Evaluating the results of a memory-based word-expert approach to unrestricted word sense disambiguation.

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,Antal.vdnBosch}@kub.nl

Abstract

In this paper, we evaluate the results of the Antwerp University word sense disambiguation system in the English all words task of SENSEVAL-2. In this approach, specialized memory-based word-experts were trained per word-POS combination. Through optimization by cross-validation of the individual component classifiers and the voting scheme for combining them, the best possible word-expert was determined. In the competition, this word-expert architecture resulted in accuracies of 63.6% (fine-grained) and 64.5% (coarse-grained) on the SENSEVAL-2 test data.

In order to better understand these results, we investigated whether classifiers trained on different information sources performed differently on the different part-of-speech categories. Furthermore, the results were evaluated in terms of the available number of training items, the number of senses, and the sense distributions in the data set. We conclude that there is no information source which is optimal over all word-experts. Selecting the optimal classifier/voter for each single word-expert, however, leads to major accuracy improvements. We furthermore show that accuracies do not so much depend on the available number of training items, but largely on polysemy and sense distributions.

1 Introduction

The task of word sense disambiguation (WSD) is to assign a sense label to a word in context. Both knowledge-based and statistical methods have been applied to the problem. See (Ide and Véronis, 1998) for an introduction to the area. Recently (both SENSEVAL competitions), various machine learning (ML) approaches have been demonstrated to produce relatively successful WSD systems, e.g. memory-based learning (Ng and Lee, 1996; Veenstra et al., 2000), decision lists (Yarowsky, 2000), boosting (Escudero et al., 2000).

In this paper, we evaluate the results of a memory-based learning approach to WSD. We ask ourselves whether we can learn lessons from the errors made in the SENSEVAL-2 competition. More particularly, we are interested whether there are words or categories of words which are more difficult to predict than other words. If so, do these words have certain characteristic features? We furthermore investigate the interaction between the use of different information sources and the part-of-speech categories of the ambiguous words. We also study the relation between the accuracy of the word-experts and their number of training items, number of senses and sense distribution. For these experiments, we performed all SENSEVAL-2 experiments all over again.

In the following Section, we briefly outline the WSD architecture used in the experiments, and discuss the word-expert approach and the optimization procedure. Furthermore, a brief overview is given of the results of the different components of the word-experts on the train set and the SENSEVAL-2 test material. In Section 3, we evaluate the results of the different classifiers per part-of-speech category. In the same Section, these results are further analysed in relation to the number of training items, the number of senses and the sense distribution. Section 4 gives a detailed analysis of the results of our approach on the SENSEVAL-2 test material. We end with some concluding remarks in Section 5.

2 Memory-based word-experts

Our approach in the SENSEVAL-2 experiments was to train so-called word-experts per word-POS combination. These word-experts consist of several learning modules, each of them taking different information as input, which are furthermore combined in a voting scheme.

In the experiments, the Semcor corpus included in WordNet1.6[1] was used as train set. In the corpus, every word is linked to its appropriate sense in the WordNet lexicon. This training corpus consists of 409,990 word forms, of which 190,481 are sense-tagged. The test data in the SENSEVAL-2 English all words task consist of three articles on different topics, with at total of 2,473 words to be sense-tagged. WordNet1.7 was used for the annotation of these test data. No mapping was performed between both versions of WordNet. For both the training and the test corpus, only the word forms were used and tokenization, lemmatization and POS-tagging were done with our own software. For the part of speech tagging, the memory-based tagger MBT (Daelemans et al., 1996), trained on the Wall Street Journal corpus[2], was used. On the basis of word and POS information, lemmatization (van den Bosch and Daelemans, 1999) was done.

