

Summary

of the PhD thesis

Linguistic Knowledge and Word Sense Disambiguation

by Tanja Gaustad (2004)

Supervisors: John Nerbonne and Gertjan van Noord

Humanities Computing, University of Groningen, The Netherlands

The main research question we try to answer in the present thesis is which linguistic knowledge sources are most useful for word sense disambiguation (WSD), more specifically word sense disambiguation of Dutch. Therefore, the structure of the thesis is based on the various levels of linguistic information tested for WSD, including morphology, information on the syntactic class of a particular ambiguous word, and the syntactic structure of the entire sentence containing an ambiguous word. Each source of linguistic knowledge is tested and evaluated individually in order to assess its value for WSD. Finally, combinations of knowledge sources are investigated and evaluated.

The goal of our project was to develop a tool which is able to automatically determine the meaning of a particular ambiguous word in context, a so called *word sense disambiguation system*. In order to achieve this, we make use of the information contained in the context. So we use the words surrounding the ambiguous word, and additional underlying information, such as syntactic class and structure, to build a statistical language model. This model is then used to determine the meaning of examples of that particular ambiguous word in new contexts.

After a general introduction in chapter 1 to the subject of WSD and the main research questions of the thesis, chapter 2 presents an overview of prior research in WSD divided according to the possible approaches and the information sources employed by the systems presented. By approaches or strategies we refer to the primary resource of information used to extract information about the different senses of words, in contrast to information sources which refer to the type of knowledge used to find the correct senses. Evaluation is also discussed, especially the SENSEVAL WSD evaluation framework. The general approach chosen for our own work concludes the introduction and literature overview.

In chapter 3 we show that the widely used technique of pseudowords to alleviate the need for hand annotated sense-tagged data is not a viable substitute for real ambiguous words. The main reason for this is that the “senses” of pseudowords consist of two (or more) clearly distinct words whereas real ambiguous words usually have senses and subsenses that can be closely related and are therefore more difficult to identify correctly, even for humans.

Then the experimental setup of the supervised corpus-based WSD system is introduced in chapter 4, including a presentation of the corpus, the classification algorithm used for disambiguation, as well as its implementation. We also present first results on the tuning data using a leave-one-out approach with only “basic” features, such as the context surrounding the ambiguous word and its lemma. From these results, we can conclude that maximum entropy works well as a classification algorithm for WSD when compared to the frequency baseline.

The results of the various experiments with these basic features decide which settings can best be used when more kinds of linguistic knowledge are included in the system. It is investigated whether it is beneficial to use a frequency threshold with regard to the number of training instances of each ambiguous word found in the corpus. Our results show that maximum entropy (in combination with smoothing using Gaussian priors) is robust enough to deal with infrequent data and for this reason no threshold was applied. Moreover, various context sizes have been tested (only taking into account the context words contained in the same sentence as the ambiguous word). We have found that a context of three words to the right and the left perform better than bigger context sizes, confirming earlier findings in the WSD literature. The last important result from chapter 4 is that using context lemmas for generalization in combination with the relative position of the context to the ambiguous word achieves better accuracy than context words and/or treating the context as a bag of words.

After the presentation of our WSD system for Dutch and the experimental setup, chapter 5 introduces a novel approach to building classifiers and, at the same time, includes the first type of linguistic knowledge we investigated, namely morphological information. Instead of building a classifier for each individual word form (as has traditionally been done), we build classifiers on the basis of the more general lemmas. An ambiguous word is then classified on the basis of its lemma.

Lemmatization allows for more compact and generalizable data by clustering all inflected forms of an ambiguous word together. The more inflection in a language, the more lemmatization will help to compress and generalize the data. Therefore, more training material is available to each classifier and the resulting WSD system is smaller and more robust.

Our comparison of the lemma-based approach with the traditional word form-based approach on the Dutch SENSEVAL-2 test data set clearly shows that using lemmatization significantly improves accuracy. Also, in comparison to earlier results with a Memory-Based WSD system, the lemma-based approach performs equally well when using the same features. involving less work (no parameter optimization).

