IMS HotCoref DE: A Data-Driven Co-Reference Resolver for German

Ina Rösiger and Jonas Kuhn
Institute for Natural Language Processing
University of Stuttgart
Germany
[roesigia|kuhn]@ims.uni-stuttgart.de

Abstract

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

Keywords: Co-reference Resolution, German, Adaptation, Tool

1. Introduction

Noun phrase co-reference resolution is the task of determining which noun phrases (NPs) in a text or dialogue refer to the same discourse entities (Ng, 2010). Coreference resolution has been extensively addressed in NLP research, e.g. in the CoNLL shared task 2012 and 2011 (Pradhan et al., 2012; Pradhan et al., 2011) or in the SemEval shared task 2010 (Recasens et al., 2010).

A lot of research focuses on English co-reference, resulting in a number of high performing English co-reference systems, e.g. Clark and Manning (2015), Durrett and Klein (2014) or Björkelund and Kuhn (2014).

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

This paper presents a data-driven co-reference resolution system that is based on the English IMS HotCoref system (Björkelund and Kuhn, 2014). It describes the adaptation process, the specific requirements for co-reference resolution in German text as well as the tool that is freely available for download[1].

2. Noun Phrase Coreference Resolution

Coreferent links exist between two NPs if the first NP refers back to a discourse entity that has already been introduced in the discourse and is thereby known to the reader. The referring entity in the text is called an anaphor while the entity to which the anaphor refers back is called the antecedent. Coreferent entities include pronominal NPs (1), nominal NPs (2) and named entities (3).

(1) Pronominal:
DE: Peter ging in den Supermarkt.
Er kaufte eine Pizza.[2]
EN: Peter went into the supermarket.
He bought a pizza.

(2) Nominal:
DE: Peter kaufte gestern ein neues Buch.
Der Roman war sehr unterhaltsam.
EN: Peter bought a new book yesterday.
The novel turned out to be very entertaining.

(3) Named entities:
DE: Barack Hussein Obama ist der 44. Präsident der Vereinigten Staaten. **Obama** erhielt 2009 den Friedensnobelpreis.
EN: Barack Hussein Obama is the 44th President of the Unites States. **Obama** received the Nobel Peace Prize in 2009.

3. System and Data

IMS HotCoref As a basis for the adaptation, we chose the English IMS HOTCoref system (Björkelund and Kuhn, 2014). It models co-reference within a document as a directed rooted tree. For learning, it adopts the idea of latent antecedents and exploits the tree structure for the purpose of non-local features, i.e. features that are not restricted to only the current pair of mentions. The learning algorithm has not been changed; for a detailed description of the system and the machine learning involved, please refer to the original paper.

Data The reference corpus for co-reference resolution experiments on German is TüBa-D/Z[3] (Naumann, 2006), a gold annotated newspaper corpus of 1.8 M tokens. To evaluate our system, we use version 10 as the newest dataset available as well as version 8 as this was used in the SemEval shared task. We adopt the official test, development and training set splits for the shared task data. For version 10, there was no standard split available, so we split the data ourselves.[4]

[1] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/HotCorefDe

[2] Anaphors are typed in bold face, their antecedents are underlined.

[3] http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html

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

4. Adapting the System to German

4.1. Small Adjustments

Markables The markables to be extracted can be defined by the user. The default markables for German are NPs with label NP or PN in the parse bit, personal pronouns (PPER), possessive pronouns (PPOSAT), relative pronouns (PRELS), demonstrative pronouns (PDS), reflexive pronouns (PRF) and named entities with the label LOC, PER, GPE and ORG. Using predicted annotations (tools involved are described in Section 7., the recall of the mention extraction module for TüBa-D/Z v10, is about 78%. The remaining 20% are not extracted mainly due to parsing errors. With gold annotations the recall is about 99%.

Number and gender information In the English version, this information comes in the form of a lookup from lists created by Bergsma and Lin (2006). For German (and other languages that feature grammatical gender), this type of information is much more essential, which is why we decided not to implement it as a lookup function, but rather include gender and number prediction in the pre-processing and rely on this predicted information. We have included short lookup lists for personal and possessive pronouns in case the morphological analyser does not predict a label.

Head rules The system includes a module that tries to identify the syntactic head of certain syntactic phrases. The adapted rule for German noun phrases is to take the rightmost noun, if present, and if this fails, to look for the rightmost personal pronoun. If this also fails, there is a number of backup strategies to come up with the most proper solution.

4.2. Features to Capture the Challenges When Resolving German Text

IMS HotCoref offers a wide range of language-independent features (single and pair-based). We ran a number of feature selection experiments and came up with a final set of features that performed best (included in the release). We additionally added a number of new features or changes that are explained in the following. There is also a number of new features implemented that are not explained here in detail. Please, have a look at the source code to see the different features available.

