Real Anaphora Resolution Is Hard

Real Anaphora Resolution is Hard

Klenner, M; Fahrni, A; Sennrich, R

Real Anaphora Resolution is Hard

Abstract

We introduce a system for anaphora resolution for German that uses various resources in order to develop a real system as opposed to systems based on idealized assumptions, e.g. the use of true mentions only or perfect parse trees and perfect morphology. The components that we use to replace such idealizations comprise a full-fledged morphology, a Wikipedia-based named entity recognition, a rule-based dependency parser and a German wordnet. We show that under these conditions coreference resolution is (at least for German) still far from being perfect.

Real Anaphora Resolution is Hard. The Case of German

Manfred Klenner & Angela Fahrni & Rico Sennrich

Institute of Computational Linguistics, Binzmuehlestrasse 14, CH-8050 Zurich
{klenner sennrich}@cl.uzh.ch
angela.fahrni@swissonline.ch

Abstract. We introduce a system for anaphora resolution for German that uses various resources in order to develop a real system as opposed to systems based on idealized assumptions, e.g. the use of true mentions only or perfect parse trees and perfect morphology. The components that we use to replace such idealizations comprise a full-fledged morphology, a Wikipedia-based named entity recognition, a rule-based dependency parser and a German wordnet. We show that under these conditions coreference resolution is (at least for German) still far from being perfect.

1 Introduction

Anaphora and coreference resolution is a central task in the course of text understanding. At the sentence level, the resolution of anaphora is a prerequisite for semantic interpretation and at the text level it contributes to coherence and discourse structure. Although a lot of work has been done in the field of coreference resolution, real systems carrying out full fledged coreference resolution including pronominal and nominal anaphora are the exception. Most of the time, researchers (including the authors of this paper) try to cut away the complexity of the task and work under idealized conditions. One can find this kind of simplifications in almost every paper presented at renowned international conferences. Among the idealization, the following are the most prominent:

1. perfect anaphoricity determination (i.e. true mentions only)
2. perfect parse trees
3. perfect functional information
4. perfect morphological analysis
5. perfect named-entity recognition

The most unrealistic and most simplifying idealization is to use true mentions (1) instead of all noun phrases (henceforth 'markables'). True mentions are those markables that are - according to the gold standard - part of a coreference chain. The majority of noun phrases in a text, however, are not in a coreference set. The determination whether a NP is anaphoric (i.e. a true mention) or not is a demanding problem, the so called anaphoricity classification problem. There are a few systems that incorporate anaphoricity classification, the majority of

systems leaves this as an implicit task to the anaphora resolution component. Separate anaphoricity classification has not (really) proven to be more successful than its implicit counterpart. Anaphoricity determination of markables is a non-trival task and cutting it away makes a system an artificial one.

Syntactic information in form of parse trees is used in state of the art systems in a number of ways. Since most of the approaches (including ours) cast anaphora resolution as a (pairwise) classification task, features are needed. Among them are e.g. the depth of embedding of a markable, the part of speech of the head of a markable and even information related to intrasentential binding constraints (*c*-command). Working with idealized syntactic information pushes performance at unrealistic heights.

One of the most discriminative information is functional, namely grammatical roles. For example, parallelism of grammatical functions of a pronoun and its antecedent candidate is a powerful feature. Fortunately, dependency parsers are quite good in the recognition of grammatical functions (a subclass of dependency labels). Thus, this kind of idealization is less serious.

Especially in medium and highly inflectional languages such as German, morphological information establishes a powerful filter. E.g. personal pronouns must unify in person, number and gender. One can get rid of all pairs that do not fulfill this condition. This reduces the number of training examples, and thereby improves the quality of the classifier (by removal of safe negative examples).

Finally, named entity recognition is crucial for coreference resolution since—at least in newspaper texts—persons, groups and institutions play an important role. They are very likely to be referred to by pronouns or nominal anaphora. To know that a markable is e.g. a person helps the classifier a lot. Again, perfect information obscures the quality of a system for real applications.

