Dutch Word Sense Disambiguation: Data and Preliminary Results

Iris Hendrickx* and **Antal van den Bosch***,+
* ILK / Computational Linguistics, Tilburg University, NL-5000 LE Tilburg, The Netherlands
+ WhizBang! Labs–Research, 4616 Henry Street, Pittsburgh PA 15213, USA

Abstract

We describe the Dutch word sense disambiguation data submitted to SENSEVAL-2, and give preliminary results on the data using a WSD system based on memory-based learning and statistical keyword selection.

1 Introduction

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

Here we introduce the first electronic Dutch word-sense annotated corpus, that was collected under a sociolinguistic research project (Schrooten and Vermeer, 1994), and was kindly donated by the team coordinators to the WSD systems community. In this paper we describe the original data and the preprocessing steps that were applied to it before submission to the SENSEVAL-2, in Section 2. We also present the first, preliminary, results obtained with MBWSD-D, the Memory-Based Word-Sense Disambiguation system for Dutch, that uses statistical keyword selection, in Section 3.

2 Data: The Dutch child book corpus

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.

Sense tags consist of the word's lemma and a sense description of one or two words (*drogen_nat*) 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 (*is/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

# tokens	152.758
# types	10.263
# sentences	12.287
# words per sentence	12.4
# unambiguous words	9.095
# words that occurs once	4.949
# sense tags	9319
# word/sense combinations occuring once	6.702
% of ambiguous tokens in corpus	54

Table 1: Basic corpus statistics

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/= ./=

The dataset needed some adaptations to make it fully usable for computational purposes. First, spelling and consistency errors have been corrected for most part, but in the data submitted to SENSEVAL-2, a certain amount of errors is still present. Second, in Dutch, prepositions are often combined with verbs as particles and these combinations have other meanings than the two separate words. Unfortunately the annotations of these cases were rather inconsistent and for that reason it was decided to give all prepositions the same sense tag *"/prepositie"* after their lemma.

The dataset consists of approximately 150,000 tokens (words and punctuation tokens) and about 10,000 different word forms. Nine thousand of these words have only one sense, leaving a thousand word types to disambiguate. These ambiguous types account for 54 % of the tokens in the corpus. The basic numbers can be found in Table 1.

For the SENSEVAL-2 competition, the dataset was divided in two parts. The training set consisted of 76 books and approximately 115.000 words. The test set consisted of the remaining 26 books and had about 38.000 words.

3 The MBWSD-D system and preliminary results

We first describe the representation of the corpus data in examples presented to a memory-based learner in Subsection 3.1. We then describe the architecture of the system in Subsection 3.2, and we then present its preliminary results in Subsection 4.

3.1 Representation: Local and keyword features

As a general idea, disambiguation information is assumed to be present in the not-too-distant context of ambiguous words; the present instantiation of MBWSD-D limits this to the sentence the ambiguous word occurs in. Sentences are not represented as is, but rather as limited sets of features expected to give salient information about which sense of the word applies.

The first source of useful disambiguation information can be found immediately adjacent to the ambiguous word.It has been found that a four-word window, two words before the target word and two words after gives good results; cf. (Veenstra et al., 2000).

Second, information about the grammatical category of the target word and its direct context words can also be valuable. Consequently, each sentence of the Dutch corpus was tagged and the part-of-speech (POS) tags of the word and its direct context (two left, two right) are included in the representation of the sentence. Part-of-speech tagging was done with the Memory Based Tagger (Daelemans et al., 1996).

Third, informative words in the context ('keywords') are detected based on the statistical chi-squared test. Chi-square estimates the significance, or degree of surprise, of the number of keyword occurrences with respect to the expected number of occurrences (apriori probability):

$$X^2 = \sum_{k=1}^{n} \frac{(f_k - e_k)^2}{e_k} \qquad (1)$$

where f_i is the keyword frequency and e_i is the expected frequency. f_i is the word frequency and e_i is the expected word frequency. The expected frequency of the keyword is given in equation 3.1. It must be noted that the Chi-Square method cannot be considered reliable when the expected frequency has a value below 5: $e_i = (f_w i / f_w) * f_k$, where f_i is the frequency the ambiguous word w of sense i, f_w is the frequency of word w and f_k is the frequency of the keyword.

The number of occurrences of a very good keyword will have a strong deviation of its expected number of occurrences divided over the senses. The expected probability with respect to all senses can be seen as a distribution of the keyword. A good keyword is a word that differs from the expected distribution and always co-occurs with a certain sense, or never co-occurs with a certain sense.

In sum, a representation of an instance of an ambiguous word consists of the two words before the target word, two words after the word, the POS tags of these words and of the target word itself, a number of selected keywords, and of course the annotated sense of the word as the class label.

3.2 System architecture

Following the example of ILK's previous word-sense disambiguation system for English (Veenstra et al., 2000), it was decided to use word experts. Berleant (Berleant, 1995) defines a word expert as follows: "A word expert is a small expert system-like module for processing a particular word based on other words in its vicinity" (1995, p.1). Word experts are common in the field of word sense disambiguation, because words are very different from each other. Words all have different numbers of senses, different frequencies and need different information sources for disambiguation. With word experts, each word can be treated with its own optimal method.

