Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia, July 2002, pp. 61-66. Association for Computational Linguistics.

Dutch Word Sense Disambiguation: Optimizing the Localness of Context

Véronique Hoste and **Walter Daelemans**
CNTS - Language Technology Group
University of Antwerp, Belgium
{hoste,daelem}@uia.ua.ac.be

Iris Hendrickx and **Antal van den Bosch**
ILK Computational Linguistics
Tilburg University, The Netherlands
{I.H.E.Hendrickx,antalb}@kub.nl

Abstract

We describe a new version of the Dutch word sense disambiguation system trained and tested on a corrected version of the SENSEVAL-2 data. The system is an ensemble of word experts; each word expert is a memory-based classifier of which the parameters are automatically determined through cross-validation on training material. The original best-performing system, which used only local context features for disambiguation, is further refined by performing additional parallel cross-validation experiments for optimizing algorithmic parameters and the amount of local context available to each of the word experts' memory-based kernels. This procedure produces an accuracy of 84.8% on test material, improving on a baseline score of 77.2% and the previous SENSEVAL-2 score of 84.2%. We show that cross-validation overfits; had the local context been held constant at two left and right neighbouring words, the system would have scored 85.0%.

1 Introduction

Solving lexical ambiguity, or word sense disambiguation (WSD), is an important task in Natural Language Processing systems (Kilgarriff and Palmer, 2000). Much like syntactic word-class disambiguation, it is not a end in itself, but rather a subtask of other natural language processing tasks. The problem is far from solved, and research and competition in the development of WSD systems in isolation remains meritable, preferrably on many different languages and genres.

This paper describes a refinement of an existing all-words WSD system for Dutch (Hoste et al., 2002b) that is an ensemble of word experts, each specialised in disambiguating the senses for one particular ambiguous wordform. Each word expert has a memory-based classification kernel. The system was developed on the basis of Dutch WSD data made available for the SENSEVAL-2 competition. The data, a collection of 102 children's books for the age range of 4 to 12, is annotated according to a non-hierarchical sense inventory that is based on a children's dictionary (for a detailed description of the data, cf. (Hendrickx and van den Bosch, 2002)).

Since SENSEVAL-2, both the data and the system have been refined. The data has been cleaned by hand to remove annotation errors. Subsequently, cross-validation experiments were performed to optimize the amount of local context around the ambiguous word, which had been set arbitrarily constant in previous studies (Veenstra et al., 2000; Hendrickx and van den Bosch, 2002; Hoste et al., 2002a). Cross-validation focused on local context as opposed to non-local context (e.g. keyword features), since a post-SENSEVAL-2 study described in (Hoste et al., 2002b) indicated that for the Dutch data, WSD on local context, the immediate three left and right neighbouring words of the ambiguous words, yielded the best performance among all variants tested. Local context alone proved to be better than keyword vector representations of the wider

textual context, and better than classifier combination schemes.

The paper is structured as follows. First, in Section 2 we briefly review the Dutch WSD system and the data it is based on. Section 3 describes the new cross-validation experiments that focus on optimising the amount of local context per word expert. Section 4 discusses the new results and puts them in perspective of related studies.

2 The Dutch WSD system: Algorithms, data, instance generation

The memory-based WSD system for Dutch, henceforth referred to as MBWSD-D, is built from the viewpoint of WSD as a classification task. Given an ambiguous word and its context as input features, a data-trained classifier assigns the contextually correct class (sense) to it. Our approach to memory-based all-words WSD follows the memory-based approach of (Ng and Lee, 1996), and the work by (Veenstra et al., 2000) on a memory-based approach to the English lexical sample task of SENSEVAL-1. We borrow the classification-based approach, and the word-expert concept of the latter: for each wordform, a word expert classifier is trained on disambiguating its one particular wordform.

In this section we give an overview of the learning algorithms used, the data, and how this data was converted into instances of ambiguous words in context, to make the WSD task learnable for the memory-based word experts.

