A Machine Learning Approach to German Pronoun Resolution

Beata Kouchnir
Department of Computational Linguistics
Tübingen University
72074 Tübingen, Germany
kouchnir@sfs.uni-tuebingen.de

Abstract

This paper presents a novel ensemble learning approach to resolving German pronouns. Boosting, the method in question, combines the moderately accurate hypotheses of several classifiers to form a highly accurate one. Experiments show that this approach is superior to a single decision-tree classifier. Furthermore, we present a standalone system that resolves pronouns in unannotated text by using a fully automatic sequence of preprocessing modules that mimics the manual annotation process. Although the system performs well within a limited textual domain, further research is needed to make it effective for open-domain question answering and text summarisation.

1 Introduction

Automatic coreference resolution, pronominal and otherwise, has been a popular research area in Natural Language Processing for more than two decades, with extensive documentation of both the rule-based and the machine learning approach. For the latter, good results have been achieved with large feature sets (including syntactic, semantic, grammatical and morphological information) derived from handannotated corpora. However, for applications that work with plain text (e.g. question answering, text summarisation), this approach is not practical.

The system presented in this paper resolves German pronouns in free text by imitating the manual annotation process with off-the-shelf language sofware. As the avalability and reliability of such software is limited, the system can use only a small number of features. The fact that most German pronouns are morphologically ambiguous proves an additional challenge.

The choice of boosting as the underlying machine learning algorithm is motivated both by its theoretical concept as well as its performance for other NLP tasks. The fact that boosting uses the method of ensemble learning, i.e. combining the decisions of several classifiers, suggests that the combined hypothesis will be more accurate than one learned by a single classifier. On the practical side, boosting has distinguished itself by achieving good results with small feature sets.

2 Related Work

Although extensive research has been conducted on statistical anaphora resolution, the bulk of the work has concentrated on the English language. Nevertheless, comparing different strategies helped shape the system described in this paper.

(McCarthy and Lehnert, 1995) were among the first to use machine learning for coreference resolution. RESOLVE was trained on data from MUC-5 English Joint Venture (EJV) corpus and used the C4.5 decision tree algorithm (Quinlan, 1993) with eight features, most of which were tailored to the joint venturte domain. The system achieved an F-measure of 86.5 for full coreference

resolution (no values were given for pronouns). Although a number this high must be attributed to the specific textual domain, RESOLVE also outperformed the authors' rule-based algorithm by 7.6 percentage points, which encouraged further reseach in this direction.

Unlike the other systems presented in this section, **(Morton, 2000)** does not use a decision tree algorithm but opts instead for a maximum entropy model. The model is trained on a subset of the Wall Street Journal, comprising 21 million tokens. The reported F-measure for pronoun resolution is 81.5. However, (Morton, 2000) only attempts to resolve singular pronouns, and there is no mention of what percentage of total pronouns are covered by this restriction.

(Soon et al., 2001) use the C4.5 algorithm with a set of 12 domain-independent features, ten syntactic and two semantic. Their system was trained on both the MUC-6 and the MUC-7 datasets, for which it achieved F-scores of 62.6 and 60.4, respectively. Although these results are far worse than the ones reported in (McCarthy and Lehnert, 1995), they are comparable to the best-performing rule-based systems in the respective competitions. As (McCarthy and Lehnert, 1995), (Soon et al., 2001) do not report separate results for pronouns.

(Ng and Cardie, 2002) expanded on the work of (Soon et al., 2001) by adding 41 lexical, semantic and grammatical features. However, since using this many features proved to be detrimental to performance, all features that induced low precision rules were discarded, leaving only 19. The final system outperformed that of (Soon et al., 2001), with F-scores of 69.1 and 63.4 for MUC-6 and MUC-7, respectively. For pronouns, the reported results are 74.6 and 57.8, respectively.

The experiment presented in **(Strube et al., 2002)** is one of the few dealing with the application of machine learning to German coreference resolution covering definite noun phrases, proper names and personal, possessive and demonstrative pronouns. The research is based on the Heidelberg Text Corpus (see Section 4), which makes it ideal for comparison with our system. (Strube et al., 2002) used 15 features modeled after those used by state-of-the-art resolution systems for English. The results for personal and possessive pronouns are 82.79 and 84.94, respectively.

3 Boosting

All of the systems described in the previous section use a single classifier to resolve coreference. Our intuition, however, is that a combination of classifiers is better suited for this task. The concept of ensemble learning (Dietterich, 2000) is based on the assumption that combining the hypotheses of several classifiers yields a hypothesis that is much more accurate than that of an individual classifier.