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

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

F1: Gender agreement Number agreement is one of the standard features used to find suitable antecedents for pronouns. For German, we additionally need gender agreement. Contrary to English, non-animate entities are often feminine or masculine. This makes the resolution more difficult as it introduces ambiguity (see Example (5)). Note that this is mainly relevant for pronominal reference as nominal cases do not need to have the same gender (see Example (6)).

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

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

F2: Compound head match Whereas English compounds are multi words where a simple (sub-)string match feature suffices to find similar compounds, German compounds are single words. Therefore matching a compound and its head as shown in Example (7) is a little more complicated.

(7) DE: Menschenrechtskomitteevorsitzender ... **der Vorsitzende**
EN: human rights committee chairman ... **the chairman**

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

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

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

F4: Distributional information Another source of semantic knowledge comes from distributional models, where similarity in a vector space can be used to find similar concepts. This type of information is particularly important in cases where string match does not suffice (see Example (8)) and GermaNet does not contain both markables.

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

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

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

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

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

5. Evaluation

On the newest dataset available (TüBa-D/Z, version 10), our resolver currently achieves a CoNLL score of 65.76[5]. Table 1 compares the performance of our system using gold annotations with our system trained on predicted annotations (Section 7. lists the tools involved).

IMS HotCoref DE using ...	CoNLL
gold annotations	65.76
predicted annotations	48.54

Table 1: Performance of IMS HotCoref DE on TüBa-D/Z version 10: gold vs. predicated annotations

In a post-task SemEval 2010 evaluation[6] our system achieves a CoNLL score of 48.61 in the *open, regular* track and a CoNLL score of 63.61 in the *open, gold* track. Table 2 compares our scores with the three best performing systems in the shared task, BART (Broscheit et al., 2010a; Broscheit et al., 2010b), SUCRE (Kobdani and Schütze, 2010) and TANL-1 (Attardi et al., 2010) as well as with CorZu (Klenner and Tuggener, 2011; Tuggener and Klenner, 2014).[7] The CoNLL scores for all systems have been computed using the official CoNLL scorer v8.01 and the system outputs provided on the SemEval webpage. The scores differ from those published on the SemEval website due to the newer, improved scorer script and because we did not include singletons in the evaluation.

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

Table 2: SemEval Shared Task 2010 post-task evaluation for track *regular* and *gold* (on TüBa 8), excluding singletons

The difference in CoNLL score between CorZu and our system is statistically significant. We mark statistical significance with a star.[9]

6. Ablation Experiments

For the features presented in Section 4.2., we perform ablation experiments using the gold annotations of TüBa-D/Z version 10. Statistical significance is computed for all comparisons against the best performing version.

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

Table 3: Performance of IMS HotCoref DE on TüBa-D/Z version 10: ablation experiments

Table 3 shows the results when leaving out one of the previously described features at a time. Computing all the features on a word form rather than lemma basis results in the biggest decrease in performance (about 2 CoNLL points), followed by leaving out gender agreement, GermaNet and the animacy features. Two features, compound head match and distributional information, only had a minor influence on the performance. We include them here because they

[5] On the test data, using the official CoNLL scorer v8.01, not including singletons as TüBa 10 does not contain them.

[6] http://stel.ub.edu/semeval2010-coref/

[7] Performance of CorZu:
Don Tuggener, personal communication

[8] Using gold constituency parses as available for TüBa 8

[9] We compute significance using the Wilcoxon signed rank test (Siegel and Castellan, 1988) at *the 0.01 or ** the 0.05 level.

have proven to be effective in other settings, e.g. when using regular annotations.

7. Running the System on New Texts

Pre-processing The system requires preprocessed text with the following annotations in CoNLL-12 format: part-of-speech (POS) tags, lemmata, constituency parse bits, number and gender information and (optionally) named entities. The mention extraction module, i.e. the part in the resolver that chooses the markables which we want to resolve in a later step, is based on the constituency parse bits and POS tags. It can be specified which POS tags and which non-terminal categories should be extracted. Per default, noun phrases, named entities and personal, possessive, demonstrative, reflexive and relative pronouns as well as a set of named entity labels are extracted. Note that most parsers for German do not annotate NPs inside PPs, i.e. they are flat, so these need to be inserted before running the tool. The tool works best on new texts if the same tools are used with which the training corpus has been preprocessed.

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

Format The tool takes input in CoNLL-12 format. The CoNLL-12 format is a standardised, tab-separated format in a one-word-per-line setup. Table 1 shows the information contained in the respective columns. An example document can be found on the webpage.

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

Table 4: CoNLL-12 format overview: tab-separated columns and content

Annotating co-reference in new texts This section explains how to use the pre-trained models to annotate co-reference in new documents. A manual on how to train a model is contained in the webpage documentation.

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

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

8. Related Work

In the SemEval shared task, a number of systems participated in the German track: BART (Broscheit et al., 2010a; Broscheit et al., 2010b), SUCRE (Kobdani and Schütze, 2010), TANL-1 (Attardi et al., 2010) and UBIU (Zhekova and Kübler, 2010). There were four different settings evaluated, using external resources (open) or not (closed) combined with gold vs. regular preprocessing. The performance of the three best-performing systems is summarised in Section 5.