A second source of linguistic information that is tested for its value for WSD is part-of-speech (PoS) (chapter 6). The PoS of an ambiguous word itself presents important information because the Dutch SENSEVAL-2 data had to be disambiguated morpho-syntactically as well as with regard to meaning. Two hypotheses are tested. On the one hand, it is investigated what effect the quality of the PoS tagger used to tag the data has on the results of the WSD system including PoS information. The results confirm the expectation that the most accurate PoS tagger (on a stand-alone task) also outperforms less accurate taggers in the application-oriented evaluation in our WSD system for Dutch. On the other hand, the experiments conducted allow us to test whether adding features explicitly encoding certain types of knowledge increases disambiguation accuracy. Our results show that this is definitely the case.

We not only include the PoS of the ambiguous words, but also add the PoS of the context as an extra feature. Both sources of knowledge lead to significant improvements in the performance of the maximum entropy WSD system.

The third kind of information, and second kind of syntactic knowledge, that is included are dependency relations (described in chapter 7). This implicitly tests whether deep linguistic knowledge is beneficial for a WSD application. After an overview of previous research in WSD using syntactic information, we introduce dependency relations and their merit for NLP, as well as Alpino, the dependency parser which was used to annotate the data. Two different kinds of features including dependency relations are experimented with. On the one hand, we test the configuration with two features containing the name of all relations of a given ambiguous word. One feature contains the head relations while the other feature contains the dependent relations of the ambiguous word. On the other hand, we test the configuration with the same two features but this time combining the name of the relation with the word completing the dependency triple.

The results in chapter 7 show that the addition of deep linguistic knowledge to a statistical WSD system for Dutch results in a significant rise in disambiguation accuracy compared with all results on the tuning data discussed so far. Dependency relations on their own already perform significantly better than the baseline, the combination of the lemma and PoS of the ambiguous word together with dependency relations even outperforming the model using context information. The best results (on the tuning data) at 86.08% are achieved including the lemma, the PoS as well as the dependency relations linked to the ambiguous words in combination with the context lemmas.

In chapter 8 we report our results on the (unseen) SENSEVAL-2 test data

with the best feature models determined during tuning. Several conclusions can be drawn from the experiments conducted on the test data. First of all, adding structural syntactic information in the form of dependency relations instead of PoS of the context leads to an error-rate reduction of 8% for the word form model. Furthermore, the lemma-based approach outperforms the word form-based approach independently of the features included in the model. The best overall performance on the test data is achieved using the lemma-based approach with the feature model including information on the PoS of the ambiguous word form/lemma, its dependency relation labels, as well as the context lemmas. We can observe an error rate reduction of 10% with regard to the lemma-based model including PoS in context, and a reduction of 6% of errors with regard to the best model based on word forms.

Comparing our results on the test data to results obtained with a different system, using Memory-Based Learning (MBL) as a classification algorithm (Hendrickx et al., 2002), both the word form-based classifiers and the lemma-based classifiers from our system produce higher accuracy. This is mainly due to the fact that our feature model includes deep linguistic information in the form of dependency relations whereas Hendrickx et al. include PoS of the context. The lemma-based model actually leads to an error rate reduction of 10% if compared to the MBL WSD system. Our maximum entropy system is thus state-of-the-art for Dutch word sense disambiguation, showing that the combination of building classifiers based on lemmas instead of word forms and including dependency relation labels as linguistic features (along with context lemmas) works best.

As a general conclusion, the results from our research suggest that in the case of a statistical disambiguation algorithm the *combination* of several orthogonal linguistic features yields the best results. This means that WSD for Dutch profits from various sources of linguistic knowledge. Thus, there is not a single best linguistic knowledge source, but rather a number of (carefully) selected features that work best in combination.

Especially the addition of deep linguistic knowledge greatly improves accuracy. In combination with an approach taking advantage of morphological information, the lemma-based approach, the best results for WSD of Dutch on the SENSEVAL-2 data set are obtained. Our system achieves significantly higher disambiguation accuracy than any results for Dutch that have been reported in the literature up to now.

References

- Hendrickx, I., van den Bosch, A., Hoste, V., and Daelemans, W. (2002). Dutch word sense disambiguation: Optimizing the localness of context. In *Proceedings of the ACL 2002 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia.