Extending BART to Provide a Coreference Resolution System for German

Samuel Broscheit*, **Simone Paolo Ponzetto***,
Yannick Versley†, **Massimo Poesio**‡

*Seminar für Computerlinguistik, University of Heidelberg
{broscheit, ponzetto}@cl.uni-heidelberg.de
† Seminar für Sprachwissenschaft, University of Tübingen
versley@sfs.uni-tuebingen.de
‡ DISI, University of Trento
poesio@disi.unitn.it

Abstract

We present a flexible toolkit-based approach to automatic coreference resolution on German text. We start with our previous work aimed at reimplementing the system from Soon et al. (2001) for English, and extend it to duplicate a version of the state-of-the-art proposal from Klenner and Ailloud (2009). Evaluation performed on a benchmarking dataset, namely the TüBa-D/Z corpus (Hinrichs et al., 2005b), shows that machine learning based coreference resolution can be robustly performed in a language other than English.

1. Introduction

Coreference resolution is the task of identifying noun phrases that are used to refer to the same extralinguistic entity in a text (Strube, 2007). Most of the work on supervised coreference resolution has been developed for English (Soon et al., 2001; Ng and Cardie, 2002; Yang et al., 2003; Luo et al., 2004, inter alia), due to the availability of large corpora such as ACE (Walker et al., 2006) and OntoNotes (Weischedel et al., 2008). However, given the current availability of a large coreferentially annotated corpus for German, namely the TüBa-D/Z corpus (Hinrichs et al., 2005b), the development of a toolkit for rapid prototyping and experimentation enables new research directions in coreference for German.

The past years have shown increasing efforts to develop robust coreference resolution engines for German, which produced systems for the resolution of pronominal anaphora (Stuckardt, 2004; Schiehlen, 2004; Kouchnir, 2004; Hinrichs et al., 2005a), names and definite noun phrases (Versley, 2006), as well as tackling the full coreference resolution task (Hartrumpf, 2001; Strube et al., 2002; Klenner and Ailloud, 2009). Hartrumpf (2001) uses a statistical backoff model to choose between antecedent candidates that have been identified by manually designed rules. The rules identifying antecedent candidates rely on semantic knowledge from HaGenLex (Hartrumpf et al., 2003). Candidates are then ranked using a statistical backoff model that backs off to subsets of the candidates until matching examples are found. Strube et al. (2002) adapt the coreference algorithm of Soon et al. (2001) to German data: in addition to features like grammatical function and a coarse-grained semantic classification (both of which were added to the data by hand), they use minimum edit distance between mentions to improve the recall on definite descriptions and names. Klenner and Ailloud (2009) use a constraint propagation approach to globally optimize the consistency of coreference sets, based on the output of a memory-based learner using syntactic, semantic and distance features.

While all these systems achieve state-of-the-art performance, we note that, in contrast to English, none of them is freely available. This poses a high entrance barrier for researchers who want to explore coreference techniques for a language other than English. We accordingly present in the following an extension of our previous proposal to provide a flexible toolkit for coreference resolution in German.

2. System Architecture

Our starting point is the toolkit from Versley et al. (2008, BART), originally conceived as a modularized version of previous efforts from Ponzetto and Strube (2006), Poesio and Kabadjov (2004), and Versley (2006). BART's overall aim to bring together state-of-the-art approaches, including syntax-based and semantic features, has led to a design that is very modular. This design provides effective separation across several tasks, including engineering *new features* that exploit different sources of knowledge, and improving the way that coreference resolution is mapped to a *machine learning* problem. In this work we extend BART to perform coreference resolution in German.

2.1. TüBa-D/Z coreference corpus

We design and evaluate our system using version 4 of TüBa-D/Z (Hinrichs et al., 2005b). The corpus contains 32,945 sentences from which we extract 144.942 markables, i.e. referring expressions (REs) to be analyzed for a potential coreference relation. Among these, we find 52,386 coreferential links in 14,073 coreference sets.

2.2. Preprocessing and markable extraction

We start with the TüBa-D/Z corpus and convert it to the data format used by BART, namely MMAX2's (Müller and Strube, 2006) standoff XML format. As a preliminary step before the actual coreference resolution is performed, the parse trees from the treebank are used to identify minimal and maximal noun projections, as well as additional features such as number, gender, and semantic class. We create a *markable* for every nominal projection if its grammatical function is not included among the following ones:

- Appositions and additional name parts. Since TüBa-D/Z includes the hierarchical structure of nominal phrases, a noun phrase such as

[[Ute	*Wedemeier],*	*[stellvertretende*	*Vorsitzende*
[[Ute	Wedemeier],	[deputy	chair person
der	*AWO]]*		
of the	AWO]]		

is transformed into a single markable with "*Ute Wedemeier*" as its minimal span. The appositive noun phrase "*stellvertretende Vorsitzende der AWO*" is included in the maximal span of the markable, but does not get a markable of its own.