One of the most popular ensemble learning methods is boosting (Schapire, 2002). It is based on the observation that finding many weak hypotheses is easier than finding one strong hypothesis. This is achieved by running a base learning algorithm over several iterations. Initially, an importance weight is distributed uniformly among the training examples. After each iteration, the weight is redistributed, so that misclassified examples get higher weights. The base learner is thus forced to concentrate on difficult examples.

Although boosting has not yet been applied to coreference resolution, it has outperformed stateof-the-art systems for NLP tasks such as partofspeech tagging and prepositional phrase attachment (Abney et al., 1999), word sense disambiguation (Escudero et al., 2000), and named entity recognition (Carreras et al., 2002).

The implementation used for this project is BoosTexter (Schapire and Singer, 2000), a toolkit freely available for research purposes. In addition to labels, BoosTexter assigns confidence weights that reflect the reliability of the decisions.

4 System Description

Our system resolves pronouns in three stages: preprocessing, classification, and postprocessing. Figure 1 gives an overview of the system architecture, while this section provides details of each component.

4.1 Training and Test Data

The system was trained with data from the Heidelberg Text Corpus (HTC), provided by the European Media Laboratory in Heidelberg, Germany.

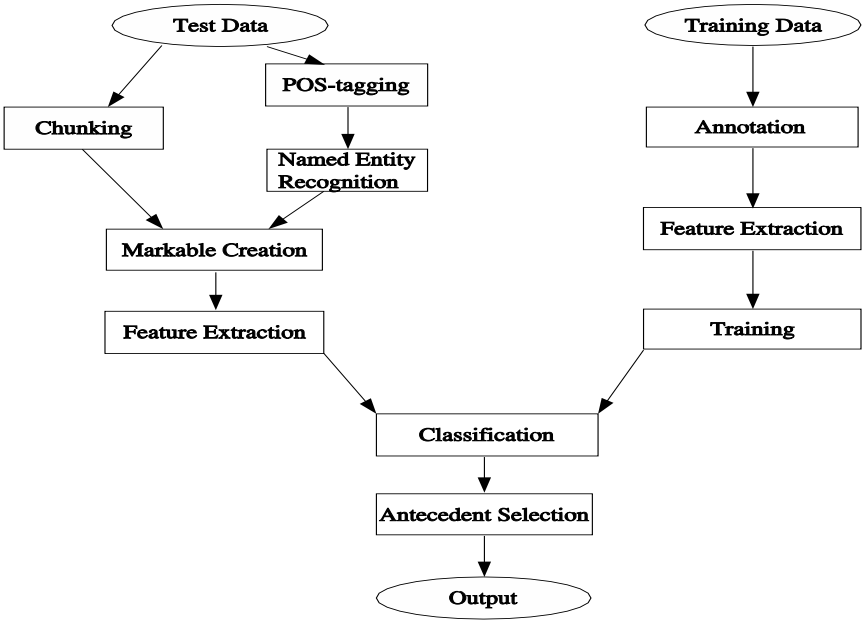

Figure 1: System Architecture

The HTC is a collection of 250 short texts (30-700 tokens) describing architecture, historical events and people associated with the city of Heidelberg. To examine its domain (in)dependence, the system was tested on 40 unseen HTC texts as well as on 25 articles from the Spiegel magazine, the topics of which include current events, science, arts and entertainment, and travel.

4.2 The MMAX Annotation Tool

The manual annotation of the training data was done with the MMAX (**M**ulti-**M**odal **A**nnotation in **X**ML) annotation tool (Müller and Strube, 2001). The fist step of coreference annotation is to identify the markables, i.e. noun phrases that refer to real-word entities. Each markable is annotated with the following attributes:

- **np_form**: proper noun, definite NP, indefinite NP, personal pronoun, possessive pronoun, or demonstrative pronoun.
- **grammatical_role**: subject, object (direct or indirect), or other.
- **agreement**: this attribute is a combination of person, number and gender. The possible values are 1s, 1p, 2s, 2p, 3m, 3f, 3n, 3p.
- **semantic_class**: human, physical object (includes animals), or abstract. When the semantic class is ambiguous, the ”abstract” option is chosen.
- **type**: if the entity that the markable refers to is new to the discourse, the value is ”none”. If the markable refers to an already mentioned entity, the value is ”anaphoric”. An anaphoric markable has another attribute for its relation to the antecedent. The values for this attribute are ”direct”, ”pronominal”, and ”ISA” (hyponym-hyperonym).

