The Influence of Minimum Edit Distance on Reference Resolution

Michael Strube
European Media Laboratory GmbH
Villa Bosch
Schloß-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
strube@eml.villa-bosch.de

Stefan Rapp
Sony International (Europe) GmbH
Advanced Technology Center Stuttgart
Heinrich-Hertz-Straße 1
70327 Stuttgart, Germany
rapp@sony.de

Christoph Müller
European Media Laboratory GmbH
Villa Bosch
Schloß-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
mueller@eml.villa-bosch.de

Abstract

We report on experiments in reference resolution using a decision tree approach. We started with a standard feature set used in previous work, which led to moderate results. A closer examination of the performance of the features for different forms of anaphoric expressions showed good results for pronouns, moderate results for proper names, and poor results for definite noun phrases. We then included a cheap, language and domain independent feature based on the minimum edit distance between strings. This feature yielded a significant improvement for data sets consisting of definite noun phrases and proper names, respectively. When applied to the whole data set the feature produced a smaller but still significant improvement.

1 Introduction

For the automatic understanding of written or spoken natural language it is crucial to be able to identify the entities referred to by referring expressions. The most common and thus most important types of referring expressions are pronouns and definite noun phrases (NPs). Supervised machine learning algorithms have been used for pronoun resolution (Ge et al., 1998) and for the resolution of definite NPs (Aone and Bennett, 1995; McCarthy and Lehnert, 1995; Soon et al., 2001). An unsupervised approach to the resolution of definite NPs was applied by Cardie and Wagstaff (1999). However, though machine learning algorithms may deduce to make best use of a given set of features for a given problem, it is a linguistic question and a non-trivial task to identify a set of features which describe the data sufficiently.

We report on experiments in the resolution of anaphoric expressions in general, including definite noun phrases, proper names, and personal, possessive and demonstrative pronouns. Based on the work mentioned above we started with a feature set including NP-level and coreference-level features. Applied to the whole data set these features led only to moderate results. Since the NP form of the anaphor (i.e., whether the anaphoric expression is realized as pronoun, definite NP or proper name) appeared to be the most important feature, we divided the data set into several subsets based on the NP form of the anaphor. This led to the insight that the moderate performance of our system was caused by the low performance for definite NPs. We adopted a new feature based on the *minimum edit distance* (Wagner and Fischer, 1974) between anaphor and antecedent, which led to a significant improvement on definite NPs and proper names. When applied to the whole data set the feature yielded a smaller but still significant improvement.

In this paper, we first discuss features that have been found to be relevant for the task of reference resolution (Section 2). Then we describe our corpus, the corpus annotation, and the way we prepared the data for use with a binary machine learning classifier (Section 3). In Section 4 we first describe the feature set used initially and the results it produced.

We then introduce the minimum edit distance feature and give the results it yielded on different data sets.

2 Features for Reference Resolution in Previous Work

Driven by the necessity to provide robust systems for the MUC system evaluations, researchers began to look for those features which were particular important for the task of reference resolution. While most features for pronoun resolution have been described in the literature for decades, researchers only recently began to look for *robust* and *cheap* features, i.e., features which perform well over several domains and can be annotated (semi-) automatically. In the following, we describe a few earlier contributions to reference resolution with respect to the features used.

Decision tree algorithms were used for reference resolution by Aone and Bennett (1995, C4.5), McCarthy and Lehnert (1995, C4.5) and Soon et al. (2001, C5.0). This approach requires the definition of a set of features describing pairs of anaphors and their antecedents, and collecting a training corpus annotated with them. Aone and Bennett (1995), working on reference resolution in Japanese newspaper articles, use 66 features. They do not mention all of these explicitly but emphasize the features *POS-tag*, *grammatical role*, *semantic class* and *distance*. The set of semantic classes they use appears to be rather elaborated and highly domain-dependent. Aone and Bennett (1995) report that their best classifier achieved an F-measure of about 77% after training on 250 documents. They mention that it was important for the training data to contain transitive positives, i.e., all possible coreference relations within an anaphoric chain.

McCarthy and Lehnert (1995) describe a reference resolution component which they evaluated on the MUC-5 English Joint Venture corpus. They distinguish between features which focus on individual noun phrases (e.g. *Does noun phrase contain a name?*) and features which focus on the anaphoric relation (e.g. *Do both share a common NP?*). It was criticized (Soon et al., 2001) that the features used by McCarthy and Lehnert (1995) are highly idiosyncratic and applicable only to one particular domain. McCarthy and Lehnert (1995) achieved results of about 86% F-measure (evaluated according to Vilain et al. (1995)) on the MUC-5 data set. However, only a defined subset of all possible reference resolution cases was considered relevant in the MUC-5 task description, e.g., only *entity* references. For this case, the domain-dependent features may have been particularly important, making it difficult to compare the results of this approach to others working on less restricted domains.