There are other dimensions that prevent current systems from really being useful. To mention but one: there are performance problems arising from theoretically interesting but rather time consuming approaches, e.g. coreference resolution on the basis of integer linear programming (ILP). It is appealing to have the means to express global constraints (e.g. transitivity of the anaphoric relation as a means to propagate binding constraints within a coreference set). But transitivity with ILP is (at least for medium and longer texts) rather time-consuming, since ten thousands of equations need to be solved.

We are not saying that these explorations under idealized conditions are all in vain. We are just arguing that it is useless to tune a system with gold standard information if one intends to (later) switch to a real-world system. One never foresees the amount of noise that is introduced by real components.

In this article we introduce a realistic system for coreference resolution for German and describe its various components. We discuss our filter-based approach to pairwise classification, give empirical results and discuss the reason for the drop of performance from an idealized setting to a real world setting. We start by describing our filter-based approach to pairwise classification and the features we are using for machine learning. They are derived on the basis of real-word preprocessing components.

2 Our Filter-Based Approach

It is common practice to cast anaphora resolution as pairwise classification. Systems differ in the features they use, but also in their training procedures (fixed window of n sentences, Soon-style flexible window) and the kind of coreference clustering (best-first, closest-first, aggressive merging) they do in order to merge positively classified pairs into a partition of coreference sets.

In a former paper we have argued that coreference clustering based on the so-called Balas order coupled with intensional constraints to ensure consistency of coreference sets performs best [Klenner and Ailloud, 2009]. In this paper, we concentrate on the features, their derivation and their quality. We do not discuss problems of coreference clustering. Just one hint why coreference clustering improves coreference resolution. The local perspective on pairs bears the danger of implicitly incompatible markables. Take the following markable chain: 'Hillary Clinton ... she ... Angela Merkel'. 'she' is compatible with 'Hillary Clinton', 'Angela Merkel' is compatible with 'she', but 'Merkel' and 'Clinton' are incompatible. Since transitivity is outside the scope of a pairwise classifier, it might classify both compatible pairs as positive without noticing that this leads to an implicit inconsistency.

Our system is filter-based, that is, only those pairs are considered as candidates that pass all filters. We have morphological, syntactic and semantic filters.

The morphological filters refer to person, number and gender. Personal pronouns must unify in each of them, while possessive pronouns only unify in person and gender, e.g. 'Er hat seine Brüder getroffen' ('He^i has met his^i brothers'), but not in number. 'seine' ('his') is plural, 'Er' ('He') is singular. Nominal anaphora in German only unify in number (and trivially in person), but not necessarily in gender ('Der Weg$^i_{masc}$ ist lang. Ich bin diese Strecke$^i_{fem}$ ...'). Each of these cases is covered by a rule and there are some rules for special cases, e.g. the rule for reported speech, where a third person pronoun is coreferent with a first person pronoun, e.g. 'Er sagte, ich ...' ('He said: I ...').

Among the syntactic filters, the subclause filter is the most prominent. It can be used to operationalize binding constraints and helps to reduce the amount of negative pairs. The constraint here is: two personal pronouns (or nouns) in the same subclause cannot be coreferent ('Sie^i gibt ihr^j das Buch', where $i \neq j$; 'She^i gives her^j the book'—in English, a reflexive pronoun is necessary to establish coreference). With possessive pronouns this is different, a possessive pronoun and its antecedent might be in the same subclause. For reflexive pronouns the antecedent even should be in the same subclause, but there are exceptions (sentences where the reflexive pronoun is not anaphoric at all).

Semantic filters are based on GermaNet [Hamp and Feldweg, 1997], the German wordnet. Two nominal markables must be semantically compatible, which means that they must be both e.g. animate or inanimate, or stand in a hyponym or synonym relation. If one of the markables is not in GermaNet, the pair does not pass the filter (reducing recall). We have also experimented with selectional restrictions available from verb frames. If a personal pronoun fills e.g. the subject slot of a verb, semantic information becomes available by the selectional

restriction of the verb slot (e.g. the subject of 'to sleep' is animate, neglecting metaphorical usages). This way, the number of valid candidate antecedents (noun phrases that are of type animate) can be further restricted.