To mark coreference, MMAX uses coreference sets, such that every new reference to an already mentioned entity is added to the set of that entity. Implicitly, there is a set for every entity in the discourse - if an entity occurs only once, its set contains one markable.

4.3 Feature Vector

The features used by our system are summarised in Table 4.3. The individual features for anaphor

Feature	Description
pron	the pronoun
ana_npform	NP form of the anaphor
ana_gramrole	grammatical role of the anaphor
ana_agr	agreement of the anaphor
ana_semclass*	semantic class of the anaphor
ante_npform	NP form of the antecedent
ante_gramrole	grammatical role of the antecedent
ante_agr	agreement of the antecedent
ante_semclass*	semantic class of the antecedent
dist	distance in markables between anaphor and antecedent (1 .. 20)
same_agr	same agreement of anaphor and antecedent?
same_gramrole	same grammatical role of anaphor and antecedent?
same_semclass*	same semantic class of anaphor and antecedent?

Table 1: Features used by our system. *-ed features were only used for 10-fold cross-validation on the manually annotated data

and antecedent - NP form, grammatical role, semantic class - are extracted directly from the annotation. The relational features are generated by comparing the individual ones. The binary target function - coreferent, non-coreferent - is determined by comparing the values of the member attribute. If both markables are members of the same set, they are coreferent, otherwise they are not.

Due to lack of resources, the semantic class attribute cannot be annotated automatically, and is therefore used only for comparison with (Strube et al., 2002).

4.4 Noun Phrase Chunking, NER and POS-Tagging

To identify markables automatically, the system uses the noun phrase chunker described in (Schmid and Schulte im Walde, 2000), which displays case information along with the chunks. The chunker is based on a head-lexicalised probabilistic context free grammar (H-L PCFG) and achieves an F-measure of 92 for range only and 83 for range and label, whereby a range of a noun chunk is defined as "all words from the beginning of the noun phrase to the head noun". This is different from manually annotated markables, which can be complex noun phrases.

Despite good overall performance, the chunker fails on multi-word proper names in which case it marks each word as an individual chunk.[1] Since many pronouns refer to named entities, the chunker needs to be supplemented by a named entity recogniser. Although, to our knowledge, there currently does not exist an off-the-shelf named entity recogniser for German, we were able to obtain the system submitted by (Curran and Clark, 2003) to the 2003 CoNLL competition. In order to run the recogniser, the data needs to be tokenised, tagged and lemmatised, all of which is done by the TreeTagger (Schmid, 1995).

4.5 Markable Creation

After the markables are identified, they are automatically annotated with the attributes described in Section 4.4. The NP form can be reliably determined by examining the output of the noun chunker and the named entity recogniser. Pronouns and named entities are already labeled during chunking. The remaining markables are labelled as definite NPs if their first words are definite articles or possessive determiners, and as indefinite NPs otherwise. Grammatical role is determined by the case assigned to the markable - subject if nominative, object if accusative. Although datives and genitives can also be objects, they are more likely to be adjuncts and are therefore assigned the value "other".

For non-pronominal markables, agreement is determined by lexicon lookup of the head nouns. Number ambiguities are resolved with the help of the case information. Most proper names, except for a few common ones, do not appear in the lexicon and have to remain ambiguous. Although it is impossible to fully resolve the agreement ambiguities of pronominal markables, they can be classi-

[1] An example is [**Verteidigunsminister Donald**] [**Rumsfeld**] ([**Minister of Defense Donald**] [**Rumsfeld**]).

fied as either feminine/plural or masculine/neuter. Therefore we added two underspecified values to the agreement attribute: 3f 3p and 3m 3n. Each of these values was made to agree with both of its subvalues.

4.6 Antecedent Selection

After classification, one non-pronominal antecedent has to be found for each pronoun. As BoosTexter assigns confidence weights to its predictions, we have a choice between selecting the antecedent closest to the anaphor (closest-first) and the one with the highest weight (best-first). Furthermore, we have a choice between ignoring pronominal antecedents (and risking to discard all the correct antecedents within the window) and resolving them (and risking multiplication of errors). In case all of the instances within the window have been classified as non-coreferent, we choose the negative instance with the lowest weight as the antecedent. The following section presents the results for each of the selection strategies.

5 Evaluation

Before evaluating the actual system, we compared the performance of boosting to that of C4.5, as reported in (Strube et al., 2002). Trained on the same corpus and evaluated with the 10-fold crossvalidation method, boosting significantly outperforms C4.5 on both personal and possessive pronouns (see Table 2). These results support the intuition that ensemble methods are superior to single classifiers.