Some cases of post-modification, such as "AWO *Bremen*" or "Jahresbericht 1999" ("*annual report 1999*"), where *Bremen* or *1999* are introduced as separate noun phrases with the function label "-", are treated similarly as they are seen more as name parts rather than referential noun phrases.

- Items occurring as predicates in copula constructions (`PRED` dependency label).

John	*ist*	*[ein*	*Bauer].*
John	is	[a	farmer].

- Noun phrases governed by *als* with a comparative or predicating function.

Peter	*arbeitet*	*[als*	*Bauarbeiter].*
Peter	works	[as	a construction worker].

- Vorfeld-*es* and correlates.

Ich	*finde*	*[es]*	*schade,*	*dass*	*nichts*	*passiert.*
I	find	[it]	a pity,	that	nothing	happens.

Pronouns such as *it* in English or *es* in German are often used non-referentially (cf. Boyd et al. (2005) for English). In the case of TüBa-D/Z, virtually all cases of non-referring *es* pronouns can be easily identified by their grammatical function labels.

2.3. Baseline features

We view coreference resolution as a binary classification problem. Following similar proposals for English (Ng and Cardie, 2002), we use the learning framework proposed by Soon et al. (2001) as a baseline. Each classification instance consists of two markables, i.e. an anaphor and potential antecedent. Instances are modeled as feature vectors and are handed over to a binary classifier that decides, given the features, whether the anaphor and the candidate are coreferent or not.

Our baseline feature set is a reimplementation of the one used by Klenner and Ailloud (2009) for coreference resolution in German – including distance, part of speech, grammatical function, and head matching – together with the semantic class distinctions from Versley (2006). The semantic classes are identified using the following methods.

- We first lookup the semantic class in a computational lexicon for German (Lemnitzer and Kunze, 2002, GermaNet). We take the head lemma of the markable and search for a set of pre-defined synsets in the taxonomy, including e.g. `nMensch.1` (*Person*), `nArtefakt.2719` (*Verkehrsweg/traffic route*), `nGruppe.752` (*Organization*) and others from a set of 28 top-nodes.
- In the case of named entities, we check for honorifics, organizational suffixes, and perform a gazetteer lookup[1].
- Finally, we apply knowledge-poor methods to capture morphological patterns such as acronyms (which often appear as organization names), binnen-*I* gender-neutral forms (as in *SchneiderInnen*), and head-constructions like *43-jähriger*.

In contrast to Klenner and Ailloud (2009) who model binding and agreement as ILP constraints, we follow the original proposal from Soon et al. (2001) and include them as simple features for the classifier.

2.4. Feature engineering for German coreference

Given BART's flexible architecture, we explore the contribution of *new features* for coreference in German. Given a potential antecedent RE$_i$ and a potential anaphor RE$_j$, we compute the following features:

1/2 Person: for each RE$_i$ and RE$_j$ in turn, TRUE if it is first or second person, FALSE otherwise.

Speech: for each RE$_i$ and RE$_j$ in turn, TRUE if it is inside quoted speech, FALSE otherwise.

Node_Dist: the number of clause nodes (SIMPX, R-SIMPX) and prepositional phrase nodes (PX) along the path between RE$_j$ and RE$_i$ in the parse tree.

Partial_Match: TRUE if the head of RE$_j$ is contained in the head of RE$_i$ or vice versa, FALSE otherwise.

GermaNet_Relatedness: the semantic relatedness between RE$_i$ and RE$_j$, as found in GermaNet.

Semantic relatedness in GermaNet is computed using the Pathfinder library (Finthammer and Cramer, 2008), which uses the GermaNet API by Gurevych and Niederlich (2005). Raw relatedness scores are discretized into three categories, i.e. NOT_RELATED, SIGNIFICANTLY_RELATED or STRONGLY_RELATED, based on the study from Cramer and Finthammer (2008). In our experiments we use the measure from Wu and Palmer (1994), which has been found to be the best performing on our development data (Section 3.1).

