Integrating Predictions from Neural-Network Relation Classifiers into Coreference and Bridging Resolution

Ina Rösiger, Maximilian Köper, Kim Anh Nguyen and **Sabine Schulte im Walde**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
{roesigia,koepermn,nguyenkh,schulte}@ims.uni-stuttgart.de

Abstract

Cases of coreference and bridging resolution often require knowledge about semantic relations between anaphors and antecedents. We suggest state-of-the-art neural-network classifiers trained on relation benchmarks to predict and integrate likelihoods for relations. Two experiments with representations differing in noise and complexity improve our bridging but not our coreference resolver.

1 Introduction

Noun phrase (NP) coreference resolution is the task of determining which noun phrases in a text or dialogue refer to the same discourse entities (Ng, 2010). The most difficult cases in NP coreference are those which require semantic knowledge to infer the relation between the anaphor and the antecedent, as in Example (1) where we need to know that *Malaria* is a *disease*.

(1) Malaria is a mosquito-borne infection. **The disease** is transmitted via a bite ...

Related, but even more complicated is the task of bridging resolution: it requires linking anaphoric noun phrases and their antecedents which however do not refer to the same referent, but are related in a way that is not explicitly stated (Poesio and Artstein, 2005; Poesio and Vieira, 1998). Bridging anaphors are discourse-new but still depend on the preceding context. For example, for resolving *the windows* in (2) to *the room*, we need to know that a room typically has windows.

(2) I went into the room. **The windows** were broken.

The semantic relation information necessary for anaphora resolution is typically integrated into a system through a knowledge base, by relying on WordNet, Wikipedia or similar resources (cf. Vieira and Poesio (2000), Ponzetto and Strube (2007), a.o.). Up to date, few approaches have tried to integrate automatically induced information about semantic relations (e.g. Poesio et al. (2002); Feuerbach et al. (2015)). In the current study, we suggest state-of-the-art neural-network classifiers to predict semantic relations between noun pairs, and integrate the relation predictions into existing systems for coreference and bridging resolution.

2 Relation Hypotheses

Coreference signals a relation of identity, so we assume that coreference resolution should benefit from relations that link identical or highly similar entities. Obviously, synonymy is a member of this set of relations, as exemplified in Example (3):

(3) I live on Shortland Street. **The road** will be closed for repair work next week.

Hypernymy can also be used to refer to a previously introduced entity, as in Example (4):

(4) My neighbour's dog has been getting on my nerves lately. **The stupid animal** kept barking all night.

Note that the direction of this relation is important, as we can introduce a hyponym and then later refer to it via a hypernym, but not vice versa[1].

The relations between a bridging anaphor and its antecedent are assumed to be more diverse. The prototypical bridging relation is represented by meronymy:

[1] Although, in news text, you might find a certain writing style which allows for hypernyms to later be referred to via a hyponym, e.g. in "Today we are celebrating a great athlete. **The olympic swimmer** has always been one of our personal favorites."

Proceedings of the Workshop on Computational Models of Reference, Anaphora and Coreference, pages 44–49
New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

(5) My car broke down yesterday. It turned out to be a problem with **the engine**.

However, other relations come into play, too, such as attribute-of and part-of-event (Hou, 2016).

3 Experimental Setup

Data We based our experiments on the benchmark dataset for coreference resolution, the OntoNotes corpus (Weischedel et al., 2011). For bridging, we used the ISNotes corpus, a small subset of OntoNotes annotated with information status (Markert et al., 2012). In order to obtain candidate pairs for semantic relation prediction, we considered all heads of noun phrases in the OntoNotes corpus (Weischedel et al., 2011) and combined them with preceding heads of noun phrases in the same document. Due to the different corpus sizes, the generally higher frequency of coreferent anaphors and the transitivity of the coreference relation, we obtained many more coreference pairs (65,113 unique pairs) than bridging pairs (633 in total, including 608 unique pairs).

Bridging resolver As there is no publicly available bridging resolver, we re-implemented the rule-based approach by Hou et al. (2014). It contains eight rules which all propose anaphor-antecedent pairs, independently of the other rules. The rules are applied in order of their precision. Apart from information on the connectivity of two nouns, which is derived from counting how often two nouns appear in a *$noun_1$ preposition $noun_2$* pattern in a large corpus, the tool does not contain information about general relations.

Coreference resolver We used the IMS Hot-Coref resolver (Björkelund and Kuhn, 2014) as a coreference resolver, because it allows an easy integration of new features. While its performance is slightly worse than the state-of-the-art neural coreference resolver (Clark and Manning, 2016), the neural resolver relies on very few basic features and word embeddings, which already implicitly contain semantic relations.

