Linguistic Knowledge and Word Sense Disambiguation
Gaustad, Tanja

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2004

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 8

Final Results on Dutch

Senseval-2 Test Data

The general idea of testing is to assess how well a given model works and that can only be done properly on data that has not been seen before. Supervised models often have a tendency to be overtrained, i.e. they expect future events to resemble training events and are not able to generalize well to new data. Therefore, it is essential to test on different data to determine the real performance of a system.

The main goal of the research reported on in this thesis is to test various sources of linguistic knowledge on their value for WSD, independently and in combination. Consequently, many different feature models have been tested. In order to avoid over-using the test data by testing each feature model on it, it was necessary to evaluate the different kinds of linguistic knowledge in a tuning scheme first. So far, most of our results (with the exception of the results presented in chapter 5) have been produced using the leave-one-out approach described in section 4.5. Now the best feature model in the tuning setup has been determined, we can proceed to evaluate our WSD system for Dutch on the (unseen) test data. These final results will also allow us to validate the conclusions we have drawn from the results on the tuning data and to compare the accuracy of our WSD system with other published results.

We will first summarize the findings on the tuning data presented so far (section 8.1). Then, in section 8.2, the settings we used for the test run will be explained. Moreover, the final results on the test data will be presented and evaluated with respect to earlier results, results on the training data and other WSD systems for Dutch.
8.1 Summary of Findings on Tuning Data

In chapter 3 we have shown 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 was introduced in chapter 4, including a presentation of the corpus, the classification algorithm used for disambiguation, as well as its implementation. We also presented first results with only “basic” features, such as the context surrounding the ambiguous word and its lemma. From these results, we could conclude that maximum entropy works well as a classification algorithm for WSD when compared to the frequency baseline.

We additionally ran several experiments taking into account these basic features to decide which settings could best be used when more kinds of linguistic knowledge were included. It was investigated whether it was 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 introduced a novel approach to building classifiers and, at the same time, included the first type of linguistic knowledge we investigated, namely morphological information. The lemma-based approach uses the advantage of more concise and more generalizable information contained in lemmas as key feature: classifiers for individual ambiguous items are built on the basis of their lemmas, instead of word forms as has traditionally been done. 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
8.1. Summary of Findings on Tuning Data

Each classifier and the resulting WSD system is smaller and more robust.

Our comparison of the lemma-based approach with the traditional word form approach on the Dutch Senseval-2 test data set clearly showed 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 has been tested for its value for WSD is PoS (chapter 6). The PoS of an ambiguous word itself presented important information because the Dutch Senseval-2 data had to be disambiguated morpho-syntactically as well as with regard to meaning. Two hypotheses were tested. On the one hand, it was investigated what effect the quality of the PoS tagger used to tag the data had on the results of the WSD system including PoS information. The results confirmed 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 allowed us to test whether adding features explicitly encoding certain types of knowledge increased disambiguation accuracy. Our results show that this is definitely the case.

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

The third kind of information and second kind of syntactic knowledge that has been included are dependency relations (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 introduced 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 were experimented with: on the one hand, two features containing the name of all relations of which a given ambiguous word is the head or the dependent, respectively, and, on the other hand, the same two features but with the name of the relation combined 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 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 informa-
Table 8.1: Results (in %) on the tuning and test data; † denotes a significant improvement over the model including PoS in context (to be read vertically).

<table>
<thead>
<tr>
<th>Data section</th>
<th>ambiguous tune</th>
<th>test</th>
<th>all tune</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>75.64</td>
<td>78.47</td>
<td>86.15</td>
<td>89.44</td>
</tr>
<tr>
<td>lemma, pos, con. lemmas, pos in con.</td>
<td>84.36</td>
<td>83.66</td>
<td>91.10</td>
<td>92.37</td>
</tr>
<tr>
<td>lemma, pos, head, dep, con. lemmas</td>
<td>86.08†</td>
<td>84.78†</td>
<td>92.07†</td>
<td>93.18†</td>
</tr>
</tbody>
</table>

8.2 Results and Evaluation

As we have already mentioned above, the combination of (carefully selected) linguistic features performs best on the tuning data. Especially deep linguistic knowledge in the form of dependency relation names significantly increases accuracy. The results on the Senseval test data for Dutch were therefore computed using the lemma of the ambiguous word, its PoS, dependency relation names, as well as the context lemmas as features. Following the results presented in section 4.6, we did not use a threshold on the number of training instances of a particular ambiguous word and kept an ordered context of three words to the left and to the right of an ambiguous word.