Soon et al. (2001) use twelve features (see Table 1). Soon et al. (2001) show a part of their decision tree in which the *weak string identity* feature (i.e. identity after determiners have been removed) appears to be the most important one. They also report on the relative contribution of the features where the three features *weak string identity, alias* (which maps named entities in order to resolve dates, person names, acronyms, etc.) and *appositive* seem to cover most of the cases (the other nine features contribute only 2.3% F-measure for MUC-6 texts and 1% F-measure for MUC-7 texts). Soon et al. (2001) include all noun phrases returned by their NP identifier and report an F-measure of 62.6% for MUC-6 data and 60.4% for MUC-7 data. They only used pairs of anaphors and their *closest* antecedents as positive examples in training, but evaluated according to Vilain et al. (1995).

Cardie and Wagstaff (1999) describe an unsupervised clustering approach to noun phrase coreference resolution in which features are assigned to single noun phrases only. They use the features shown in Table 2, all of which are obtained automatically without any manual tagging. The feature *semantic class* used by Cardie and Wagstaff (1999) seems to be a domain-dependent one which can only be used for the MUC domain and similar ones. Cardie and Wagstaff (1999) report a performance of 53,6% F-measure (evaluated according to Vilain et al. (1995)).

3 Data

3.1 Text Corpus

Our corpus consists of 242 short German texts (total 36924 tokens) about sights, historic events and persons in Heidelberg. The average length is 151 to-

– distance in sentences between anaphor and antecedent
– antecedent is a pronoun?
– anaphor is a pronoun?
– weak string identity between anaphor and antecedent
– anaphor is a definite noun phrase?
– anaphor is a demonstrative pronoun?
– number agreement between anaphor and antecedent
– semantic class agreement between anaphor and antecedent
– gender agreement between anaphor and antecedent
– anaphor and antecedent are both proper names?
– an alias feature (used for proper names and acronyms)
– an appositive feature

Table 1: Features used by Soon et al.

– position (NPs are numbered sequentially)
– pronoun type (nom., acc., possessive, ambiguous)
– article (indefinite, definite, none)
– appositive (yes, no)
– number (singular, plural)
– proper name (yes, no)
– semantic class (based on WordNet: time, city, animal, human, object; based on a separate algorithm: number, money, company)
– gender (masculine, feminine, either, neuter)
– animacy (anim, inanim)

Table 2: Features used by Cardie and Wagstaff

kens. The texts were POS-tagged using *TnT* (Brants, 2000). A basic identification of markables (referring expressions, i.e. NPs) was obtained by using the NP-Chunker *Chunkie* (Skut and Brants, 1998). The POS-tagger was also used for assigning attributes like e.g. the NP form to markables. The automatic annotation was followed by a manual correction and annotation phase in which the markables were annotated with further tags (e.g. semantic class). In this phase manual coreference annotation was performed as well. In our annotation coreference is represented in terms of a *member* attribute on markables. Markables with the same value in this attribute are considered coreferring expressions. The annotation was performed by two students. The reliability of the annotations was checked using the kappa statistic (Carletta, 1996).

3.2 Data Generation

The problem of coreference resolution can easily be formulated as a binary classification: Given a pair of potential anaphor and potential antecedent, classify as positive if the antecedent is in fact the closest antecedent, and as negative otherwise. In anaphoric chains only the immediately adjacent pairs are classified as positive. We generated data suitable as input to a machine learning algorithm from our corpus using a straightforward algorithm which combined potential anaphors and their potential antecedents. We then applied the following filters to the resulting pairs: Discard an antecedent-anaphor pair

- if the anaphor is an indefinite NP,
- if one entity is embedded into the other, e.g. if the potential anaphor is the head of the potential antecedent NP (or vice versa),

- if both entities have different values in their semantic class attributes[1],
- if either entity has a value other than 3rd person singular or plural in its agreement feature,
- if both entities have different values in their agreement features[2].