After this preprocessing stage, all word-experts were built. This process was guided by WordNet1.7: for every combination of a word form and a POS, WordNet1.7 was consulted to determine whether this combination had one or more possible senses. In case of only one possible sense (about 20% of the test words), the appropriate sense was assigned. In case of more possible senses, a minimal threshold of ten occurrences in the Semcor training data was determined, since 10-fold cross-validation was used for testing in all experiments. This threshold was then varied between 10 and 100 training items in order to determine the optimal number of training instances. For all words of which the frequency was lower than the threshold (also about 20% of the test words), the most frequent sense according to WordNet1.7 was predicted. The cross-validation results in Figure 2 clearly show that accuracy drops when the contribution of the baseline classifier increases. The application of the WordNet baseline classifier yields a 61.7% accuracy. The "best" graph displays the accuracy when applying the optimal classifier for each single word-expert: with a threshold of 10, a 73.8% classification accuracy is obtained. On the basis of these results, we set the threshold for the construction of a word-expert to 10 training items. For all words below this threshold, the most frequent sense according to WordNet1.7 was assigned as sense-tag. For the other words in the test set (1,404 out of 2,473), word-experts were built for each word form-POS combination, leading to 596 word-experts for the SENSEVAL-2 test data.

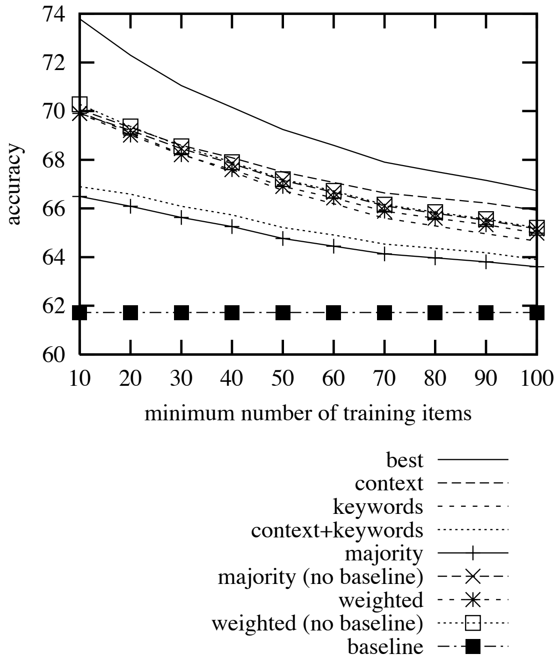

Figure 1: Accuracy of the different classifiers and voting techniques in relation to a threshold varying between 10 and 100. This accuracy is calculated on the words with more than one sense which qualify for the construction of a word-expert.

[1] Available from http://www.cogsci.princeton.edu/~wn/. Further information on WordNet can be found in Fellbaum (1998).

[2] ACL Data Collection Initiative CD-Rom 1, September 1991

The word-experts consist of different trained subcomponents which make use of different knowledge: (i) a classifier trained on the local context of the ambiguous focus word, (ii) a learner trained on keywords, (iii) a classifier trained on both of the previous information sources, (iv) a baseline classifier always providing the most frequent sense in the sense lexicon and (v) four voting strategies which vote on the outputs of the previously mentioned classifiers. For the experiments with the single classifiers, we used the MBL algorithms implemented in TIMBL[3]. In this memory-based learning approach to WSD, all instances are stored in memory during training and during testing (i.e. sense-tagging), the instance most similar (Hamming distance) to that of the focus word and its local context and/or keyword information is selected and the associated class is returned as sense-tag. For an overview of the algorithms and metrics, we refer to Daelemans et al. (2001).

- The first classifier in a word-expert takes as input a vector representing the **local context** of the focus word in a window of three words to the left and three to the right. For the focus word, both the lemma and POS are provided. For the context words, POS information is given. E.g., the following is a training instance: *American JJ history NN and CC most most JJS American JJ literature NN is VBZ most%3:00:01::.*
- The second classifier in a word-expert is trained with information about **possible disambiguating content keywords** in a context of three sentences (focus sentence and one sentence to the left and to the right). The method used to extract these keywords for each sense is based on the work of Ng and Lee (1996). In addition to the keyword information extracted from the local context of the focus word, possible disambiguating content words were also extracted from the examples in the sense definitions for a given focus word in WordNet.
- The third subcomponent is a learner combining both of the previous information sources.