We strive to integrate as much linguistic knowledge as possible into the filters. Alternatively, one could use this kind of linguistic knowledge as a feature. But our experiments have shown that a filter based approach is more reliable. There are only a few exceptions of these regularities (at least at the morphological and syntactic level). It's better to erroneously filter such pairs out as to let everything pass.

Any pair that has passed all filters gets classified by a machine learning programme. We use the memory-based learner TiMBL [Daelemans et al., 2004] as a classifier. This is done on the basis of following features:

- distance in sentences
- distance in markables
- part of speech of the heads (tagger)
- grammatical functions (parser)
- parallelism of grammatical functions (parser)
- salience of the grammatical functions of the heads (see below)
- depth of embedding of the heads (parser)
- whether an NP is definite or not (Gertwol)
- the semantic class (GermaNet)
- whether an NP is animate or not (GermaNet)
- whether the markables are in the same subclause (parser)

Salience of a grammatical function is estimated (on the basis of the training set) in the following way: the number of cases a grammatical function realizes a true mention divided by the number of true mentions (it's the conditional probability of a grammatical function given an anaphoric markable). The function 'subject' is the most salient function followed by 'direct object'.

3 System Components

The preprocessing step prior to pair-wise classification of anaphora candidates is crucial, since it produces the features used to describe the markables and thus indirectly determines the quality of the classifier. Fortunately, we have high performance tools available: the TreeTagger, GermaNet, Gertwol and Pro3GresDe (the parser). After tokenization and tagging the morphological analysis takes place. We use Gertwol, a commercial system based on two-level morphology. Gertwol is fast and also able to do noun decomposition which is rather helpful, since in German compounds are realized as single wordforms (e.g. Computerexperte, English: computer expert). Compounds (which are quite frequent in German) might become very complex, but often the head of the compound is sufficient to semantically classify the whole compound via GermaNet. For instance, 'Netzwerkcomputerexperte' ('expert for network computers') is an expert and, thus, is animate. Gertwol decomposes the compound and the head can be classified with

the aid of GermaNet. The other important task of Gertwol is to determine the number, person and gender information of a word. Unfortunately, ambiguity rate is high, since e.g. some personal pronouns are highly ambiguous. For instance, the German pronoun 'sie' ('she') might be singular/feminine or plural (without gender restriction). The pronoun 'ich' does not impose any gender restrictions and moreover often refers (in reported speech) to a speaker which is third person.

3.1 Named-Entity Recognition

Our Named-Entity Recognition (NER) is pattern-based, but also makes use of extensive resources. We have a large list of (international) first names (53'000) where the gender of each name is given. From Wikipedia we have extracted all multiword article names (e.g. 'Berliner Sparkasse', a credit institute from Berlin) and, if available, their categories (e.g. 'Treptower Park' has 'Parkanlage in Berlin | Bezirk Treptow-Köpenick' as its category tree; 'Parkanlage' being the crucial information').

The pattern-based NER uses GermaNet and Wikipedia and the information of the POS tagger. For instance, 'Grünen Bewegung Litauens' is a multiword named entity. 'Litauens' is genitive, thus it is not the head of the noun phrase, 'Bewegung' (here: 'group') is the head, so the whole compound denotes a group of people not a country. Since 'Grünen' is an adjective in initial caps (which is unusual), it is considered as part of the name.

Our parser takes advantage of NER, since it reduces ambiguity and grouping problems.

3.2 Pro3gresDe: the Parser

Pro3GresDe is a hybrid dependency parser for German that is based on the English Pro3Gres parser [Schneider, 2008]. It combines a hand-written grammar and a statistical disambiguation module trained on part of the TüBa-D/Z treebank [Telljohann et al., 2004].[1]. This hybrid approach has proven especially useful for the functional disambiguation of German noun phrases. While the function of noun phrases is marked morphologically in German, many noun phrases are morphologically ambiguous, especially named entities. We use both morphological unification rules and statistical information from TüBa-D/Z (i.e. data about possible subcategorisation frames of verbs) to resolve functional ambiguities. We have shown that this approach performs better at functionally disambiguating noun phrases than purely statistical parsers.

