Classifier optimization and combination in the English all words task.

Véronique Hoste and **Anne Kool** and **Walter Daelemans**
CNTS - Language Technology Group
University of Antwerp
Universiteitsplein 1, 2610 Wilrijk
hoste@uia.ua.ac.be, kool@uia.ua.ac.be, daelem@uia.ua.ac.be

Proceedings of SENSEVAL-2, Preiss & Yarowsky, eds., p. 83-86, 2001

Abstract

We report on the use of machine learning techniques for word sense disambiguation in the English all words task of SENSEVAL2. The task was to automatically assign the appropriate sense to a possibly ambiguous word form given its context. A "word expert" approach was adopted, leading to a set of classifiers, each specialized in one single word form-POS combination. Experts consist of multiple classifiers trained on Semcor using two types of learning techniques, viz. memory-based learning and rule-induction. Through optimization by cross-validation of the individual classifiers and the voting scheme for combining them, the best possible word expert was determined. Results show that especially memory-based learning in a word-expert approach is a feasible method for unrestricted word-sense disambiguation, even with limited training data.

1 Introduction

We report on the use of machine learning, especially memory-based learning and classifier combination, for word sense disambiguation (WSD) in the English all words task of SENSEVAL2. WSD can be described as the problem of assigning the appropriate sense to a given word in a given context. Machine learning techniques show state-of-the-art accuracy on WSD, e.g. memory-based learning (Ng and Lee, 1996; Veenstra et al., 2000), decision lists (Yarowsky, 2000), and combination methods (Escudero et al., 2000).

Results of the first SENSEVAL exercise for English (Killgarriff and Rosenzweig, 2000), in which only a restricted set of words had to be disambiguated, showed that supervised learning systems outperform unsupervised ones, even when little corpus training material was available. In our submission to SENSEVAL2, we investigated whether the supervised learning approach can be scaled to the all-words task. As a back-off for word-tag pairs for which no or not enough training data was available, we used the most frequent sense in the WordNet1.7 sense lexicon (Fellbaum, 1998) as default classifier in the disambiguation process. Sense disambiguation was mainly performed by a memory-based learning classifier. Also the use of rule induction was explored. Furthermore, the outputs of these different classifiers were combined in order to study the usefulness of different voting strategies. Results show that all classifiers outperform the WordNet baseline and that memory-based learning compares favorably to rule induction and different voting strategies.

In the remainder of this paper, we first outline the sense-disambiguation architecture used in the experiments, and discuss the word expert approach and the optimization procedure. Then we report on the generalization accuracy achieved for the SENSEVAL2 test data.

2 Experimental Setup

2.1 Preprocessing

In the experiments, the Semcor corpus included in WordNet1.6 was used as training corpus. In the corpus, every word is linked to its appropriate sense in the lexicon. Texts that were used to create the semantic concordances were extracted from the Brown Corpus and then linked to senses in the WordNet lexicon. The training corpus consists of 409,990 wordforms, of which 190,481 are sense-tagged. For each word form in the corpus, a lemma and a part of speech is given.

The test data in the English all words task consist of three articles on different topics, with at total of 2,473 words to be sense-tagged. For