To put the performance of our system into perspective, we established a baseline and an upper bound for the task. The baseline chooses as the antecedent the closest non-pronominal markable that agrees in number and gender with the pronoun. The upper bound is the system's performance on the manually annotated (gold standard) data without the semantic features.

For the baseline, accuracy is significantly higher for the gold standard data than for the two test sets (see Table 3). This shows that agreement is the most important feature, which, if annotated correctly, resolves almost half of the pronouns. The classification results of the gold standard data, which are much lower than the ones in Table 2 also demonstrate the importance of the semantic features. As for the test sets, while the classifier significantly outperformed the baseline for the HTC set, it did nothing for the Spiegel set. This shows the limitations of an algorithm trained on overly restricted data.

	PPER	**PPOS**
(Strube et al., 2002)	82.8	84.9
our system	87.4	86.9

Table 2: Comparison of classification performance ($F_{\beta=1}$) with (Strube et al., 2002)

Among the selection heuristics, the approach of resolving pronominal antecedents proved consistently more effective than ignoring them, while the results for the closest-first and best-first strategies were mixed. They imply, however, that the bestfirst approach should be chosen if the classifier performed above a certain threshold; otherwise the closest-first approach is safer.

Overall, the fact that 67.2 of the pronouns were correctly resolved in the automatically annotated HTC test set, while the upper bound is 82.0, validates the approach taken for this system.

6 Conclusion and Future Work

The pronoun resolution system presented in this paper performs well for unannotated text of a limited domain. While the results are encouraging considering the knowledge-poor approach, experiments with a more complex textual domain show that the system is unsuitable for wide-coverage tasks such as question answering and summarisation.

To examine whether the system would yield comparable results in unrestricted text, it needs to be trained on a more diverse and possibly larger corpus. For this purpose, Tüba-D/Z, a treebank consisting of German newswire text, is presently being annotated with coreference information. As the syntactic annotation of the treebank is richer than that of the HTC corpus, additional features may be derived from it. Experiments with Tüba-D/Z will show whether the performance achieved for the HTC test set is scalable.

For future versions of the system, it might also

	HTC-Gold	HTC-Test	Spiegel
Baseline accuracy	46.7%	30.9%	31.1%
Classification $F_{\beta=1}$ score	77.9	62.8	30.4
Best-first, ignoring pronominal ant.	82.0%	67.2%	28.3%
Best-first, resolving pronominal ant.	72.2%	49.1%	21.7%
Closest-first, ignoring pronominal ant.	82.0%	57.3%	34.4%
Closest-first, resolving pronominal ant.	72.2%	49.1%	22.8%

Table 3: Accuracy of the different selection heuristics compared with baseline accuracy and classification F-score. HTC-Gold and HTC-Test stand for manually and automatically annotated test sets, respectively.

be beneficial to use full parses instead of chunks. As most German verbs are morphologically unambiguous, an analysis of them could help disambiguate pronouns. However, due to the relatively free word order of the German language, this approach requires extensive reseach.

References

Steven Abney, Robert E. Schapire, and Yoram Singer. 1999. Boosting applied to tagging and PP attachment. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2002. Named entity extraction using AdaBoost. In *Proceedings of CoNLL-2002*, pages 167–170, Taipei, Taiwan.

James R. Curran and Stephen Clark. 2003. Language-independent NER using a maximum entropy tagger. In *Proceedings of CoNLL-2003*, pages 164–167, Edmonton, Canada.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *First International Workshop on Multiple Classifier Systems*, Lecture Notes in Computer Science, pages 1–15. Springer, New York.

Gerard Escudero, Lluís Màrquez, and German Rigau. 2000. Boosting applied to word sense disambiguation. In *Proceedings of the 12th European Conference on Machine Learning*, pages 129–141.

Joseph F. McCarthy and Wendy G. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 1050–1055, Montreal, Canada.

Thomas S. Morton. 2000. Coreference for nlp applications. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*, Hong Kong.

Christoph Müller and Michael Strube. 2001. Annotating anaphoric and bridging relations with MMAX. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, pages 90–95, Aalborg, Denmark.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 104–111, Philadelphia, PA, USA.

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

Robert E. Schapire and Yoram Singer. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

Robert E. Schapire. 2002. The boosting approach to machine learning: an overview. In *Proceedings of the MSRI Workshop on Nonlinear Estimation and Classification*.

Helmut Schmid and Sabine Schulte im Walde. 2000. Robust German noun chunking with a probabilistic context-free grammar. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pages 726–732, Saarbrücken, Germany.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

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 *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, pages 312–319, Philadelphia, PA, USA.