In order to improve the predictions of the different learning algorithms, algorithm parameter optimization was performed where possible. Furthermore, the possible gain in accuracy of different voting strategies was explored. On the output of these three (optimized) classifiers and the WordNet1.7. most frequent sense, both majority voting and weighted voting was performed. In case of majority voting, each sense-tagger is given one vote and the tag with most votes is selected. In weighted voting, the accuracies of the taggers on the validation set are used as weights and more weight is given to the taggers with a higher accuracy. In case of ties when voting over the output of 4 classifiers, the first decision (TIMBL) was taken as output class. Voting was also performed on the output of the three classifiers without taking into account the WordNet class.

For a more complete description of this word-expert approach, we refer to (Hoste et al., 2001) and (Hoste et al., 2002).

3 Evaluation of the results

For the evaluation of our word sense disambiguation system, we concentrated on the words for which a word-expert was built. We first evaluated our approach using cross-validation on the **training data**, giving us the possiblity to evaluate over a large set (2,401) of word-experts. The results on the **test set** (596 word-experts) are discussed in Section 4.

3.1 Parts-of-speech vs. information sources

In a first evaluation step, we investigated the interaction between the use of different information sources and the part-of-speech category of the ambiguous words. Table 1 shows the results of the different component classifiers and voting mechanisms per part-of-speech category. This table shows the same tendencies among all classifiers and voters: the best scores are obtained for the adverbs, nouns and adjectives. Their average scores range between 64.2% (score of the baseline classifier on the nouns) and 76.6% (score of the context classifier on the adverbs). For the verbs, accuracies drop by nearly 10% and range between 56.9% (baseline classifier) and 64.6% (weighted voters). A similar observation was made by Kilgarriff and Rosenzweig (2000) in the SENSEVAL-1 competition in which a restricted set of words had to be disambiguated. They also showed that in English the verbs were the hardest

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

Pos	Baseline	local context	keywords	local context + keywords	majority voting	majority voting (no baseline)	weighted voting	weighted voting (no baseline)
NN	64.19	71.36	**74.20**	69.34	69.31	72.69	73.39	73.75
VB	56.87	64.33	63.82	60.09	60.84	63.55	**64.56**	64.55
JJ	66.26	72.16	**73.80**	70.39	70.37	72.79	73.34	73.61
RB	69.95	**76.64**	74.51	73.05	72.48	74.90	75.51	75.42
ALL	61.73	70.06	69.96	66.89	66.49	69.91	69.91	**70.28**

Table 1: Results on the train set of the component classifiers and voters per part-of-speech category

category to predict.

Each row in Table 1 shows results of the different word-expert components per part-of-speech category. This comparison reveals that there is no optimal classifier/voter per *part-of-speech*, nor an overall optimal classifier. However, making use of different classifiers/voters which take as input different information sources does make sense, if the selection of the classifier/voter is done at the *word* level. We already showed this gain in accuracy in Figure 2: selecting the optimal classifier/voter for each single word-expert leads to an overall accuracy of 73.8% on the train set, whereas the second best method (weighted voting without taking into account the baseline classfier) yields a 70.3% accuracy.

3.2 Number of training items

We also investigated whether the words with the same part-of-speech have certain characteristics which make them harder/easier to disambiguate. In other words, why are verbs harder to disambiguate than adverbs? For this evaluation, the results of the context classifier were taken as a test case and evaluated in terms of (i) the number of training items, (ii) the number of senses in the training corpus and (iii) the sense distribution within the word-experts.

With respect to the number of training items, we observed that their frequency distribution is Zipf-like (Zipf, 1935): many training instances only occur a limited number of times, whereas few training items occur frequently. In order to analyze the effect of the number of training items on accuracy, all word-experts were sorted according to their performance and then divided into equally-sized groups of 50. Figure 2 displays the accuracy of the word-experts in relation to their number of training items. The Figure shows that the fluctuations in accuracy are higher for the experts with a limited number of training items and that these fluctuations decrease as the number of training items increases. The average accuracy level of 70% can be situated somewhere in the middle of this fluctuating line.