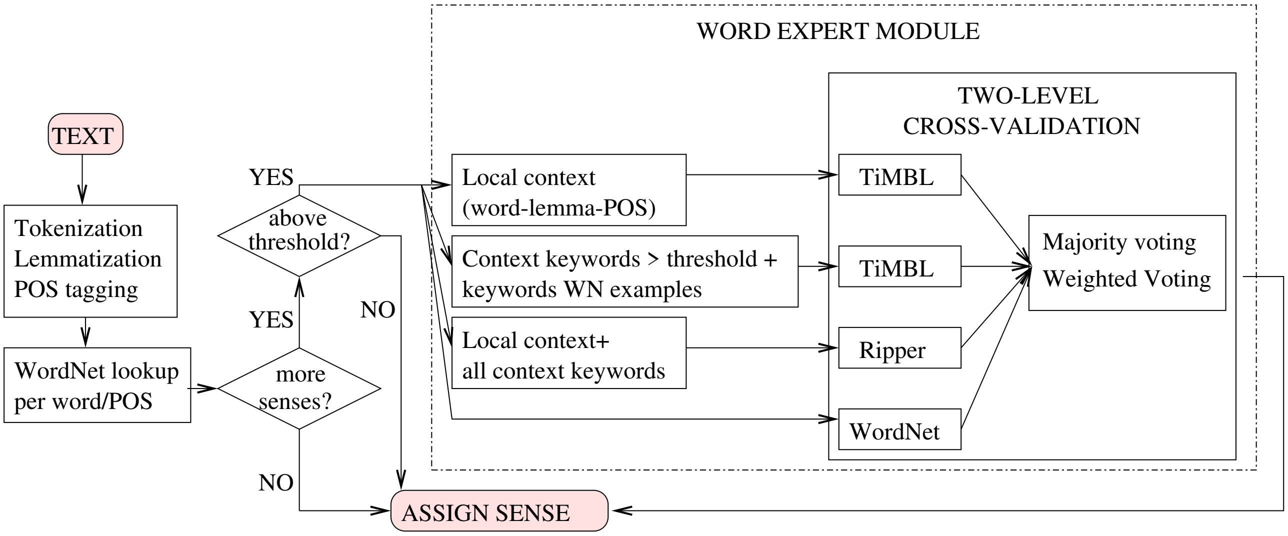

Figure 1: Disambiguation process.

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[1], was used. On the basis of word and POS information, lemmatization was done[2].

2.2 Word experts

After the preprocessing stage, WordNet1.7 was used to guide the sense disambiguation process. For every combination of a word form and a POS, WordNet 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 WordNet sense was assigned. In case of more possible senses, a threshold of 11 occurrences in the Semcor training data was determined. For all words below this threshold, the most frequent sense according to WordNet was assigned as sense-tag. For the other words, which represent more than 60% of the word forms to be sense-tagged, word experts were built for each word form-POS combination, leading to 568 word experts for the SENSEVAL2 test data.

These word experts consist of different trained subcomponents (see Figure 1) which make use of different knowledge.

The first subcomponent is trained using TiMBL, a package containing several memory-based learning algorithms and metrics (Daelemans et al., 2000). It takes as input a vector representing the local context of the focus word in a window of three word forms to the left and three to the right. For the focus word, also the lemma and POS are provided. For the context word forms, POS information is given. E.g., the following is a training instance:

```
many JJ times NNS , , yet yet RB on
IN each JJ occasion NN yet%4:02:02::.
```

During training, those instances are stored in memory and during sense-tagging, the instance most similar to that of the ambiguous word and its context is selected and the associated class is returned as sense-tag.

A second subcomponent of each word expert trained with TiMBL is trained with information about possible disambiguating content keywords in a context of three sentences. The method used to extract these keywords for each sense is based on the work of (Ng and Lee, 1996). They determine the probability of a sense s of a focus word f given keyword k by dividing $N_{s,kloc}$ (the number of occurrences of a possible local context keyword k with a particular focus word-POS combination w with a particular sense s) by N_{kloc} (the number of occurrences of a possible local context keyword $kloc$ with a particular focus word-POS combi-

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

[2]With a memory-based lemmatizer trained by Antal van den Bosch, see `http://ilk.kub.nl/`

nation w ignoring its sense). In addition, we also took into account the frequency of a possible keyword in the complete training corpus N_{kcorp}.

$p(s|k) = \frac{N_{s,kloc}}{N_{kloc}} \times (\frac{1}{N_{kcorp}})$

A word is a keyword for a given sense if (i) the word occurs more than M_1 times in that sense s, where M_1 is a predefined minimum number of times and if (ii) $p(s|k) \geq M_2$ for that sense s, where M_2 is some predefined minimum probability. Due to time restrictions M_1 was not optimized by cross-validation, but arbitrarily set to 3 and M_2 to 0.001.

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 that accompany the different sense definitions for a given focus word in WordNet. For each combination of a word form, POS and sense, all content words were extracted and added to the input vector of the memory-based learner. Both the contextual keywords and the example keywords were represented as binary features, with a value of 1 when the keyword was present in the example and 0 if not[3].