2.1 Learning algorithms

The distinguishing feature of memory-based learning (MBL) in contrast with minimal-description-length-driven or "eager" ML algorithms is that MBL keeps all training data in memory, and only abstracts at classification time by extrapolating a class from the most similar item(s) in memory to the new test item. This strategy is often referred to as "lazy" learning. In recent work (Daelemans et al., 1999) we have shown that for typical natural language processing tasks, this lazy learning approach performs well because it allows extrapolation from low-frequency or exceptional cases, whereas eager methods tend to treat these as discardable noise. Also, the automatic feature weighting in the similarity metric of a memory-based learner makes the approach well-suited for domains with large numbers of features from heterogeneous sources, as it embodies a smoothing-by-similarity method when data is sparse (Zavrel and Daelemans, 1997). For our experiments, we used the MBL algorithms implemented in TIMBL[1]. We give a brief overview of the algorithms and metrics here, and refer to (Daelemans et al., 1997; Daelemans et al., 2001) for more information.

IB1 – The distance between a test item and each memory item is defined as the number of features for which they have a different value (Aha et al., 1991). Classification occurs via the *k-nearest-distances* rule: all memory items which are equally near at the nearest k distances surrounding the test item are taken into account in classification. The classification assigned to the test item is simply the majority class among the memory items at the k nearest distances.

Feature-weighted IB1 – In most cases, not all features are equally relevant for solving the task; different types of weighting are available in TIMBL to assign differential cost to a feature value mismatch during comparison. Some of these are information-theoretic (based on measuring the reduction of uncertainty about the class to be predicted when knowing the value of a feature): information gain and gain ratio. Others are statistical (based on comparing expected and observed frequencies of value-class associations): chi-squared and shared variance.

Distance-weighted IB1 – Instead of simply taking the majority class among all memory items in the k nearest distances, the class vote of each memory item is weighted by its distance. The more distant a memory item is to the test item, the lower its class vote is. This can be implemented by using several mathematical functions; the TIMBL software implements linear inversed distance weights, inversed distance weights, and exponentially decayed distance weights.

[1] Available from http://ilk.kub.nl

Value-difference weighted IB1 – For typical symbolic (nominal) features, values are not ordered. In the previous variants, mismatches between values are all interpreted as equally important, regardless of how similar (in terms of classification behaviour) the values are. We adopted the *modified value difference metric* (Cost and Salzberg, 1993) to assign a different distance between each pair of values of the same feature. This algorithm can also be combined with the different feature weighting methods.

2.2 Data

The Dutch WSD corpus was built as a part of a sociolinguistic project, led by Walter Schrooten and Anne Vermeer (1994), on the active vocabulary of children in the age of 4 to 12 in the Netherlands. The aim of developing the corpus was to have a realistic wordlist of the most common words used at elementary schools. This wordlist was further used in the study to make literacy tests, including tests how many senses of ambiguous words were known by children of different ages. The corpus consists of texts of 102 illustrated children books in the age range of 4 to 12. Each word in these texts is manually annotated with its appropriate sense. The data was annotated by six persons who all processed a different part of the data.

Each word in the dataset has a non-hierarchical, symbolic sense tag, realised as a mnemonic description of the specific meaning the word has in the sentence, often using a related term. As there was no gold standard sense set of Dutch available, Schrooten and Vermeer have made their own set of senses, based on a children's dictionary (Van Dale, 1996). Sense tags consist of the word's lemma and a sense description of one or two words (*berg_stapel*) or a reference of the grammatical category (*fiets_N*, *fietsen_V*). Verbs have as their tag their lemma and often a reference to their function in the sentence (*bent/zijn_kww*). When a word has only one sense, this is represented with a simple "=". Names and sound imitations also have "=" as their sense tag.

The dataset also contains senses that span over multiple words. These multi-word expressions cover idiomatic expressions, sayings, proverbs, and strong collocations. Each word in the corpus that is part of such multi-word expression has as its meaning the atomic meaning of the expression.