Since then, only a few systems have been developed or improved. Ziering (2011) improved the scores of SUCRE by integrating linguistic features. This results in an improvement of the average of MUC and B3 of about 5 points. It is however difficult to compare these numbers as the scorer scripts have changed and the system output as well as the system are not publicly available.

Klenner and Tuggener (2011) implemented a rule-based incremental entity-mention co-reference-system that has since the SemEval shared task received the best results on newspaper data for German (it ws improved in Tuggener and Klenner (2014)). Krug et al. (2015) compared their rule/pass-based system tailored to the domain of historic novels with CorZu in this specific domain, restricting co-reference resolution to the resolution of persons, and found that their own system outperformed the rule-based CorZu.

Mikhaylova (2014) adapted the IMS Coref system, a predecessor of IMS HotCoref, to German as part of a Master thesis. To the best of our knowledge, however this system was not made publicly available.

For co-reference resolution of newspaper text, our system achieves state-of-the-art results. For other domains on which the system has not been trained, it is however difficult to say which co-reference system performs best. We

are positive that some of the features translate well into other domains, but this hypothesis needs to be tested for every domain. In some cases a rule-based system might be more stable.

9. Conclusion

We have presented IMS HotCoref DE, a German coreference system that has been adapted from an English coreference system, IMS HotCoref. Our results show that the system achieves state-of-the-art performance on the reference dataset TüBa-D/Z, and that integrating linguistic features designed for co-reference resolution of German text increases performance. The tool is publicly available.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments as well as Anders Björkelund and Arndt Riester for their feedback on an earlier version of this paper. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) via the SFB 732, project A6.

10. Bibliographical References

Attardi, G., Simi, M., and Dei Rossi, S. (2010). Tanl-1: Coreference resolution by parse analysis and similarity clustering. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 108–111, Uppsala, Sweden, July. Association for Computational Linguistics.

Bergsma, S. and Lin, D. (2006). Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Sydney, Australia, July. Association for Computational Linguistics.

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

Bohnet, B. and Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea, July. Association for Computational Linguistics.

Broscheit, S., Poesio, M., Ponzetto, S. P., Rodriguez, K. J., Romano, L., Uryupina, O., Versley, Y., and Zanoli, R. (2010a). Bart: A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 104–107, Uppsala, Sweden, July. Association for Computational Linguistics.

Broscheit, S., Ponzetto, S. P., Versley, Y., and Poesio, M. (2010b). Extending BART to provide a coreference resolution system for german. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

Cap, F. (2014). Morphological processing of compounds for statistical machine translation. Dissertation, Institute for Natural Language Processing (IMS), University of Stuttgart.

Clark, K. and Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In *Association of Computational Linguistics (ACL)*.

Dinu, G., Pham, N. T., and Baroni, M. (2013). DISSECT – DIStributional SEmantics Composition Toolkit. In *Proceedings of ACL*, Sofia, Bulgaria.

Durrett, G. and Klein, D. (2014). A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Faaß, G. and Eckart, K. (2013). SdeWaC – a corpus of parsable sentences from the web. In *Language Processing and Knowledge in the Web*, Lecture Notes in Computer Science. Springer.

Faruqui, M. and Padó, S. (2010). Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Klenner, M. and Tuggener, D. (2011). An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility. In *Proceedings of RANLP*, pages 178–185, Hissar, Bulgaria.

Kobdani, H. and Schütze, H. (2010). Sucre: A modular system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 92–95, Uppsala, Sweden, July. Association for Computational Linguistics.

Krug, M., Puppe, F., Jannidis, F., Macharowsky, L., Reger, I., and Weimer, L. (2015). Rule-based coreference resolution in german historic novels. In *Computational Linguistics for Literature*.

Mikhaylova, A. (2014). Koreferenzresolution in mehreren sprachen. Msc thesis, Center for Information and Language Processing, University of Munich.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, AZ, USA.

Naumann, K. (2006). Manual for the annotation of in-document referential relations. University of Tübingen.

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

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). CoNLL-2011 Shared Task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference*

on Computational Natural Language Learning: Shared Task, pages 1–27, Stroudsburg, PA, USA.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of NeMLaP*, Manchester, UK.

Tuggener, D. and Klenner, M. (2014). A hybrid entity-mention pronoun resolution model for german using markov logic networks. In *Proceedings of KONVENS 2014*, pages 21–29.

Zhekova, D. and Kübler, S. (2010). Ubiu: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99, Uppsala, Sweden, July. Association for Computational Linguistics.

Ziering, P. (2011). Feature engineering for coreference resolution in german: Improving the link feature set of sucre for german by using a more linguistic background. Diploma thesis, Institute for Natural Language Processing, University of Stuttgart.