Evaluation metrics For coreference resolution, we report the performance as CoNLL score, version 8.01 (Pradhan et al., 2014). For bridging resolution, we report performance in precision, recall and F1. For bridging evaluation, we take coreference chains into account during the evaluation, i.e. the predicted antecedent is considered correct if it is in the same coreference chain as the gold antecedent. We applied train-development-test splits, used the training and development set for optimisation, and report performance on the test set.

4 First Experiment

4.1 Semantic Relation Classification

We used the publicly available relation resource BLESS (Baroni and Lenci, 2011), containing 26,546 word pairs across the six relations co-hyponymy/coordination, attribute, meronymy, hypernymy, and random. As classification method, we relied on the findings from Shwartz and Dagan (2016), and used a plain distributional model combined with a non-linear classifier (neural network) with only word representations. As many of our target word pairs rarely or never occurred together in a shared sentence, we could not integrate intervening words or paths as additional features.

We took the publicly available 300-dimensional vectors from ConceptNet (Speer et al., 2017), combined the word representations with the semantic relation resources, and trained a feed-forward neural network for classification. The input of the network is simply the concatenation of the two words, and the output is the desired semantic relation. At test time we present two words and output the class membership probability for each relation. In addition we provide information about the semantic similarity by computing the cosine.

We relied on the training, test and validation split from Shwartz and Dagan (2016). The hyper-parameter were tuned on the validation set and obtained the best performance by relying on two hidden layers with 200 and 150 neurons respectively. As activation function we applied rectified linear units (ReLU). Despite, we set batch size to 100 and used a dropout rate of 20%.

Intrinsic Evaluation To validate that the semantic relation classification works to a sufficient degree, we performed an intrinsic evaluation. On the test set from Shwartz and Dagan (2016), our model achieved an accuracy of 87.8%*, which is significantly[2] better than the majority class baseline (i.e. the random class with 45%). Shwartz and Dagan report a weighted average F-score of

[2] We used the χ^2 test * with $p < 0.001$.

89, which is only marginally better than our reimplementation (88).

While this performance seems very good and confirms the quality of our reimplementation, the work by Levy et al. (2015) pointed out that such supervised distributional models often just memorise whether a word is a prototypical example for a certain relation. Indeed, we found many of these cases in our dataset. For example the term 'gas' appeared $\frac{9}{10}$ times in a meronym relation in training and $\frac{4}{4}$ times as a meronym in the test set. To encounter this effect we conducted a second evaluation where we made sure that training and test set contained different terms.

With an accuracy of 58.6%* and a weighted mean F-score of .52, the performance of this second evaluation was still significantly better than the majority class baseline but considerably worse than the reported results on the BLESS train/test split with lexical overlap. Still, we assume that this evaluation provides a more realistic view on the relation classification. Results per relation are given in Table 1. It can be seen that the model is skewed towards the majority class (random), whereas in particular the hypernym relation seems to be difficult. Here we observed many false decision between coord/hyper.

Rel.	P	R	F1
Random	63.7	93.8	75.9
Coord	46.6	41.2	43.7
Attri	68.9	18.7	29.4
Mero	31.1	22.4	26.0
Hyper	25.0	0.4	0.7

Table 1: Results of the intrinsic evaluation on BLESS (without lexical overlap).

4.2 Relation Analysis

Before using the predicted relations for coreference and bridging resolution, we analysed the distribution of relations across the bridging and coreference pairs, as well as across all other, non-related pairs. Table 2 shows the average cosine similarities (COS) of these pairs. As expected, the average cosine similarity is highest for coreference pairs and a little lower for bridging pairs, but still much higher in comparison to all other pairs. In the rows below cosine similarity, we give the averages of the output probabilities of the classifier for each relation. Random represents the class for non-related pairs without a relation. Such non-related pairs have indeed a high score for not being in a relation, whereas coreference and bridging pairs have lower scores in this category. Non-related random pairs have a high score for not being in a relation, whereas coreference and bridging pairs have lower scores in this category. Both coreference and bridging pairs have high meronym values, which is surprising for the coreference pairs. Bridging pairs also have a higher coordination value (i.e. co-hyponymy), and a slightly higher value for hypernymy.

	Coref pairs	Bridging pairs	Other pairs
COS	0.26	0.19	0.05
Random	0.39	0.49	0.78
Coord	0.22	0.13	0.03
Attri	0.07	0.07	0.06
Mero	0.22	0.23	0.10
Hyper	0.09	0.07	0.02

Table 2: Average cosine similarities and relation classifier probabilities for coreferent and bridging pairs in comparison to other pairs of nouns, experiment 1.