These are two example sentences in the corpus:

```
"/= het/het\_lidwoord raadsel/=
van/van\_prepositie de/=
verdwenen/verdwijnen regenboog/=
kan/kunnen\_mogelijkheid
alleen/alleen\_adv met/met\_prepositie
geweld/= opgelost/oplossen\_probleem
worden/worden\_hww ,"/=
zeiden/zeggen\_praten de/=
koningen/koning ./= toen/toen\_adv
verklaarden/verklaren\_oorlog ze/=
elkaar/=de/=  oorlog/= ./=
```

After SENSEVAL-2 the data was manually inspected to correct obvious annotation errors. 845 changes were made. The dataset now contains 152,728 tokens (words and punctuation tokens) from 10,258 different wordform types. 9133 of these wordform types have only one sense, leaving 1125 ambiguous wordform types.The average polysemy is 3.3 senses per wordform type and 10.7 senses per ambiguous token. The latter high number is caused by the high polysemy of high frequent prepositions which are part of many multi-word expressions. These ambiguous types account for 49.6 % (almost half) of the tokens in the corpus. As with the SENSEVAL-2 competition, the dataset was divided in two parts. The training set consists of 76 books and 114,959 tokens. The test set contains the remaining 26 books and has 37,769 tokens.

2.3 Instance generation

Instances on which the system is trained, consist only of features that are expected to give salient information about the sense of the ambiguous word. Several information sources have been suggested by the literature, such as local context of the ambiguous word, part-of-speech information and keywords.

A previous study, described in (Hoste et al., 2002b) showed that MBWSD-D trained only on local features, has a better performance on the test set than all other variants that use keyword information. In this study the local context consisted of the three neighbouring words right and left of the ambiguous word and their part-of-speech tags. It performed even better than a system that combined several classifiers, including the local classifier itself, in a voting scheme.

This suprising fact could have been caused by the

use of an ineffective keyword selection method. The keywords were selected through a selection method suggested by (Ng and Lee, 1996) within three sentences around the ambiguous word; only content words were used as candidates. So, our first step was to try two different selection methods often used for this task: information gain and loglikelihood. Although both selection methods gave better results on the training set (information gain: 86.4, log-likelihood: 86.4, local classifier: 86.1), the results on the test set (information gain: 84.1, log-likelihood: 83.9) were still not higher than the score of the local classifier (84.2).

As the use of keyword information does not seem to contribute to the Dutch WSD system, we decided to pursue optimizing the local context information. The previously used local context of three was never tested against smaller or bigger contexts, so for this study we varied the context from one word to five words left and right, plus their part-of-speech (POS) tags (i.e., we tested symmetrical contexts only). POS tags of the focus word itself are also included, to aid sense disambiguations related to syntactic differences (Stevenson and Wilks, 2001). POS tags were generated by MBT (Daelemans et al., 1996).

The following is an instance of the ambiguous word *donker* [dark] and its context "(...)zei : hmmm , het donker is ook niet zo eng(...) [said:,hmm the dark is also not so scary]":

V zei Punc : Int hmmm Punc , Art het N V is Adv ook Adv niet Adv zo Adj eng donker_duister

Instances were made for each ambiguous word, consisting of 22 features. The first ten features represent the five words left to the ambiguous focus word and their part-of-speech tags, followed by the part-of-speech tag of the focus word, in this example *N* which stands for noun. The next ten features contain the five neighbouring words and tags to the right of the focus word. The last feature shows the classification of the ambiguous word, in this case *donker_duister* [the dark].

3 Cross-validating parameters and local context

In principle, word experts should be constructed for all words with more than one sense. However, many ambiguous words occur only a few times. Word experts trained on such small amount of data may not surpass guessing the most frequent sense. In a previous experiment (Hoste et al., 2002b) it was shown that building word experts for words that occur at least ten times in the training data, yield the best results. In the training set, 484 wordforms exceeded the threshold of 10. For all words of which the frequency is lower than the threshold, the most frequent sense was predicted.