For some texts, these heuristics (which were applied to the entire corpus) reduced to up to 50% the potential anaphor-antecedent pairs all of which would have been negative cases. We consider the cases discarded as irrelevant because they do not contribute any knowledge for the classifier. After application of the filters, the remaining candidate pairs were labeled as follows:

- Pairs of anaphors and their direct (i.e. closest) antecedents were labeled P. This means that each anaphoric expression produced exactly *one* positive instance.
- Pairs of anaphors and those non-antecedents which occurred *closer* to the anaphor than the direct antecedent were labeled N. The number of negative instances that each expression produced thus depended on the number of non-antecedents occurring between the anaphor and the direct antecedent (or, the beginning of the text if there was none).

Pairs of anaphors and non-antecedents which occured further away than the direct antecedent as well as pairs of anaphors and non-direct (transitive) antecedents were not considered in the data sets. This produced 242 data sets with a total of 72093 instances of potential antecedent-anaphor pairs.

4 Results

4.1 Our Features

The features for our study were selected according to three criteria:

- relevance according to previous research,
- low annotation cost and/or high reliability of automatic tagging,
- domain-independence.

We distinguish between features assigned to noun phrases and features assigned to the potential coreference relation. All features are listed in Table 3 together with their respective possible values.

The grammatical function of referring expressions has often been claimed to be an important factor for reference resolution and was therefore included (features 2 and 6). The surface realization of referring expressions seems to have an influence on coreference relations as well (features 3 and 7). Since we use a German corpus and in this language the gender and the semantic class do not necessarily coincide (i.e., objects are not necessarily neuter as they are in English) we also provide a semantic class feature (5 and 9) which captures the difference between *human*, *concrete objects*, and *abstract objects*. This basically corresponds to the gender attribute in English, for which we introduced an agreement feature (4 and 8). The feature *wdist* (10) captures the distance in words between anaphor and antecedent, while the feature *ddist* (11) does the same in terms of sentences and *mdist* (12) in terms of markables. The equivalence in grammatical function between anaphor and potential antecedent is captured in the feature *syn_par* (13), which is true if both anaphor and antecedent are subjects or both are objects, and false in the other cases. The *string_ident* feature (14) appears to be of major importance since it provides for high precision in reference resolution (it almost never fails) while the *substring_match* feature (15) could potentially provide better recall.

4.2 Baseline Results

Using the features of Table 3, we trained decision tree classifiers using C5.0, with standard settings for pre and post pruning. As several features are discrete, we allowed the algorithm to use subsets of feature values in questions such as "Is ana_npform in {PPER, PPOS, PDS}?". We also let C5.0 construct rules from the decision trees, as we found them to give superior results. In our experiments, the value

[1]This filter applies only if none of the expressions is a pronoun. Otherwise, filtering on semantic class is not possible because in a real-world setting, information about a pronoun's semantic class obviously is not available prior to its resolution.

[2]This filter applies only if the anaphor *is* a pronoun. This restriction of the filter is necessary because German allows for cases where an antecedent is referred back to by a non-pronoun anaphor which has a different grammatical gender.

	Document level features	
1.	doc_id	document number (1 . . . 250)
	NP-level features	
2.	ante_gram_func	grammatical function of antecedent (subject, object, other)
3.	ante_npform	form of antecedent (definite NP, indefinite NP, personal pronoun, demonstrative pronoun, possessive pronoun, proper name)
4.	ante_agree	agreement in person, gender, number
5.	ante_semanticclass	semantic class of antecedent (human, concrete object, abstract object)
6.	ana_gram_func	grammatical function of anaphor (subject, object, other)
7.	ana_npform	form of anaphor (definite NP, indefinite NP, personal pronoun, demonstrative pronoun, possessive pronoun, proper name)
8.	ana_agree	agreement in person, gender, number
9.	ana_semanticclass	semantic class of anaphor (human, concrete object, abstract object)
	Coreference-level features	
10.	wdist	distance between anaphor and antecedent in words (1 . . . n)
11.	ddist	distance between anaphor and antecedent in sentences (0, 1, >1)
12.	mdist	distance between anaphor and antecedent in markables (1 . . . n)
13.	syn_par	anaphor and antecedent have the same grammatical function (yes, no)
14.	string_ident	anaphor and antecedent consist of identical strings (yes, no)
15.	substring_match	one string contains the other (yes, no)

Table 3: Our Features

of the *ana_semanticclass* attribute was reset to *missing* for pronominal anaphors, because in a realistic setting the semantic class of a pronoun obviously is not available prior to its resolution.