4.3 Relations for Bridging Resolution

As short, unmodified NPs are generally considered useful bridging anaphor candidates, because they often lack an antecedent in the form of an implicit modifier, we add the following new rule to our bridging resolver: "search for an unmodified NP, in the form of *the N*", e.g. in *the advantages*. As bridging antecedents typically appear in a rather close window (cf. Hou (2016)), we search for an antecedent within the last three sentences. As bridging pairs have a higher cosine value than non-related pairs, we experiment with an additional cosine similarity constraint: if the pair is in a certain relation and the cosine similarity is greater than 0.2, it is proposed.

Table 3 shows the results for the different relations as well as the versions with and without a cosine similarity threshold, which are explored further in Table 4. Note that both tables do not give absolute numbers of correct and wrong bridging pairs, but only the bridging pairs which were proposed by the newly added semantic rule.

Meronymy seems to be the best predictor for bridging, with a significant gain of 2.38% in F1 score[3], followed by the not-random version. The precision slightly decreased, but since the rule was designed to increase recall, this is acceptable. In the best setting (meronymy, cosine threshold of

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

Baseline	-	-	-	-	-	-	-	59.82	10.58	18.0
Relation	*without cosine threshold*					*with cosine threshold of 0.2*				
	Correct	Wrong	Precision	Recall	F1	Correct	Wrong	Precision	Recall	F1
Coord	5	41	45.57	11.37	18.20	5	32	48.3	11.37	18.41
Attri	3	46	43.48	11.06	17.63	2	8	56.56	10.9	18.28
Mero	14	101	35.69	12.80	18.84	14	36	50.00	12.80	**20.38**
Hyper	2	7	57.02	10.90	18.3	2	4	58.47	10.9	18.38
Not random	17	105	35.90	13.27	19.37	15	54	45.3	12.95	**20.15**

Table 3: Correct and wrong bridging pairs which are found by the additional semantic rule, with and without additional cosine threshold constraint (> 0.2).

Threshold	Correct	Wrong	P	R	F1
0.15	16	56	44.20	12.64	19.66
0.20	14	36	50.00	12.80	**20.38**
0.25	10	26	52.03	12.16	19.72
0.30	2	22	50.74	10.90	17.95

Table 4: Effect of the cosine threshold constraint, for the relation meronymy.

0.2) we now find 14 additional correct pairs, for example:

(6) IBM said it expects industrywide efforts to become prevalent because semiconductor manufacturing has become so expensive. A state-of-the-art plant cost 40 million in the mid-1970s but costs 500 million today because **the technology** is so complex.

We also find 36 more wrong pairs, for example:

(7) In the 1980s, the Justice Department and lower federal courts that enforce the Voting Rights Act have required state legislatures and municipal governments to create the maximum number of "safe" minority election districts – districts where minorities form between 65% and 80% of **the voting population** .

4.4 Relations for Coreference Resolution

We used the following features in the resolver:

- *Random as the highest class*: a boolean feature which returns true if the random class got assigned the highest value of all the relations.
- *Cosine binned into low, middle, high*: this is a binned version of cosine similarity. We experimented with two different bins, the first one {0-0.3,0.3-0.49,>0.49}, the second one {0-0.3,0.3-0.6,>0.6}
- *Relation with the highest value*: a multi-value feature with 6 potential values: none, mero, coord, attri, hyper and random. The class with the highest value is returned.

We added one feature at a time and analysed the change in CoNLL score. The results are not shown in detail, as the score decreased in every version. For coreference resolution, where the baseline performance is already quite high, the additional semantic information thus does not seem to improve results. This is in line with Björkelund and Kuhn (2014), where integrating a WordNet synonym/hypernym lookup did not improve the performance, as well as Durrett and Klein (2013), where increased semantic information was not beneficial either.

5 Second Experiment

The first experiment had a few major shortcomings. First, we did not have lemmatised vectors, and as a result, singular and plural forms of the same lemma had different values. Sometimes, this led to the wrong analysis, cf. Example (8), where the singular and plural versions of *novel* make different predictions, and where a lemmatised version would have preferred the correct antecedent:

W1	W2	COS	coord	attri	mero
characters	novel	0.35	**0.69**	0.02	0.27
characters	novels	0.43	0.28	0.05	**0.38**

(8) In novels of an earlier vintage$_{\text{predicted}}$, David would have represented excitement and danger; Malcom, placid, middle-class security. The irony in this novel$_{\text{gold}}$ is that ... **The characters** confront a world ...

Second, many proper nouns were assigned zero values, as they were not covered by our vector representations. These pairs thus could not be used in the new rule. Third, the relations in the benchmark dataset BLESS do not completely match our hypotheses. We thus designed a second experiment to overcome these shortcomings.