3.1 Cross-validating algorithmic parameters and local context

For each of the 484 word experts, we performed an exhaustive matrix of experiments, cross-validating on training material through 10-fold cross-validation experiments. We varied among algorithmic parameters set out in Section 2, and among local context sizes. In detail, the matrix spanned the following $10 \times 5 \times 5 \times 2 \times 2 = 1000$ variations:

- *The k parameter*, representing the number of nearest distances in which memory items are searched. In the experiments, k was varied between 1, 3, 5, 7, 9, 11, 15, 25, 35 and 45.
- *Feature weighting*: all experiments were performed without feature-weighting, and with feature-weighted IB1 using gain ratio weighting, information gain, chi-square and shared variance weighting.
- *Distance*: all experiments were performed with and without linear-inversed distance weighting.
- *Value-difference*: all experiments were performed with and without the modified value difference metric MVDM.
- *Local context size*: all experiments were performed with symmetric context widths 1 to 5, where "5" means five left and five right neighbouring words with their POS tags.

For each word expert, from these 1000 experiments the best-performing parameter setting was selected. Cross-validating on training material, the optimal accuracy of the word experts on ambiguous held-out words was 87.3%, considerably higher than the baseline of 77.0%). Subsequently, the best settings were used in a final experiment, in which all word experts were trained on all available training material and tested on the held-out test set. To further evaluate the results, described in the next section, the results were compared with a baseline score. The baseline was to select for each wordform its most frequent sense. Of the 484 wordforms for which word experts were made, 470 occured in the test set.

4 Results

The top line of Table 1 shows the mean score of all the word experts together on the test set. The score of the word experts on the test set, 84.8%, is generously higher than the baseline score of 77.2%. These are the results of the word experts only; the second row also includes the best-guess outputs for the lower-frequency words, lowering the system's performance slightly.

test selection	#words	baseline	system
word-expert words	17071	77.17	84.8
all ambiguous words	17720	76.66	84.0
all words	37769	89.04	92.5

Table 1: Summary of results on test material

We can also calculate the score on all the words in the test set, including the unambiguous words, to give an impression of the overall performance. The unambiguous words are given a score of 100%. It might be useful for a disambiguation system to tag unambiguous words with their lemma, but the kind of tagging this is not of interest in our task. The third row of Table 1 shows the results on all words in the test set.

The best context and parameter settings, determined by cross-validation for each word expert on the training set, is estimated to be the best setting for test material as well – this is a fundamental assumption of parameter cross-validation. As a post-hoc analysis, we checked the validity of this assumption. We partitioned the exhaustive matrix of experiments on all tested parameters, measuring the accuracy on test material while holding each value of the parameter constant. This means, for example, that we split the matrix of 1000 experiments per word expert into 500 experiments without the use of MVDM, and 500 experiments with MVDM. Two test scores are computed: the best settings from the first and the second 500 are used respectively (for each word expert) to determine the best parameter settings, and apply these to the test material. In other words, all parameters are optimized except MVDM, which is held constant (on or off). We performed this post-hoc test for all parameters. As it turned out, in six cases keeping the parameter constant led to (slighlty) better or equal performance as compared to the cross-validated 84.8%. Table 2 lists the six constant parameter settings. These results indicate that the parameter setting estimation by cross-validation suffers, albeit slightly, from overfitting on the training material.

cross-validated	84.8
context = 2	85.0
gain ratio	84.9
MVDM	84.8
distance weighting	84.8
k = 5	84.8
k = 11	84.8

Table 2: List of the six parameter values, along with their accuracy on test material that, held constant, equal or outperform the cross-validated test score (top).

