Semantic and Syntactic Features for Dutch Coreference Resolution

Iris Hendrickx[1], Veronique Hoste[2], and Walter Daelemans[1]

[1] CNTS - Language Technology Group,
University of Antwerp, prinsstraat 13, Antwerp
Belgium
iris.hendrickx@ua.ac.be, walter.daelemans@ua.ac.be
[2] LT3 - Language and Translation Technology Team,
University College Ghent, Groot-Brittaniëlaan 45, Ghent,
Belgium
veronique.hoste@hogent.be

Abstract. We investigate the effect of encoding additional semantic and syntactic information sources in a classification-based machine learning approach to the task of coreference resolution for Dutch. We experiment both with a memory-based learning approach and a maximum entropy modeling method.

As an alternative to using external lexical resources, such as the low-coverage Dutch EuroWordNet, we evaluate the effect of automatically generated semantic clusters as information source. We compare these clusters, which group together semantically similar nouns, to two semantic features based on EuroWordNet encoding synonym and hypernym relations between nouns.

The syntactic function of the anaphor and antecedent in the sentence can be an important clue for resolving coreferential relations. As baseline approach, we encode syntactic information as predicted by a memory-based shallow parser in a set of features. We contrast these shallow parse based features with features encoding richer syntactic information from a dependency parser. We show that using both the additional semantic information and syntactic information lead to small but significant performance improvement of our coreference resolution approach.

1 Introduction

Coreference resolution is the task of resolving different descriptions of the same underlying entity in a given text. Written and spoken texts contain a large number of coreferential relations and a good text understanding largely depends on the correct resolution of these relations. Resolving ambiguous referents in a text can be a helpful preprocessing step for many NLP applications such as text summarization or question answering.

As an alternative to the knowledge-based approaches, in which there has been an evolution from the systems which require an extensive amount of linguistic and non-linguistic information (e.g. [1]) toward more knowledge-poor approaches