphors.

SciCorp SCiCorp (Rösiger, 2016) is a corpus annotated with information status and bridging as a subclass. It contains scientific text of two disciplines, computational linguistics and genetics. 1366 bridging pairs were annotated. However, the bridging pairs contain pairs where possessive pre-modification was involved, as in (13).

(13) them ... **their interest**.

This is sometimes called *containing inferrable* (Prince, 1981) or *bridging-contained* (Baumann and Riester, 2012), as the antecedent is a syntactic argument within the markable. We filter out these cases, as we think that they should not be included in the category bridging proper, since anaphoricity is achieved by linking the pronoun *their* to its coreferential antecedent.

GUM The GUM corpus (Zeldes, 2017) is a corpus of (currently) 85,350 tokens The corpus is expanded every year by students as part of a curriculum at Georgetown University. As GUM contains bridging annotation as part of their information status classification, most bridging pairs are expected to be referential bridging. However, after a quick scan of the data, we also found some non-anaphoric, lexical bridging (or lexically given) pairs, e.g. in (14) or (15).

(14) However, there are four hotels in Hadibo – **Taj Socotra Hotel**, **Hafiji Hotel**, **Socotra Hotel** and **Summer land Hotel**.

(15) Name your language. This is the most fundamental property in **all languages**. You have many names to choose from [...] **most languages** [...]

GUM also contains a number of bridging-contained cases as well as some cases of aggregation, which we see as a special case of coreference, e.g. in (16).

(16) Mix .2 grams of luminon, 4 grams of sodium carbonate, .4 grams of copper sulfate [...] in a second bowl. Unfortunately, **these hazardous chemicals** will not float freely in mid-air like this graphic suggests.

CorefPro Grishina (2016) recently described a parallel corpus of German, English and Russian texts with 432 German bridging pairs that have been transferred to their English and Russian counterparts, resulting in 188 transferred English bridging pairs. The corpus has recently been made available.[4] In contrast to the other corpora, they apply a three-way classification: anaphors can be coreferent, bridging or of the category near-identity. Only definite anaphors were annotated.

4 Experimental setup

4.1 Corpora

ISNotes Hou et al. (2014) split the corpus into a development (10 documents) and test set (40 documents). The rules were optimised on the development set and the performance of the system reported on the test set. Unfortunately, the concrete development/test split is not specified. We report numbers for our own test-development-split[5] as well as for the whole corpus.

ARRAU The data was obtained from the LDC and consists of training, development and test sets for the three domains newspaper, narrative text and dialogue, with most of the text being news text. As the number of bridging anaphors in the narrative and dialogue part is quite small, we report numbers only on the test set of the news part (RST). This is also done in order to be compatible to the scores of the shared task associated with this data, which also focused on the RST domain.

BASHI The data was downloaded from the webpage.[6] As we simply apply our systems to this data, we report performance on the whole corpus.

SCiCorp The data was downloaded from the webpage.[7] Again, we report numbers on the whole corpus.

[4] https://github.com/yuliagrishina/corefpro

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

[6] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/bashi.html

[7] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/scicorp.html

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

Table 2: Overview of rules in Hou et al. (2014).

4.2 Evaluation metrics

The evaluation of bridging resolution is computed using the widely known precision and recall measures (and the harmonic mean between them, F1). The bridging anaphor is considered a mention, while the antecedent is considered an entity. This is taken into account by including gold coreference chains during the evaluation. If the predicted antecedent is one of the mentions in the coreference chain of the gold antecedent, the bridging pair is considered correct. The evaluation is rather strict, as overlapping markables are considered wrong. Furthermore, bridging anaphors with more than one link, e.g. comparative anaphora in the sense of (17), cannot be predicted with our system, as it is based on predicting just one antecedent.

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

It is also unclear how these multiple antecedents should be evaluated in case of partial correctness. In the current study, when the pair *the US* and *other countries* is suggested by the system, it is considered wrong. The same holds for antecedents with discontinuous or empty antecedents, which are contained in the ARRAU corpus.