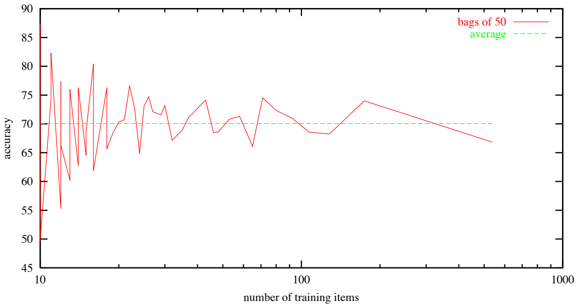

Figure 2: Number of training items over all word-experts in relation to the accuracy of the context classifier (logscale).

This tendency of performance being independent of the number of training items is also confirmed when averaging over the number of training items per part-of-speech category. The adjectives have on average 49.0 training items and the nouns have an average of 52.9 training items. The highest average number of training items is for the verbs (86.7) and adverbs (82.1). When comparing these figures with the scores in Table 1, in which it is shown that the verbs are hardest to predict, whereas the accuracy levels on the adverbs, nouns, adjectives are close, we can conclude that the mere number of training items is not an accurate predictor of accuracy. This again confirms the usefulness of training classifiers even on very small data sets, also shown in Figure 1.

3.3 Polysemy and sense distribution

For the English lexical sample task in SENSEVAL-1, Kilgarriff and Rosenzweig (2000) investigated the effect of polysemy and entropy on accuracy. Polysemy can be described as the number of senses

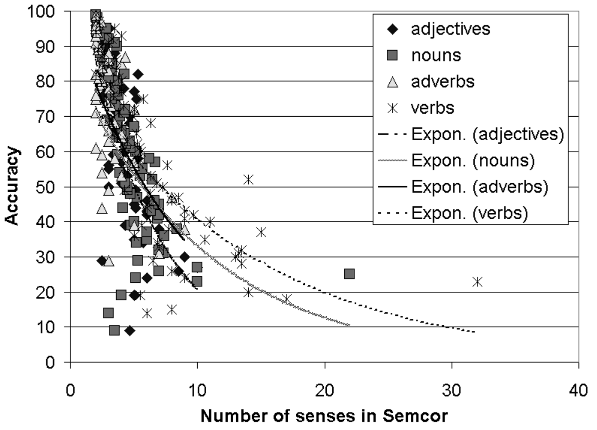

Figure 3: Scatter plot displaying the number of senses and the exponential trendline per POS in relation to the accuracy of the context classifier.

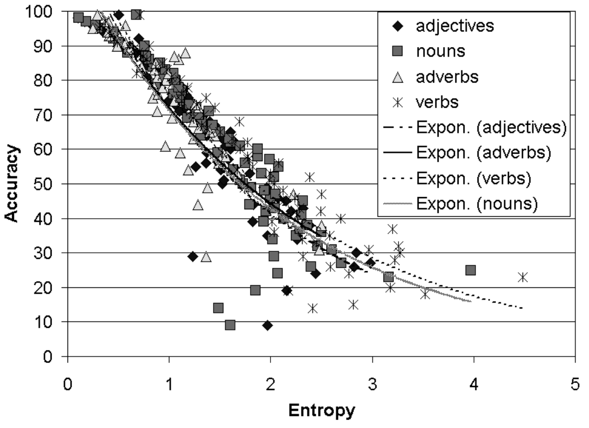

Figure 4: Scatter plot displaying the sense distributions and the exponential trendline per POS in relation to the accuracy of the context classifier.

of a word-POS combination; entropy measures the frequency distribution of the sense tags per word-POS combination. If the corpus instances are evenly spread across the lexicon senses, entropy will be high. The sense distribution of ambiguous words can also be highly skewed, giving rise to low entropy scores. Kilgarriff and Rosenzweig (2000) found that the nouns on average had higher polysemy than the verbs and the verbs had higher entropy. Since verbs were harder to predict than nouns, they came to the conclusion that entropy was a better measure of task difficulty than polysemy. Since we were interested whether the same could be concluded for the English all-words task, we investigated this effect of polysemy and entropy in relation to the accuracy of one classifier in our word-expert, namely the context classifier.