The parser give access to the following features: e.g. grammatical function, depth of embedding, subclause information.

4 Empirical Evaluation

We have evaluated our base system only, i.e. without our clustering method described in [Klenner and Ailloud, 2009]. It's the baseline performance drop that

[1] For a full discussion of Pro3GresDe, see [Sennrich et al., 2009].

we are interested in. The performance drop is measured in terms of save (gold standard) versus noisy (real-world components) morphological, functional and syntactic information. The gold standard information stems from the TüBa-D/Z treebank (phrase structure trees, topological fields, head information, morphology) which also is annotated with coreference links [Naumann, 2006]. Our experiments are restricted to nominal anaphora and personal pronouns, i.e. we exclude the very simple cases of reflexive and relative pronouns, but also possessive pronouns, since we are focusing on the most demanding classes.

We have run the system with all markables and without any gold standard information (see Fig. 1). The f-measure of these runs (5-fold cross validation) is 58.01%, with a precision of 70.89% and a recall of 49.01%. The performance is low because recall is low. Precision on the other hand is good. The recall is low, since our filters for nominal anaphora are quite restrictive (fuzzy string match, GermaNet hyponomy and synonymy restrictions). Most of the false negatives stem from such filtered out nominal pairs. Refining our filters for nominal anaphora would clearly help to improve recall. Nominal anaphora are, however, the most challenging part of coreference resolution. Another reason for low recall is: we are working with a fixed window of 3 sentences in order to limit the number of candidate pairs. Only named-entities are allowed to refer back further than 3 sentences, but not personal pronouns and normal nouns. This way, we miss some long distance anaphoric relations. Our experiments have, however, shown that it is better to restrict the search than to generate any reachable pairs: performance drops to a great extent the larger the window. If we take gold standard

	gold standard info	- morphological	- functional	- subclause (=real)
F-measure	61.49%	59.01%	58.20%	58.01%
Precicion	68.55%	69.78%	69.12%	70.89%
Recall	55.73%	51.12%	50.56%	49.01%

Fig. 1. Performance Drop

information, especially perfect morphology, perfect syntax and perfect functional information, the f-measure value is 61.49%, about 3.5% above the real-world setting. Precision drops: 68.55%, but recall significantly increases to 55.73%. Thus, the reason for performance increase is the increase of recall. How can we explain it? Let us first see how the different gold standard resources contribute to this increase. If we turn grammatical functions from 'parser given' to 'gold standard given', the increase on the baseline is small: f-measure raises from 58.01% to 58.20%. Our dependency parser is good enough to almost perfectly replace gold standard information. The same is true with syntactic information concerning the depth of embedding and subclause detection. Here as well, only a small increase occurs: the f-measure is 59.01%. But if we add perfect morphology, an increase of 3.5% pushes the results to the final 61.49%.

The reason for the increase in recall (and f-measure) is our filter-based method. Only those pairs are generated that pass the filter. If the morphol-

ogy is noisy, pairs erroneously might pass the filter and others pairs erroneously do not pass the filter. The first one spoils precision the second hampers recall.

We were quite surprised that the replacement of syntactic and functional information by real components was not the problem. Morphology is (mainly) responsible for the drop. See the next section for a comparison of our results with previous work.

5 Related Work

The work of [Soon et al., 2001] is a prototypical and often reimplemented machine-learning approach in the paradigm of pair-wise classification. Our system has a similar architecture, and our features do overlap to a great extent.

Work on coreference resolution for German is rare, most of it uses the coreference annotated treebank TüBa-D/Z. [Klenner and Ailloud, 2008] and [Klenner and Ailloud, 2009] are concerned with the consistency of coreference sets using idealized input from the TüBa-D/Z treebank.