5 Discussion

In this paper we reported on a refined version of MBWSD-D, a memory-based WSD system for Dutch. As compared to an earlier version, built on data made available to the SENSEVAL-2 competition, we have made manual corrections in the annotations of the data, and on the corrected data we have additionally cross-validated the amount of local context, which in previous work had been left arbitrarily constant at three left and right neighbouring words and their POS tags (Hendrickx and van den Bosch, 2002; Hoste et al., 2002b). Also, we did not in-

clude keyword features that were used in the mentioned studies, but were shown in those studies not to contribute to accuracy on test material. Our cross-validation experiments lead to a score on test material of 84.8%. As we have done these exeriments on a cleaned version of the data, the results described so far cannot be compared to the results described in (Hendrickx and van den Bosch, 2002), which were obtained on the previous version of the data and with different parameter optimalisations. In those experiments an optimized memory-based classifier trained only on local context of three neighbouring words right and left, achieved a score of 84.2 % on the word-expert words in the test set.

To make a comparison between the results on the old version of the data and the new version, we have conducted an experiment on the new data, using the same cross-validation procedure as we have used in (Hendrickx and van den Bosch, 2002) which led to a score of 84.3% on the test set. This shows that the cleaning of the data did not give significant better results.

Additional post-hoc analyses show that when local context is not cross-validated but held constant at two left and right neighbouring words, an accuracy of 85.0% can be obtained. This suggests that the cross-validation method has overfitted its estimations on the training material slightly; this is also witnessed by the higher cross-validated optimal accuracy on held-out training material (87.3%).

References

D. W. Aha, D. Kibler, and M. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.

S. Cost and S. Salzberg. 1993. A weighted nearest neighbour algorithm for learning with symbolic features. *Machine Learning*, 10:57–78.

W. Daelemans, J. Zavrel, and P. Berck. 1996. Part-of-speech tagging for dutch with MBT, a memory-based tagger generator. In K. van der Meer, editor, *Informatiewetenschap 1996, Wetenschappelijke bijdrage aan de Vierde Interdisciplinaire Onderzoeksconferentie Informatiewetenchap*, pages 33–40, The Netherlands. TU Delft.

W. Daelemans, A. van den Bosch, and A. Weijters. 1997. IGTree: using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review*, 11:407–423.

W. Daelemans, A. van den Bosch, and J. Zavrel. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning, Special issue on Natural Language Learning*, 34:11–41.

W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. 2001. TiMBL: Tilburg memory based learner, version 4.0, reference guide. ILK Technical Report 01-04, Tilburg University. available from http://ilk.kub.nl.

I. Hendrickx and A. van den Bosch. 2002. Dutch word sense disambiguation: Data and preliminary results. To be published in the Proceedings of the SENSEVAL-2 Workshop, Toulouse, France, 2001.

V. Hoste, W. Daelemans, I. Hendrickx, and A. van den Bosch. 2002a. Evaluating the results of a memory-based word-expert approach to unrestricted word-sense disambiguation. To be published in the Proceedings of the SENSEVAL-2 Workshop, Toulouse, France, 2001.

V. Hoste, I. Hendrickx, W. Daelemans, and A. van den Bosch. 2002b. Parameter optimization for machine-learned word sense disambiguation. To be published in Natural Language Engineering.

A. Kilgarriff and M. Palmer. 2000. Introduction to the special issue on SENSEVAL. *Computing in the Humanities*, 34(1–2):1–13.

H. T. Ng and H. B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proc. of 34th meeting of the Assiociation for Computational Linguistics*.

W. Schrooten and A. Vermeer. 1994. *Woorden in het basisonderwijs. 15.000 woorden aangeboden aan leerlingen*. TUP(Studies in meertaligheid 6).

M. Stevenson and Y. Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–351.

Van Dale. 1996. *Van Dale Basiswoordenboek van de Nederlandse taal*. Van Dale, Utrecht.

J. Veenstra, A. van den Bosch, S. Buchholz, W. Daelemans, and J. Zavrel. 2000. Memory-based word sense disambiguation. *Computers and the Humanities, special issue on Senseval, Word Sense Disambiguations*, 34(1-2).

J. Zavrel and W. Daelemans. 1997. Memory-based learning: Using similarity for smoothing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain*, Madrid.