2.5. Learning algorithm

In order to learn coreference decisions, we experiment with J48, WEKA's (Witten and Frank, 2005) implementation of the C4.5 decision tree learning algorithm (Quinlan, 1993), and a Maximum entropy classifier (Berger et al., 1996, MaxEnt) with feature combination. In addition, we explore an architecture consisting a separate classifier for pronouns and non-pronouns (i.e. common nouns and proper names, 'split' henceforth). Instances are created following Soon et al. (2001). We generate a positive training instance from each pair of adjacent coreferent markables. Negative instances are created by pairing the anaphor with

[1] The gazetteer lists are derived from the lexicon of the WCDG parser (Foth and Menzel, 2006), the UN-ECE Locode database (`http://www.unece.org/cefact/locode/`), as well as a list of person names compiled by Biemann (2002).

any markable occurring between the anaphor and the antecedent. During testing, we perform a *closest first* clustering of instances deemed coreferent by the classifier. Each text is processed from left to right: each markable is paired with any preceding markable from right to left, until a pair labeled as coreferent is output, or the beginning of the document is reached.

3. Evaluation

3.1. Evaluation metrics and results

We report in the following the MUC (Vilain et al., 1995) and Constrained Entity-Alignment F-Measure (Luo, 2005, CEAF) scores. These are computed for *true mentions only*, due to the current unavailability of a preprocessing pipeline for the automatic extraction of markables from raw text. In order to provide a fair comparison with Klenner and Ailloud (2009), we use the first 1100 documents from TüBa-D/Z and evaluate using 5-fold cross validation. The remaining documents are used as development set. Table 1 shows a comparison of the performance of different learners using our baseline feature set. Table 2 compares instead the performance between our baseline system and the ones incremented with the new features.

3.2. Discussion

The results from Table 1 show that, similar to Versley et al. (2008), J48 achieves lower performing results when compared against the MaxEnt classifier. The best results in the benchmarking evaluation are given by using the 'split' architecture, that is, performance gains can be achieved by learning a specialized classifier for different types of markables. The results show that our system, although robust, does not perform as good as the one from from Klenner and Ailloud (2009). We assume that these differences are given by the use of (1) a different clustering technique to generate the coreference sets from the markable pairs classified as coreferent (*closest first* vs. *aggressive merging*); (2) a limited context window for the generation of the training and testing instances.

When looking at the contribution of the different feature sets in Table 2, we see that the only feature yielding substantial result improvements above the baseline is PARTIAL_MATCH (+1.4% MUC F_1, +1.3% CEAF F_1). However, the best improvements are given by combining *all* features together (+2.1% MUC F_1, +3.3% CEAF F_1). These results seem therefore to indicate that, while all features (except PARTIAL_MATCH) are not effective enough for coreference when used alone, they model complementary sources of information which are indeed beneficial when exploited jointly via feature combination.

4. Conclusions and Future Work

We presented a coreference resolution system for German based on BART (Versley et al., 2008). Our effort represents the first step towards building a freely available coreference resolution system for many languages[2].

Ongoing work is currently aiming at integrating the system with a preprocessing pipeline, in order to perform end-to-end coreference resolution from raw text. Future work will concentrate on porting the systems to other languages, e.g. Italian and Spanish, as well as investigating the portability and usefulness of syntactic, morphological and semantic information across different languages, i.e. a research question which has been addressed so far only for shallow string matching features by Strube et al. (2002).

[2] BART is available at `http://bart-anaphora.org`.

5. References

Adam Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Christian Biemann. 2002. Finden von semantischen Relationen in natürlichsprachlichen Texten mit Hilfe maschinellem Lernens. Diplomarbeit, Universität Leipzig.

Adriane Boyd, Whitney Gegg-Harrison, and Donna Byron. 2005. Identifying non-referential it: a machine learning approach incorporating linguistically motivated features. In *Proceedings of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 40–47.

Irene Cramer and Marc Finthammer. 2008. An evaluation procedure for WordNet based lexical chaining: Methods and issues. In *Proc. of GWC-08*, pages 120–147.

Marc Finthammer and Irene Cramer. 2008. Exploring and navigating: Tools for GermaNet. In *Proc. of LREC '08*.

Kilian Foth and Wolfgang Menzel. 2006. Hybrid parsing: Using probabilistic models as predictors for a symbolic parser. In *Proc. of COLING-ACL-06*, pages 321–328.

Iryna Gurevych and Hendrik Niederlich. 2005. Accessing GermaNet data and computing semantic relatedness. In *Comp. Vol. to Proc. of ACL-05*, pages 5–8.