Figure 3 shows the number of senses (polysemy) over all word experts with the same part-of-speech in relation to the scores from the context classifier, whereas Figure 4 displays the sense distributions (entropy) over all word-experts with the same part-of-speech. Although it is not very clear from the scatter plot in Figure 3, the exponential trendlines show that accuracy increases as the number of senses decreases. For the sense distributions, the same tendency, but much stronger, can be observed: low entropy values mostly coincide with high accuracies, whereas high entropies lead to low accuracy scores. This tendency is also confirmed when averaging these scores over all word-experts with the same part-of-speech (see Table 2): the verbs, which are hardest to predict, are most polysemic and also show the highest entropy. The adverbs, which are easiest to predict, have on average the lowest number of senses and the lowest entropy. We can conclude that both polysemy and in particular entropy are good measures for determining task difficulty.

These results indicate it would be interesting to work towards a more coarse-grained granularity of the distinction between word senses. We believe that this would increase performance of the WSD systems and make them a possible candidate for integration in practical applications such as machine translation systems. This is also shown by Stevenson and Wilks (2001), who used the *Longman Dictionary of Contemporary English* (LDOCE) as sense inventory. In LDOCE, the senses for each word type are grouped into sets of senses with related meanings (homographs). Senses which are far enough apart are grouped into separate homographs. The vast majority of homographs in LDOCE are marked with a single part-of-speech. This makes the task of WSD partly a part-of-speech tagging task, which is generally held to be an easier task than word sense disambiguation: on a corpus of 5 articles in the *Wall Street Journal*, their system already correctly classifies 87.4% of the words when only using POS information (baseline: 78%).

As illustrated in Figure 4, the context classifier

POS	Average polysemy	Average entropy
RB	3.26±1.55	1.11±0.52
JJ	4.11±1.63	1.35±0.67
NN	4.75±2.64	1.52±0.72
VB	6.36±4.51	1.74±0.87

Table 2: Average polysemy and entropy per part-of-speech category.

performs best on word-POS combinations with low entropy values. However, since low entropy scores are caused by at the one end, many instances having the same sense and at the other, a very few instances having different senses, this implies that simply choosing the majority class for all instances already leads to high accuracies. In order to determine performance on those low entropy words, we selected 100 words with the lowest entropy values. The local context classifier has an average accuracy of 96.8% on these words, whereas the baseline classifier which always predicts the majority class has an average accuracy of 90.2%. These scores show that even in the case of highly skewed sense distributions, where the large majority of the training instances receives a majority sense, our memory-based learning approach performs well.

4 Results on the Senseval test data

In order to evaluate our word-expert approach on the SENSEVAL-2 test data, we divided the data into three groups as illustrated in Table 3. The *one-sense* group (90.5% accuracy) contains the words with one sense according to WordNet1.7. Besides the errors made for the "U" words, the errors in this group were all due to incorrect POS tags and lemmata. The *more-sense* $<$ *threshold* group (63.3% accuracy) contains the words with more senses but for which no word-expert was built due to an insufficient number (less than 10) of training instances. These words all receive the majority sense according to WordNet1.7. The *more-sense* $>$ *threshold* group (55.3% accuracy) contains the words for which a word-expert is built. In all three groups, top performance is for the nouns and adverbs; the verbs are hardest to classify. The last row of Table 3 shows the accuracy of our system on the English all words test set. Since all 2,473 word forms were covered, no distinction is made between precision and recall. On the complete test set, an accuracy of 64.4% is obtained according to the fine-grained SENSEVAL-2 scoring.

This result is slightly different from the score obtained during the competition (63.6%), since for these new experiments complete optimization was performed over all parameter settings. Moreover, in the competition experiments, Ripper (Cohen, 1995) was used as the keyword classifier, whereas in the new experiments TIMBL was used for training all classifiers. Just as in the SENSEVAL-1 task for English (Kilgarriff and Rosenzweig, 2000), overall top performance is for the nouns and adverbs. For the verbs, the overall accuracy is lowest: 48.6%. This was also the case in the train set (see Table 1). All 86 "unknown" word forms, for which the annotators decided that no WordNet1.7 sense-tag was applicable, were mis-classified.

