
Adding Semantic Information: Unsupervised Clusters for Coreference Resolution

Iris Hendrickx
Walter Daelemans

IRIS.HENDRICKX@UA.AC.BE
WALTER.DAELEMANS@UA.AC.BE

CNTS - Language Technology Group, University of Antwerp, universiteitsplein 1, Antwerp, Belgium

Abstract

We evaluate the effect of automatically generated semantic clusters as an information source in our machine learning approach to the task of coreference resolution for Dutch. We compare these clusters which group semantically similar nouns together, to two semantic features based on WordNet encoding synonym and hypernym relations between nouns. Our experiments with two learners show that the cluster-based features lead to a small improvement for memory-based learning, while combining both leads to an improvement for maximum entropy modeling.

1. Introduction

Coreference resolution is the task of resolving different descriptions in the text to the same underlying entity. Resolving ambiguous referents in a text can be a helpful preprocessing step for many NLP applications such as text summarization or question answering.

We view coreference resolution of noun phrases (NPs) as a classification task that can be solved with supervised machine learning. This approach requires a corpus annotated with coreferential links between NPs. Next, instances are created between every NP (candidate anaphor) and all of its preceding NPs (candidate antecedents). The task of the classifier is to label each pair of NPs as coreferential or not.

In this study we focus on a particular semantic information source, namely automatically generated semantic clusters (Van de Cruys, 2005) to model the semantic classes of NPs. We study the effect of using this information and we compare its effect to the use of two other semantic features based on Wordnet.

Proceedings of the 18th Benelx
P. Adriaans, M. van Someren, S. Katrenko (eds.)
Copyright © 2007, The Author(s)

In Section 2 we discuss these two types of semantic features, Section 3 describes the experimental setup, results and conclusions are presented in Section 4 and 5.

2. Semantic information sources

Semantic information can be an important source to determine whether two referents point to the same entity. For Dutch there are few sources available to obtain semantic knowledge about words. One well-known source is the Dutch part of EuroWordNet (Vossen, 1998), a multilingual lexical database.

In our approach we use WordNet to construct two binary features *is_synonym* and *is_hypernym* to code for every pair of referents whether their descriptions can be found in WordNet in some synonym or hypernym relation¹.

As a second source we use semantic clusters (Van de Cruys, 2005). These clusters were extracted with unsupervised k-means clustering on the Twente Nieuws Corpus. The corpus was first preprocessed by the Alpino parser to extract syntactic relations. The top-10,000 lemmatized nouns (including names) were clustered into a 1000 groups based on the similarity of their syntactic relations. This is an example of two clusters:

- { Disney MGM Paramount PolyGram Time_Warner Turner Viacom Walt_Disney }
- { barrire belemmering drempel hindernis hobbelpunt knelpunt obstakel struikelblok }

For each pair of referents we constructed three features as follows. For each referent the lemma of the head word is looked up in the list of clusters. The number of the matching cluster, or 0 in case of no match, is used as the feature value. We constructed two features (*clust1*, *clust2*) presenting the cluster number of

¹Two referents with complete a string match are also considered as synonyms and hypernyms.

each referent and a binary feature marking whether the head words of the referents occur in the same cluster (*same_clust*). Table 1 shows the percentages of instances in which a particular semantic feature has a non-zero value.

Table 1. Percentage of instances in which each semantic feature is active

FEATURE	% INST
IS_SYN	6.5
IS_HYP	6.6
CLUST1	59.5
CLUST2	58.4
SAME_CLUSTER	2.1

3. Experimental setup

We use a Dutch corpus of Flemish news articles, KNACK-2002, annotated with coreference information for NPs (Hoste, 2005).

We created instances between every NP (candidate anaphor) and all of its preceding NPs (candidate antecedent). Sometimes, the search scope is limited to 3 sentences through the application of distance restrictions or linguistically motivated filters. Instances describe the relation between a potential anaphor and its antecedent and are labeled positive when the NPS are coreferential and negative otherwise. For each NP pair we create a feature set encoding morphological-lexical, syntactic, semantic, string matching and positional information. Details can be found in (Hoste, 2005).

We ran ten-fold cross validation experiments using 242 documents of KNACK-2002. We tried two different machine learning algorithms, memory-based learning (Timbl (Daelemans & van den Bosch, 2005)) and maximum entropy modeling (Maxent(Le, 2004)).

4. Results

Table 2 presents the micro-averaged Fscores of Timbl and Maxent in the ten-fold cross validation experiments with four different feature sets varying the presence/absence of the WordNet- and cluster-based features. For Timbl the WordNet features only show a marginal effect while the cluster-based features do show a small improvement. Combining both features does not really have any effect compared to using the cluster-based features. For Maxent using the WordNet features or the cluster-based features gives a marginal improvement. Combining both features has a stronger

Table 2. Average Fscore in 10-fold CV experiments. WN presents the two WordNet-based features, CLUST presents the three cluster-based features

	TIMBL	MAXENT
-WN, -CLUST	46.77	43.15
+WN, -CLUST	46.79	43.41
-WN, +CLUST	47.38	43.28
+WN, +CLUST	47.36	44.35

effect and improves the Fscore of Maxent with 1%.

5. Conclusions

We evaluated the effect of the cluster-based features on performance and compare it to the effect of two WordNet-based features. Our experiments showed that the WordNet features do not seem to improve the performance of Timbl while the cluster-based features do give a small positive effect. For Maxent the WordNet- or cluster-based features separately only have a marginal effect but combining both features gives a positive effect on performance.

6. Acknowledgments

We would like to thank Tim Van de Cruys for kindly sharing his data sets of semantic clusters.

References

- Daelemans, W., & van den Bosch, A. (2005). *Memory-based Language Processing*. Cambridge University Press.
- Hoste, V. (2005). *Optimization issues in machine learning of coreference resolution*. Doctoral dissertation, Antwerp University.
- Le, Z. (2004). *Maximum entropy modeling toolkit for python and c++ (version 20041229)*. Natural Language Processing Lab, Northeastern University, China.
- Van de Cruys, T. (2005). Semantic clustering in dutch. *Proceedings of the Sixteenth Computational Linguistics in the Netherlands (CLIN)* (pp. 17–32).
- Vossen, P. (Ed.). (1998). *Eurowordnet: a multilingual database with lexical semantic networks*. Norwell, MA, USA: Kluwer Academic Publishers.