Making word experts for every ambiguous word may not be useful because many words occur only a few times in the corpus. It was decided to create word experts for wordforms with a threshold of minimal 10 occurrences in the training set. There are 524 of such words in the training set. 10 is a rather low threshold, but many words can be easily disambiguated by knowledge a single feature value, such as of their part-of-speech tag.

The software for emulating memory-based learning used in this research is TiMBL (Tilburg Memory-Based Learner). TiMBL (Daelemans et al., 2001) is a software package developed by the ILK research group at Tilburg University. TiMBL implements several memory-based classifiers. In essence, memory-based classifiers use stored classified examples to disambiguate new examples.

For each word a TiMBL word expert was trained on that portion of the training corpus that consisted of sentence representations containing that word. TiMBL was trained 300 times, each time with another combination of parameters. Each of these training sessions was evaluated with leave-one-out cross validation (Weiss and Kulikowski, 1991) to select the optimal TiMBL setting for a particular word, to be used eventually for classifying the test material.

For each word expert a total of 300 experiments were performed, each with another combination of parameter settings. In this study the following options were used (cf. (Daelemans et al., 2001) for first pointers to descriptions of these metrics and functions):

distance-weighted voting : (1) all neighbors have equal weight; (2) Inverse Distance weighting; (3) Inverse Linear weighting

feature weighting : (1) no weighting; (2) Gain Ratio; (3) Information Gain; (4) Chi Square; (5) Shared Variance

similarity metric : (1) Overlap metric; (2) MVDM

number of nearest neighbours : 1, 3, 5, 7, 9, 11, 15, 25, 45, and 75

The last step for each word expert was to test the optimal settings on the test set. To evaluate the results, described in the next Section, the results were compared with a baseline score. The baseline was to select for each word the most frequent sense.

4 Results

The top line of Table 2 shows the mean score of all the word experts together on the test set. The score of the word experts on the test set, 84.1%, is generously higher than the baseline score of 74.1%. 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.

The same results, now split on the frequency of the words in the training set, can be seen in Table 3. The first column shows the frequency groups, based on the word frequencies in the training set, the second the number of words in

test selection	#words	baseline	system
word-expert words	15365	74.1	84.1
all ambiguous words	16686	74.6	83.8
all words	37770	88.8	92.9

Table 2: Summary of results on test material

the test set, and the third column shows the mean score of the WSD system. The scores tend to get better as the frequency goes up, except for the group of 40-49, which has the lowest score of all. Note that the baseline score of the group of words with a frequency below 10 is relatively high: 80.5%.

frequency	#words	baseline	system
<10	1321	–	80.5
10-19	868	63.0	76.8
20-29	644	70.3	79.5
30-39	503	75.9	83.3
40-49	390	66.7	75.9
50-99	1873	73.7	85.4
100-199	2289	77.7	83.1
$\geq$ 200	8798	74.6	85.6
> 100	10995	75.3	85.1

Table 3: Results divided into frequency groups

We can also calculate the score on all the words in the text, including the unambiguous words, to give an impression of the overall performance. The unambiguous words are given a score of 100%, because the task was to disambiguate the ambiguous words. 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 2 shows the results on all words in which the system was applied with a threshold of 10: The system scores 4 % higher than the baseline.

5 Discussion

This paper introduced a Dutch child book corpus, generously donated to the WSD community by the team leaders of the sociolinguistic project that produced the corpus. The data is annotated with a non-hierarchical mnemonic sense inventory. The data has been cleaned up and split for the SENSEVAL-2 competition.

The data provides an arguably interesting case of a "flat" semantic tagging, where there is obviously no gain from a governing wordnet, but alternatively it is not negatively biased by an inappropriate or badly-structured wordnet either. Learnability results are therefore an interesting baseline to beat when the data would be annotated with a Dutch wordnet.

The system applied to the data as a first indication of its complexity and learnability, consisted of an ensemble of word experts trained to disambiguate particular ambiguous word forms. The score of the system on the 16686 ambiguous words in the test set was 83.8% compared to a baseline score of 74.6%. On free heldout text the system achieved a result of 92.9%; 4% over the baseline of 88.8%, or in other words yielding an error reduction of about 37%. These absolute and relative figures are roughly comparable to performances of other systems on other data, indicating at least that the data represents learnability properties typical for the WSD area.

References

D. Berleant. 1995. Engineering word-experts for word disambiguation. *Natural Language Engineering*, pages 339–362.

W. Daelemans, J. Zavrel, and P. Berck. 1996. Part-of-speech tagging of dutch with mbt, a memory-based tagger generator. In *Congresboek van de Interdisciplinaire Onderzoeksconferentie Informatiewetenchap.*

W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. 2001. Timbl: Tilburg memory based learner, version 4.0, reference guide. Technical report, Tilburg University.

A. Kilgarriff and J. Rozenzweig. 2000. Framework and results for english senseval. *Computers and the Humanities*, 34.

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

J. Veenstra, A. van den Bosch, S. Buchholz, W. Daelemans, and J. Zavrel. 2000. Memory-based word sense disambiguation. *Computers and the Humanities*, 34.

S. Weiss and C. Kulikowski. 1991. *computer systems that learn.* Morgan Kaufmann.