The results in table 8.1 confirm our findings on the tuning data that maximum entropy classification works well for WSD of Dutch and significantly outperforms the frequency baseline. The results with the best settings determined during tuning also work best on the test data in comparison to the settings used in chapter 5. An error-rate reduction of 8% can be observed when adding structural syntactic information in the form of dependency relations instead of PoS of the context.

If we compare our results on the tuning and on the test data (also in table 8.1), several aspects are worth mentioning. During training, 953 classifiers were built. There are 512 unique ambiguous words in the test data and for 410 of them training data and a trained classifier exist. This means that 102 unique ambiguous words were seen for the first time and where assigned a random guess. In the case of our tuning data, all tested instances had a trained classifier since we used a leave-one-out approach. Based on this fact, the results on the test data are expected to be lower than the results on the
8.2. Results and Evaluation

<table>
<thead>
<tr>
<th>Data Approach</th>
<th>ambiguous word</th>
<th>ambiguous lemma</th>
<th>all word</th>
<th>all lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline test data</td>
<td>78.47</td>
<td>89.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lemma, pos, con. lemmas, pos in con.</td>
<td>83.66</td>
<td>84.15†</td>
<td>92.37</td>
<td>92.45†</td>
</tr>
<tr>
<td>lemma, pos, head, dep., con. lemmas</td>
<td>84.78</td>
<td>85.74†</td>
<td>92.81</td>
<td>93.37†</td>
</tr>
</tbody>
</table>

Table 8.2: Comparison of results (in %) on the test section of the Dutch Senseval-2 data with the word form and the lemma-based approach; † denotes a significant improvement over the word form approach.

Notwithstanding words without trained classifier, the ranking of the accuracy achieved with the different feature settings remains the same on the test data: deep linguistic knowledge still outperforms shallower PoS information of the context. Due to the fact that fewer instances are being classified during testing, the difference in performance between the feature models is also less pronounced (1.12%) than during leave-one-out tuning on the training data (1.72%). The difference between the two feature models on the test and tuning data is statistically significant, however, using a paired sign test with a confidence level of 95%.

In chapter 5, we have presented results on the test data using the lemma-based approach introduced in the same chapter. We will now also give new results with the lemma-based approach using the feature model which worked best with the word form-based approach, namely including dependency relations. This allows us to verify whether the lemma-based approach outperforms classifiers built on the basis of word forms (with the same settings). The results on the test data presented in chapter 5 will be repeated, as well, which enables us to compare the lemma-based approach on two different feature models.

As we can see in table 8.2, the lemma-based approach outperforms the word form-based approach independently of the features included in the model. Also, 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.

In table 8.3, we present a comparison with another system for Dutch already described in section 5.5. The results published in Hendrickx et al.
Table 8.3: Comparison of results (in %) from different systems on the test section of the Dutch Senseval-2 data.

<table>
<thead>
<tr>
<th></th>
<th>ambiguous</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline on test data</td>
<td>78.5</td>
<td>89.4</td>
</tr>
<tr>
<td>word form-based classifiers</td>
<td>84.8</td>
<td>92.8</td>
</tr>
<tr>
<td>lemma-based classifiers</td>
<td>85.7</td>
<td>93.4</td>
</tr>
<tr>
<td>(Hendrickx et al., 2002)</td>
<td>84.0</td>
<td>92.5</td>
</tr>
</tbody>
</table>

As we can see, both the word form-based classifiers and the lemma-based classifiers produce higher accuracy than the results from the MBL system by Hendrickx et al. (2002). We think that this is mainly due to the fact that our feature model includes deep linguistic information in the form of dependency relations whereas they include PoS of the context. From our comparison of results on the tuning data (see chapter 7), we conclude that dependency relations provide more useful clues for disambiguation than PoS of the context surrounding an ambiguous word. The same conclusion can be drawn with regard to the results on the test data and even with regard to two different ML algorithms, namely maximum entropy and MBL.

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.

To conclude, it is important to mention that what has already been shown in the leave-one-out tuning setup can also be found in the results on the test data. Clearly a combination of different sources of linguistic knowledge leads to the best results. Especially the addition of deep linguistic knowledge in the form of dependency relations in combination with building classifiers based on lemmas instead of word forms greatly improves accuracy over earlier results. Future work will include a more detailed error analysis of our results with dependency relations as features with respect to e.g. trends in accuracy on different PoS or ambiguous words with (very) skewed sense distributions.

\(^1\)Unfortunately, there has been very little interest in Dutch WSD and therefore very little results to compare our approach to.