[Versley, 2006] uses a maximum entropy model for nominal anaphora resolution, his major insight is that if information from GermaNet is available then it outperforms the statistical model. We took this finding seriously and have tried to use Wikipedia to complement GermaNet (we map Wikipedia multiword items via Wikipedia categories to GermaNet classes). [Hinrichs et al., 2005] introduce anaphora resolution (only pronouns) on the basis of a former version of the TüBa-D/Z. They also work with TiMBL. Their results are based on gold standard information and are – compared to subsequent work [Wunsch et al., 2009] that also utilized gold standard information – surprisingly high (f-measure 73.40% compared to 58.40%). We take the f-measure of the latter, namely 58.40%, as more realistic, since it is more in line with our results (61.40%). A study concerning the influence of different knowledge sources and preprocessing components on pronoun resolution was carried out by [Schiehlen, 2004].

6 Conclusion

We have introduced a realistic system for coreference resolution that makes extensive use of non-statistical resources (rule-based dependency parsing, a German wordnet, Wikipedia, two-level morphology) but at the same time is based on a state of the art machine learning approach. The system is not subject to any idealized assumptions related to the various preprocessing steps (i.e. no gold standard information is used), its empirical performance is, thus, not breathtaking. This is, however, not an embarrassing flaw. Rather, we think it is time to move away from idealized prototypes to assessing the performance of coreference resolution under real-world conditions.

We have shown that the performance drop, at least for coreference resolution in German, is mainly based on the morphological ambiguity introduced by

replacing perfect morphological descriptions with the output of a real morphological analyzer. Most surprising to us was the finding that using a parser instead of gold standard information only had a small negative effect on the results.

Acknowledgments. Our project is funded by the Swiss National Science Foundation (grant 105211-118108).

References

[Daelemans et al., 2004] Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A. (2004). TiMBL: Tilburg Memory-Based Learner.

[Hamp and Feldweg, 1997] Hamp, B. and Feldweg, H. (1997). GermaNet—a Lexical-Semantic Net for German. In *In Proc. of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

[Hinrichs et al., 2005] Hinrichs, E., Filippova, K., and Wunsch, H. (2005). A Data-driven Approach to Pronominal Anaphora Resolution in German. In Proc. of RANLP.

[Klenner and Ailloud, 2008] Klenner, M. and Ailloud, E. (2008). Enhancing Coreference Clustering. In Johansson, C., editor, *Proc. of the Second Workshop on Anaphora Resolution (WAR II)*, volume 2 of *NEALT Proceedings Series*, Bergen, Norway.

[Klenner and Ailloud, 2009] Klenner, M. and Ailloud, E. (2009). Optimization in Coreference Resolution Is Not Needed: A Nearly-Optimal Zero-One ILP Algorithm with Intensional Constraints. In *Proceedings of the EACL*.

[Lingsoft, 1994] Lingsoft (1994). Gertwol. Questionnaire for Morpholympics. In *LDV-Forum, 11(1)*, pages 17–29.

[Naumann, 2006] Naumann, K. (2006). Manual for the annotation of indocument referential relations. Electronic document: `http://www.sfs.uni-tuebingen.de/de_tuebadz.shtml`.

[Schiehlen, 2004] Schiehlen, M. (2004). Optimizing algorithms for pronoun resolution. In *Proceed. of the 20th International Conference on Computational Linguistics*.

[Schneider, 2008] Schneider, G. (2008). *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, Univ. of Zurich.

[Sennrich et al., 2009] Sennrich, R., Schneider, G., Volk, M., and Warin, M. (2009). A New Hybrid Dependency Parser for German. In *Proc. of the German Society for Computational Linguistics and Language Technology 2009 (GSCL 2009)*, pages 115–124, Potsdam, Germany.

[Soon et al., 2001] Soon, W., Ng, H., and Lim, D. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544.

[Telljohann et al., 2004] Telljohann, H., Hinrichs, E. W., and Kübler, S. (2004). The TüBa-D/Z Treebank: Annotating German with a Context-Free Backbone. In *Proc. of the Fourth Intern. Conf. on Language Resources and Evaluation*, Lisbon, Portugal.

[Versley, 2006] Versley, Y. (2006). A Constraint-based Approach to Noun Phrase Coreference Resolution in German Newspaper Text. In Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS).

[Wunsch et al., 2009] Wunsch, H., Kübler, S., and Cantrell, R. (2009). Instance Sampling Methods for Pronoun Resolution. In *Proc. of RANLP*, Borovets, Bulgaria.