5 Re-implementation of Hou et al. (2014)

As a basis for our experiments, we re-implement the rule-based system proposed by Hou et al. (2014).[8] For more details on the re-implementation, please refer to the supplementary material of this paper.[9]

The re-implementation comprises all three components of the original paper: pre-processing, rule adaptation and post-processing. During pre-processing, markables are extracted, which are then passed on to eight rules which predict bridging anaphor and antecedent pairs. These eight hand-crafted rules are based on linguistic intuitions about (referential) bridging. Most of the rules are rather specific, aiming at high precision, while some of the rules are designed to capture more general bridging cases, thus increasing the recall. Finally, the rules are applied in order of their precision. We extract NPs as our predicted markables. We also extract the markables of the information status annotation as our set of gold markables, where available. These form the initial set of anaphors and antecedents. We follow Hou et al.'s (2014) suggestion to exclude NPs whose head appeared before in the document, as these cases are typically involved in coreference chains. We also experiment with filtering out predicted and gold coreferent anaphors before applying the rules, described in Section 5.4.

5.1 Rules

Each rule is applied separately to the list of extracted markables and proposes pairs of bridging anaphors and antecedents. Table 2 gives an overview of the rules implemented. Two measures are computed independently of the actual bridging resolver and are needed as input for several rules, the semantic connectivity and the argument taking ratio. In order to find more general bridging cases, we have also implemented one additional rule, which is explained in Section 5.3.

[8]The re-implementation was necessary because the source code has not been made publicly available.

[9]Code and supplementary material: `https://github.com/InaRoesiger/BridgingSystem`

Computing the semantic connectivity The semantic connectivity between two words can be approximated by the number of times two words occur in a noun preposition noun pattern in a big corpus. This means that two nouns like *window* and *room* have a high semantic connectivity because they often occur as *windows in the room*, whereas other nouns do not appear often in such a construction and are therefore not highly semantically connected. Following Hou et al. (2014), we take the GigaWord corpus as the basis for the computation of the scores.

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

5.2 Results

Setting	Corpus	Anaphor recognition			Full bridging		
		Precision	Recall	F1	Precision	Recall	F1
Hou (2014), gold markables	test set	61.7	18.3	28.2	42.9	11.9	18.6
Re-implementation with gold markables							
	test set	73.4	12.6	21.6	60.6	10.4	17.8
	whole corpus	65.9	14.1	23.2	57.7	10.1	17.2
Re-implementation with predicted markables							
	test set	69.3	12.2	20.7	57.7	10.1	17.2
	whole corpus	65.2	13.6	22.5	49.2	10.3	17.0
Extended re-implemenation (+new rule) using gold markables							
	test set	51.7	17.1	25.7	36.8	12.2	18.3
	whole corpus	45.9	18.3	26.2	32.0	12.8	18.3
Filtering out coreferent anaphors							
No coreference	whole corpus	45.9	18.3	26.2	32.0	12.8	18.3
Predicted coreference	whole corpus	68.6	18.3	28.9	47.9	12.8	20.2
Gold coreference	whole corpus	71.6	18.3	29.2	50.0	12.8	20.4

Table 3: Performance of the re-implementation of Hou et al. (2014), with different settings

Hou et al. (2014) state a precision of 42.9, a recall of 11.9 and an F1 score of 18.6 for full bridging resolution and a precision of 61.7, a recall of 18.3 and F1 score of 28.2 for anaphor detection. In both settings, they use gold markables but no coreference information. Table 3 contains the scores of the re-implementation for the test and the whole corpus when using gold or predicted markables. As mentioned above, we have defined a different test-development-split, which is why the results are not directly comparable. In general, however, we think that our re-implementation achieves comparable results, as our rules also achieve similar precision values and firing rates than in Hou (2016b), which are not shown here due to lack of space. As we have simply re-implemented the system from the original paper without any hand-tuning on the development set, we also report the numbers on the whole ISNotes corpus. Here, our re-implementation yields a precision of 57.7, a recall of 10.1 and an F1 score of 17.2 for full bridging resolution. Compared to the original numbers in Hou et al. (2014), we achieve higher precision, but lower recall, resulting in an overall lower F1 measure.