Using 10-fold cross validation (with about 25 documents for each of the 10 bins), we achieved an overall error rate of 1.74%. Always guessing the by far more frequent negative class would give an error rate of 2.88% (70019 out of 72093 cases). The precision for finding positive cases is 88.60%, the recall is 45.32%. The equally weighted F-measure[3] is 59.97%.

Since we were not satisfied with this result we examined the performance of the features. Surprisingly, against our linguistic intuition the *ana_npform* feature appeared to be the most important one. Thus, we expected considerable differences in the performance of our classifier with respect to the NP form of the anaphor under consideration. We split the data into subsets defined by the NP form of the anaphor and trained the classifier on these data sets. The results confirmed that the classifier performed poorly on definite NPs *(defNP)* and demonstrative pronouns *(PDS)*, moderately on proper names *(NE)* and quite good on personal pronouns *(PPER)* and possessive pronouns *(PPOS)* (the results are reported in Table 4). As definite NPs account for 792 out of 2074 (38.19%) of the positive cases (and for 48125 (66.75%) of all cases), it is evident that the weak performance for the resolution of definite NPs, especially the low recall of only 8.71% clearly impairs the overall results. Demonstrative pronouns appear only in 0.87% of the positive cases, so the inferior performance is not that important. Proper names *(NE)* however are more problematic, as they have to be considered in 644 or 31.05% of the positive cases (22.96% of all).

	P	R	F
defNP	87.34%	8.71%	15.84%
NE	90.83%	50.78%	65.14%
PDS	25.00%	11.11%	15.38%
PPER	88.12%	78.07%	82.79%
PPOS	82.69%	87.31%	84.94%
all	88.60%	45.32%	59.97%

Table 4: Baseline results using features 2–15.

[3] computed as $F = 2PR/(P + R)$

Antecedent	Anaphor
“Philips”	“Kurfürst Philip”
“vier Schülern”	“die Schüler”
“die alte Universität”	“der alten Universität”
“im Studentenkarzer in der Augustinergasse”	“des Studentenkarzers”
“diese hervorragende Bibliothek”	“dieser Bibliothek”

Table 5: Anaphors and their direct antecedents

	New coreference-level features	
16.	ante_med	minimum edit distance to anaphor $ante\_med = 100 \cdot \frac{m-(s+i+d)}{m}$
17.	ana_med	minimum edit distance to antecedent $ana\_med = 100 \cdot \frac{n-(s+i+d)}{n}$

Table 6: Additional Features (m, n, s, i, d: see text)

4.3 Additional features

Since definite noun phrases constitute more than a third of the anaphoric expressions in our corpus, we investigated why the resolution performed so poorly for these cases. The major reason may be that the resolution algorithm relies on surface features and does not have access to world or domain knowledge, which we did not want to depend upon since we were mainly interested in cheap features. However, the *string_ident* and *substring_match* features did not perform very well either. The *string_ident* feature had a very high precision (it almost never failed) but a poor recall. The *substring_match* feature was not too helpful either as it does not trigger in many cases. So, we investigated ways to raise the recall of the *string_ident* and *substring_match* features without losing too much precision.

A look at some relevant cases (Table 5) suggested that a large number of anaphoric definite NPs shared some substring with their antecedent, but they were not identical nor completely included. What is needed is a weakened form of the *string_ident* and *substring_match* features. Soon et al. (2001) removed determiners before comparing the strings. Other researchers like Vieira and Poesio (2000) used information about the syntactic structure and compared only the syntactic heads of the phrases. However, the feature used by Soon et al. (2001) is neither sufficient nor language dependent, the one used by Vieira and Poesio (2000) is not cheap since it relies on a syntactic analysis.

We were looking for a feature which gave us the improvements of the features used by other researchers without their associated costs. Hence we considered the *minimum edit distance (MED)* (Wagner and Fischer, 1974), which has been used for spelling correction and in speech recognizer evaluations (termed “accuracy” there) in the past. The MED computes the similarity of strings by taking into account the minimum number of editing operations (substitutions, insertions, deletions) needed to transform one string into the other (see also Jurafsky and Martin (2000, p.153ff. and p.271)).

We included MED into our feature set by computing one value for each editing direction. Both values share the number of editing operations but they differ when anaphor and antecedent have a different length. The features *ante_med* (16) and *ana_med* (17) are computed from the number of substitutions s, insertions i, deletions d and the length of the potential antecedent m or anaphor n as in Table 6.