Sven Hartrumpf, Herrmann Helbig, and Rainer Osswald. 2003. The semantically based corpus HaGenLex - structure and technological environment. *Traitement automatique des langues*, 44(2):81–105.

Sven Hartrumpf. 2001. Coreference resolution with syntactico-semantic rules and corpus statistics. In *Proc. of CoNLL-01*, pages 137–144.

Erhard Hinrichs, Katja Filippova, and Holger Wunsch. 2005a. What treebanks can do for you: Rule-based and machine-learning approaches to anaphora resolution in German. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories*, pages 77–88.

Erhard Hinrichs, Sandra Kübler, and Karin Naumann. 2005b. A unified representation for morphological, syntactic, semantic and referential annotations. In *Proceedings of the ACL-05 Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 13–20.

Manfred Klenner and Étienne Ailloud. 2009. Optimization in coreference resolution is not needed: A nearly-optimal algorithm with intensional constraints. In *Proc. of EACL-09*, pages 442–450.

Beata Kouchnir. 2004. A machine learning approach to German pronoun resolution. In *Proc. of ACL-04*, pages 55–60. Student Session.

Lothar Lemnitzer and Claudia Kunze. 2002. GermaNet – representation, visualization, application. In *Proc. of LREC '02*, pages 1485–1491.

	MUC scorer			CEAF		
	R	P	F_1	R	P	F_1
J48	60.9	70.7	65.4	52.9	56.1	54.4
MaxEnt	71.2	78.0	74.4	60.4	64.0	62.2
MaxEnt split	75.6	80.8	78.1	63.2	67.0	65.0
Klenner and Ailloud (2009)	–	–	–	69.3	73.8	71.5

Table 1: Performance for different classifiers

	MUC scorer			CEAF		
	R	P	F_1	R	P	F_1
baseline (MaxEnt split)	75.6	80.8	78.1	63.2	67.0	65.0
+ 1/2 PERSON, SPEECH	76.2	80.9	78.4	63.6	67.4	65.4
+ NODE_DIST	75.7	80.9	78.2	63.3	67.1	65.1
+ PARTIAL_MATCH	77.8	81.3	79.5	64.4	68.3	66.3
+ GERMANET_RELATEDNESS	76.4	80.6	78.5	63.0	66.8	64.8
+ all features	78.4	82.2	80.2	66.3	70.3	68.3
Klenner and Ailloud (2009)	–	–	–	69.3	73.8	71.5

Table 2: Performance for different feature sets

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell Tree. In *Proc. of ACL-04*, pages 136–143.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proc. HLT-EMNLP '05*, pages 25–32.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang: Frankfurt a.M., Germany.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proc. of ACL-02*, pages 104–111.

Massimo Poesio and Mijail A. Kabadjov. 2004. A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In *Proc. of LREC '04*, pages 663–666.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proc. of HLT-NAACL-06*, pages 192–199.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufman, San Mateo, Cal.

Michael Schiehlen. 2004. Optimizing algorithms for pronoun resolution. In *Proc. of COLING-04*, pages 390–396.

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

Michael Strube, Stefan Rapp, and Christoph Müller. 2002. The influence of minimum edit distance on reference resolution. In *Proc. EMNLP-02*, pages 312–319.

Michael Strube. 2007. Corpus-based and machine learning approaches to coreference resolution. In M. Schwarz-Friesel, M. Consten, and M. Knees, editors, *Anaphors in Text. Cognitive, formal and applied approaches to anaphoric reference*, pages 207–222. Amsterdam, The Netherlands: John Benjamins.

Roland Stuckardt. 2004. Three algorithms for competence-oriented anaphor resolution. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium*, pages 157–163.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Proc. of LREC '08*.

Yannick Versley. 2006. A constraint-based approach to noun phrase coreference resolution in German newspaper text. In *Proceedings of Konferenz zur Verarbeitung Natürlicher Sprache*, pages 143–150.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45–52, San Mateo, Cal. Morgan Kaufmann.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. LDC2006T06, Philadelphia, Penn.: Linguistic Data Consortium.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. 2008. Ontonotes release 2.0. LDC2008T04, Philadelphia, Penn.: Linguistic Data Consortium.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, Cal., 2nd edition.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proc. of ACL-94*, pages 133–138.

Xiaofeng Yang, Guodung Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competition learning approach. In *Proc. of ACL-03*, pages 176–183.