5.3 New rule and final performance

In order to include more general information and to increase recall, we apply the distributional (DS) classifier described in Shwartz and Dagan (2016) to distinguish certain semantic relations, e.g. hyponyms and meronyms. The classifiers input are word embeddings, taken from ConceptNet (Speer et al., 2017). As we expect the prototypical bridging relation to be the relation of meronymy (part-whole), we include this information in our bridging resolver, in the form of the following rule: the anaphor has to be a definite, unmodified expression in the form of *the N*. We search for an antecedent within the last three sentences

Corpus	Domain	Bridging type	Anaphor recognition Prec.	Recall	F1	Full bridging Prec.	Recall	F1
ISNotes (gold markables)	news	ref.	71.6	18.3	29.2	50.0	12.8	20.4
ISNotes (pred markables)	news	ref.	64.1	18.3	28.5	41.4	11.9	18.4
BASHI (pred)	news	ref.	49.4	20.2	28.7	24.3	10.0	14.1
ARRAU (original, gold mark.)	news	ref./lex.	13.3	0.9	1.7	2.2	0.2	0.3
ARRAU (adapted, gold mark.)	news	ref./lex.	29.2	32.3	30.8	18.5	20.6	19.5
SciCorp (pred)	scientific	ref.	17.7	0.9	8.1	3.2	0.9	1.5

Table 4: Performance of the rule-based method on other corpora. We use gold markables for ARRAU in order to be compatible with the shared task results, and predicted mentions for BASHI and SciCorp as they do not contain gold markables.

for which the classifier has predicted the pair of antecedent-anaphor to be an instance of meronymy. Additionally, the constraint that the two pairs need to have a cosine similarity score over the threshold of 0.2 has increased results (this threshold has been optimised on the development set). The new rule decreases precision, but significantly increases recall[10], which is why we get a significant increase in F1 score. The final system performance is shown in Table 3. More details on our experiment on integrating predictions from neural-net relation classifiers can be found in Rösiger et al. (2018).

5.4 Filtering out coreferent anaphors: Gold vs. predicted

Bridging anaphors are difficult to distinguish from coreferent anaphors, as they both often are short, unmodified expressions which are either interpretable because they are coreferent with a previously mentioned entity or are bridging anaphors which require an antecedent for their interpretation. Thus, we think it may be beneficial for the precision of our system to filter out coreferent anaphors before applying the bridging system. We experiment with three settings: (i) no coreference information, (ii) predicted coreference information and (iii) gold annotated coreference information. For predicted coreference, we applied the IMSHotCoref system (Björkelund and Kuhn, 2014) with its default settings on the ISNotes corpus.[11] We report the change in performance on the whole corpus, as there was no optimisation involved in the filtering of the coreferent anaphors. In Table 3, it can be seen that both predicted and gold coreference significantly improve the precision of the system, as well as the final F1 score. Surprisingly, the difference between gold and predicted coreference is small, compared to having no coreference information at all. The same effect can be observed with predicted mentions. We use gold coreference for the remaining experiments of the paper, as it is available in all corpora on which we want to apply our system.

6 Application to other corpora

6.1 BASHI (in-domain)

We first apply our re-implementation to a corpus of the exact same domain, BASHI. As can be seen in Table 4, the F1 score for anaphor recognition is 28.7, which is comparable to the score on ISNotes, although we observe a much lower precision on BASHI. Lower precision is also the reason for the overall lower score on BASHI for full bridging resolution, which means that the performance for anaphor detection is about the same, while we are worse in finding the correct antecedent. Still, the system performs relatively well on this data.

6.2 ARRAU (in-domain)

We apply our system to the test set of the RST (news) part, without making any changes. As can be seen in Table 4, the performance is extremely poor. We carefully analysed the reasons for the huge difference in performance between ISNotes/BASHI and ARRAU, which both contain Wall Street Journal articles and can thus not be explained with domain effects. We soon realised that the annotations differ quite a lot with respect to the understanding of the category bridging. We noticed that besides predicting wrong

[10] We compute significance using the Wilcoxon signed rank test (Siegel and Castellan, 1988) at the 0.05 level.

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

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 (18).

(18) As competition heats up in Spain's bank market, [...] **The government** directly owns...

Additionally, it would miss a lot of lexical bridging and subset/lexical givenness pairs, as these often involve mentions with matching heads, which are filtered out in the pre-processing step of the system, such as in (19).