Although our WSD system performed second best on the SENSEVAL-2 test data, this 64.4% accuracy is rather low. When only taking into account the words for which a word-expert is built, a 55.3% classification accuracy is obtained. This score is nearly 20% below the result on the train set (see Figure 1): 73.8%. A possible explanation for the accuracy differences between the word-expert classifiers on the test and train data, is that the instances in the Semcor training corpus do not cover all possible WordNet senses: in the training corpus, the words we used for the construction of word-experts had on average 4.8±3.2 senses, whereas those same words had on average 7.4±5.8 senses in WordNet. This implies that for many sense distinctions in the test material no training material was provided: for 603 out of 2,473 test instances (24%), the assigned sense tag (or in case of multiple possible sense tags, one of those senses) was not provided in the train set.

5 Conclusion

In this paper, we evaluated the results of the Antwerp automatic disambiguation system in the context of the SENSEVAL-2 English all words task. Our approach was to create word-experts per word-POS pair. These word-experts consist of different classifiers/voters, which all take different information sources as input. We concluded that there was no information source which was optimal for all word-experts. But we also showed that selecting the opti-

		nouns	verbs	adverbs	adjectives	U	Total
One-sense	#	263	29	110	89	22	513
	acc.	98.9	72.4	96.4	86.5	0.0	**90.5**
More-sense<threshold	#	241	120	33	132	30	556
	acc.	74.3	57.5	72.7	60.6	0.0	**63.3**
More-sense>threshold	#	563	405	158	244	34	1,404
	acc.	63.4	44.2	59.5	59.8	0.0	**55.3**
Total	#	1,067	554	301	465	86	2,473
	acc.	**74.6**	**48.6**	**74.4**	**65.2**	**0.0**	**64.4**

Table 3: Results on the SENSEVAL-2 test data.

mal classifier/voter for each single word-expert led to major accuracy improvements.

Since not all words were equally hard/easy to predict, we also evaluated the results of our WSD system in terms of the available number of training items, the number of senses and the sense distributions in the data set. Suprisingly, we observed that the available number of training items was not an accurate measure for task difficulty. But we furthermore concluded that the fluctuations in accuracy largely depend on the polysemy and entropy of the ambiguous words. On the basis of these results, we conclude that a more coarse-grained granularity of the distinction between word senses would increase performance of the WSD systems and make them a possible candidate for integration in practical applications such as machine translation systems.

When evaluating our system on the test set, accuracy dropped by nearly 20% compared to scores on the train set, which could be largely explained by lack of training material for many senses. So the creation of more annotated data is necesssary and will certainly cause major improvements of current WSD systems and NLP systems in general (see also (Banko and Brill, 2001)).

References

M. Banko and E. Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 26–33.

W.W. Cohen. 1995. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, pages 115–123.

W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. 1996. Mbt: A memory-based part of speech tagger-generator. In E. Ejerhed and I. Dagan, editors, *Fourth Workshop on Very Large Corpora*, pages 14–27.

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, ILK Technical Report 01-04.

G. Escudero, L. Marquez, and G. Rigau. 2000. Boosting applied to word sense disambiguation. In *European Conference on Machine Learning*, pages 129–141.

C. (ed.) Fellbaum. 1998. *WordNet : An Electronic Lexical Database*. MIT Press.

V. Hoste, A. Kool, and W. Daelemans. 2001. Classifier optimization and combination in the english all words task. In *Proceedings of Senseval-2*, pages 83–86.

V. Hoste, I. Hendrickx, W. Daelemans, and A. van den Bosch. 2002. Parameter optimization for machine-learning of word sense disambiguation. *Natural Language Engineering*, 8(3):to appear.

N. Ide and J. Véronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for english senseval. *Computers and the Humanities. Special Issue on SENSEVAL*, 34(1-2):15–48.

H.T. Ng and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 40–47.

M. Stevenson and Y. Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–349.

A. van den Bosch and W. Daelemans. 1999. Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 285–292.

J. Veenstra, A. Van den Bosch, S. Buchholz, W. Daelemans, and J. Zavrel. 2000. Memory-based word sense disambiguation. *Computers and the Humanities*, 34(1/2):171–177.

D. Yarowsky. 2000. Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, 34(1/2):179–186.

G. K. Zipf. 1935. *The psycho-biology of language: an introduction to dynamic philology*. Cambridge, MA: MIT Press.