5.1 Semantic Relation Classification

To address the problem with out-of-vocabulary words we relied on fasttext (Bojanowski et al., 2016), which uses subword information to create representations for unseen words. We created 100-dimensional representations by applying a window of 5 to a lemmatised and lower-cased version of DECOW14 (Schäfer, 2015). The semantic relations were induced from WordNet (Fellbaum, 1998), by collecting all noun pairs from the relations: synonymy, antonymy, meronymy, hyponymy, hypernymy. To obtain a balanced setup, we sampled 2,010 random pairs from each relation, and in addition we created random pairs without relations across files. Hyper-parameters of the neural network were identical to the ones used in the first experiment.

Intrinsic Evaluation We obtained a similar performance as before, an accuracy of 55.8%* (exp1: 58.6) and a mean weighted f-score of 55 (exp1: 52). Results per relation are shown in Table 5. Interestingly, the performances with respect to the individual relations differ strongly from the first experiment. In this second experiment, with balanced relations, meronym and antonym are well-detected whereas random performs inferior.

Rel.	P	R	F1
Random	56.7	39.0	46.2
Ant	70.0	83.4	76.3
Syn	46.3	46.5	46.4
Mero	62.1	69.5	65.6
Hyper	48.9	49.1	49.0
Hypo	47.5	47.6	47.6

Table 5: Results of the intrinsic evaluation on WordNet.

5.2 Relation Analysis

Table 6 shows that –unexpectedly– the coreference and bridging pairs in comparison to other pairs differ much less than in the first experiment.

	Coref pairs	Bridging pairs	Other pairs
COS	0.38	0.31	0.22
Random	0.13	0.15	0.21
Mero	0.18	0.15	0.17
Hyper	0.25	0.23	0.23
Hypo	0.20	0.27	0.19
Syn	0.16	0.15	0.15
Ant	0.08	0.06	0.05

Table 6: Average relation classifier probabilities and cosine similarities for coreferent and bridging pairs in comparison to other pairs of nouns, experiment 2.

5.3 Relations for Anaphora Resolution

The two setups for integrating the relation classification into bridging and coreference resolution were exactly the same as in the first experiment. The outcome is however a little disappointing. The baseline system for bridging resolution was only improved in one condition, for the relation meronymy and with a cosine threshold of 0.3, reaching F1=18.92 (in comparison to F1=20.38 in the first experiment). Regarding coreference resolution we did not obtain any improvements over the baseline, as in the first experiment.

These results correspond to the less clear differences in the relation analysis (cf. Table 6) but are unexpected because in our opinion the setup for experiment 2 in comparison to the setup for experiment 1 was clearly improved regarding the task requirements.

6 Discussion and Conclusion

As the data for which we predicted the relations does not contain labeled relations that match the categories in our hypotheses, it is difficult to assess how well the classifiers work on this data. Despite the fact that we applied state-of-the-art methods, annotating at least a small part of the data would be necessary to assess the quality of the predictions. Our analysis shows that while some of our hypotheses have been confirmed, e.g. that meronymy is the most important relation for bridging, which can be used to improve the performance of a bridging resolver, the distribution of the relations in actual corpus data seems to be more complex than our hypotheses suggested, as we find for example also cases of meronymy in the coreference pairs.

For some of the relations, the missing direction can be problematic, as the system sometimes proposes pairs where the anaphor is a superordinate to the antecedent (e.g. residents ... **city**), although as mentioned in the introduction, it typically only works vice versa (city ... **residents**).

As the performance for coreference resolution is already quite high, the predicted relations did not improve the performance. For bridging resolution, however, the performance is typically low, and further work on finding general cases of bridging seems promising.

Acknowledgments

This work was funded by the Collaborative Research Center SFB 732, projects A6 and D12.

References

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 1–10, Stroudsburg, PA, USA.

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. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods on Natural Language Processing*, Austin, USA.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, USA.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Tim Feuerbach, Martin Riedl, and Chris Biemann. 2015. Distributional semantics for resolving bridging mentions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 192–199, Hissar, Bulgaria.

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

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.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, USA.

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.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky*, pages 76–83. Association for Computational Linguistics.

Massimo Poesio, Tomonori Ishikawa, Sabine Schulte Im Walde, and Renata Vieira. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the third international conference on language resources and evaluation (LREC)*, Las Palmas, Spain.

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

Simone Paolo Ponzetto and Michael Strube. 2007. Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res.(JAIR)*, 30:181–212.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Roland Schäfer. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28 – 34.

Vered Shwartz and Ido Dagan. 2016. Path-based vs. distributional information in recognizing lexical semantic relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V), in COLING*, Osaka, Japan.

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

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

Renata Vieira and Massimo Poesio. 2000. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.