(19) 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 (pure) lexical bridging. With the modular approach of the rule-based system, however, one can define new rules to also capture lexical bridging. The rules have been developed on the training and development set of the RST domain of the corpus and include rather specific rules, to find e.g. locations *Hollywood – LA* or *Los Angeles – California*, as well as more general rules to find the predominant relations in the corpus, namely subset and element-of. For more details on the adapted rule-based system, please refer to Rösiger (2018b). The final performance of the adapted system (F-score of 19.5) is also given in Table 4.

6.3 SciCorp (out-of-domain)

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

7 Conclusion and future work

We have presented a re-implementation of Hou et al. (2014), the state-of-the-art system for bridging resolution (which will be made publicly available), and applied it to other corpora, to test the generalisability of the approach. Besides the expected out-of-domain effects, we also observe low performance on some of the in-domain corpora. Our analysis shows that the system generalises well to in-domain corpora, if they are of the same type of bridging, namely referential bridging. We think that the distinction between referential and lexical bridging is a valuable contribution towards the understanding of the phenomenon of bridging and that it can also help design computational approaches. We also report that filtering out gold or predicted coreferent anaphors before bridging resolution helps improve bridging resolution.

As the GUM and CorefPro corpus have in the meantime been made available, we are planning to analyse the bridging annotations in these corpora according to our categorisation as well as to test the rule-based system on it.

References

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

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

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.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47–57, Baltimore, Maryland, June. Association for Computational Linguistics.

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.

Claire Gardent, Hélène Manuélian, and Eric Kow. 2003. Which bridges for bridging definite descriptions? In *Proceedings of EACL: Fourth International Workshop on Linguistically Interpreted Corpora*, pages 69–76, Budapest.

Yulia Grishina. 2016. Experiments on bridging across languages and genres. In *CORBON@ HLT-NAACL*, pages 7–15.

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

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

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

Yufang Hou, Katja Markert, and Michael Strube. 2013a. Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In *EMNLP*, pages 814–820.

Yufang Hou, Katja Markert, and Michael Strube. 2013b. Global inference for bridging anaphora resolution. In *Proceedings of NAACL-HLT*, pages 907–917.

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 *EMNLP*, pages 2082–2093.

Yufang Hou. 2016a. Incremental fine-grained information status classification using attention-based lstms. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1880–1890.

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

Saul Kripke. 1972. Naming and necessity. In Donald Davidson and Gilbert Harman, editors, *Semantics of Natural Language*, pages 253–355. Springer, Dordrecht.

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

Sebastian Löbner. 1998. Definite associative anaphora. *manuscript) http://user. phil-fak. uniduesseldorf. de/˜ loebner/publ/DAA-03. pdf.*

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. Association for Computational Linguistics.

Anna Nedoluzhko, Jiří Mírovskỳ, 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. *LREC 2004*.

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

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.

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.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In *Radical Pragmatics*, pages 223–55. Academic Press.

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.

Marta Recasens and Eduard Hovy. 2010. A typology of near-identity relations for coreference (nident). In *LREC*.

Marta Recasens, Maria Antonia Marti, and Constantin Orasan. 2012. Annotating near-identity from coreference disagreements. In *LREC*, pages 165–172.

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

Arndt Riester and Jörn Piontek. 2015. Anarchy in the NP. When new nouns get deaccented and given nouns dont. *Lingua*, 165(B):230–253.

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

Mats Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.

Ina Rösiger, Maximilian Köper, Kim Anh Nguyen, and Sabine Schulte im Walde. 2018. Integrating predictions from neural-network relation classifiers into coreference and bridging resolution. In *Proceedings of NAACL-HLT: Workshop on Computational Models of Reference, Anaphora and Coreference*, New Orleans, USA, June.

Ina Rösiger. 2016. Scicorp: A corpus of english scientific articles annotated for information status analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Ina Rösiger. 2018a. Bashi: A corpus of wall street journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ina Rösiger. 2018b. Rule- and learning-based methods for bridging resolution in the arrau corpus. In *Proceedings of NAACL-HLT: Workshop on Computational Models of Reference, Anaphora and Coreference*, New Orleans, USA, June.

Vered Shwartz and Ido Dagan. 2016. Path-based vs. distributional information in recognizing lexical semantic relations. *arXiv preprint arXiv:1608.05014*.

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

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Olga Uryupina, Ron Artstein, Antonella Bristot, Frederica 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.

Amir Zeldes. 2017. The gum corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, Sep.

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*.