4.4 Improved Results

The inclusion of the MED features 16 and 17 led to a significant improvement (Table 7). The F-measure is improved to 67.98%, an improvement of about 8%. Considering the classifiers trained and tested on the data partitions according to *ana_npform*, we can see that the improvements mainly stem from *defNP* and

NE. With respect to definite NPs we gained about 18% F-measure, with respect to proper names about 11% F-measure. For pronouns, the results did not vary much.

4.5 MUC-style results

It is common practice to evaluate coreference resolution systems according to a scheme originally developed for MUC evaluation by (Vilain et al., 1995). In order to be able to apply it to our classifier, we first implemented a simple reference resolution algorithm. This algorithm incrementally processes a real text by iterating over all referring expressions. Upon encountering a possibly anaphoric expression, it moves upwards (i.e. in the direction of the beginning of the text) and submits each pair of potential anaphor and potential antecedent to a classifier trained on the features described above. For the reasons mentioned in Section 4.2, the value of the *ana_semanticclass* attribute is reset to *missing* if the potential anaphor is a pronominal form. The algorithm then selects the first (if any) pair which the classifier labels as coreferential. Once a text has been completely processed, the resulting coreference classes are evaluated by comparing them to the original annotation according to the scheme proposed by (Vilain et al., 1995). This scheme takes into account the particularities of coreference resolution by abstracting from the question if *individual pairs* of anaphors and antecedents are found. Instead, it focusses on whether *sets of coreferring expressions* are correctly identified. In contrast to the experiments reported in Section 4.2 and 4.4, our algorithm did not use a C5.0, but a J48[4] decision tree classifier, which is a Java re-implementation of C4.5. This was done for technical reasons, J48 being more easily integrated into our system. Accompanying experimentation revealed that J48's performance is only slightly inferior to that of C5.0 for our data. Again using 10-fold cross validation, we obtained the results given in Table 8.

[4]Part of the Weka machine learning library, cf. http://www.cs.waikato.ac.nz/ml/weka

	P	R	F
defNP	69.26%	22.47%	33.94%
NE	90.77%	65.68%	76.22%
PDS	25.00%	11.11%	15.38%
PPER	85.81%	77.78%	81.60%
PPOS	82.11%	87.31%	84.63%
all	84.96%	56.65%	67.98%

Table 7: Improved results using features 2–17.

5 Conclusions

In this paper we described the influence of features based on the minimum edit distance (MED) between anaphor and antecedent on reference resolution. Though previous research used several different string similarity measures, to our knowledge, the MED feature was not used in previous work on reference resolution. We showed that the feature led to a significant improvement over the standard set of features we started with. It improved the recall for definite NPs and proper names considerably without losing too much precision. Also, it did not have any negative effect on pronouns. The MED feature is easy to compute and language and domain independent. In contrast, features used in previous work were either language dependent (e.g. the *weak string identity* feature as used by Soon et al. (2001)), domain dependent (their *alias* feature or similar features used by Cardie and Wagstaff (1999)), or relied on information on the syntactic structure (Vieira and Poesio, 2000). We consider the MED feature as a generalization of these features. It is more abstract than the features used by other researchers but delivers similar information.

We showed that our approach performs very well for personal and possessive pronouns and for proper names. For definite NPs, although they benefit from the MED features as well, there is still much room for improvement. We are curious to investigate further "cheap" features and compare them to what could be obtained when taking domain or world knowledge into account.

Features	P	R	F
2 – 15	81.31%	47.44%	59.92%
2 – 17	80.17%	55.14%	65.34%

Table 8: MUC-style results with different features.

Acknowledgments. The work presented here has been partially funded by the German Ministry of Research and Technology as part of the EMBASSI project (01 IL 904 D/2, 01 IL 904 S 8), by Sony International (Europe) GmbH and by the Klaus Tschira Foundation. We would like to thank our annotators Anna Björk Nikulásdôttir, Berenike Loos and Lutz Wind.

References

Chinatsu Aone and Scott W. Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, Mass., 26–30 June 1995, pages 122–129.

Thorsten Brants. 2000. TnT – A statistical Part-of-Speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, Seattle, Wash., 29 April – 4 May 2000, pages 224–231.

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, Md., 21–22 June 1999, pages 82–89.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, Montréal, Canada, pages 161–170.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, N.J.

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*, Montréal, Canada, 1995, pages 1050–1055.

Wojciech Skut and Thorsten Brants. 1998. A maximum-entropy partial parser for unrestricted text. In *6th Workshop on Very Large Corpora*, Montreal, Canada, pages 143–151.

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.

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

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.

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173.