The third subcomponent of each word expert was trained with Ripper (Cohen, 1995), a rule learning algorithm, allowing both single-valued and set-valued attributes. In our disambiguation task, the ripper input vector contained local context feature values (as the first TiMBL), and a set-valued feature with all content words in a context of three sentences.

3 Optimization and Voting

In order to improve the predictions of the different single learning algorithms, algorithm parameter optimization was performed where possible. Furthermore, the possible gain in accuracy of different voting strategies was explored.

3.1 Optimization

For the first TiMBL memory-based learner, backward sequential selection (BSS) (Aha and Bankert, 1994) was performed for each word form-POS combination. BSS starts from the complete feature set and generates in each iteration new subsets by discarding a feature. The feature string with the best performance is retained. Furthermore, the use of different feature weighting possibilities was explored, viz. gain ratio weighting, information gain weighting, chi-squared weighting and shared variance weighting. For each feature weighting possibility, the k value, representing the number of nearest neighbours used for extrapolation, was varied between 1 and 19. Leave-one-out was used as testing method: testing was done on each instance of the training file, while the remainder of the training file functioned as training material.

Due to the size of the feature vectors for the second memory-based learner, which takes content words from the surrounding sentences and from the example sentences in the WordNet definitions as input, no feature selection was performed. For the same reasons, 10-fold cross-validation was used as testing method: the training data was split into 10 different parts and in each iteration, one part served as test set, while the remainder was used to train the classifier. The k value was varied (1-19), different weighting techniques (gain ratio weighting, chi-squared weighting and log likelihood weighting) and different distance metrics (number of mismatches, number of matches, number of matches minus number of mismatches) were explored.

For Ripper, the default parameter settings were used, due to time constraints and the slowness of the cross-validation process. 10-fold-cross-validation was used as testing method.

3.2 Voting

On the output of these three (optimized) classifiers and the default 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, more weight is given to the taggers with a higher overall 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 learning classifiers without taking into ac-

[3] Since no length limitations were taken into account when building these vectors, they could grow very large. Therefore, a version of TiMBL was used that is optimized for sparse binary features, and allows a positional representation of the active keywords rather than a binary one, written by Jakub Zavrel.

Classifier	no. WE
Default (WordNet1.7)	16
TiMBL (context)	**155**
TiMBL (keywords)	**185**
Ripper	16
Majority Voting	33
Weighted Voting	58
Majority Voting (no WordNet)	53
Weighted Voting (no WordNet)	52
	568

Table 1: Best performing word experts on the Semcor train set

count the WordNet class. Table 1 shows the best performing classifiers per word form-POS combination of the Semcor train set: both optimized memory-based learners outperform the other classifiers.

4 Results

Table 2 shows the accuracy of our disambiguation system on the English all words test set. Since all 2,473 word forms were covered, no distinction is made between precision and recall. An accuracy of 63.61% and 64.54% were obtained according to the fine-grained and coarse-grained SENSEVAL2 scoring, respectively. Just as in the first SENSEVAL task for English (Kilgarriff and Rosenzweig, 2000), top performance was for the nouns. All 86 "unknown" word forms, for which the test set annotators decided that no WordNet1.7 sense-tag was applicable, were obviously incorrectly classified.

	key	fine %	coarse %
noun (%1)	1,067	74.51	75.45
verb (%2)	554	47.83	49.64
adj. (%3-%5)	465	62.58	63.44
adv. (%2)	301	73.42	73.42
unkn.	86	0.00	0.00
total	**2,473**	**63.61**	**64.54**

Table 2: Results on the SENSEVAL2 test data.

5 Conclusion

This paper reported on the architecture and the results of the CNTS-Antwerp automatic disambiguation system in the context of the SENSEVAL2 English all words task. Disambiguation per word form-POS pair is performed through the application of word experts trained on local context information and cross-validated on the limited available training data. Among these word experts, optimized memory-based learning proves to be more accurate than default Ripper rule-induction and various voting strategies.

Acknowledgements

We like to thank Antal van den Bosch for taking care of the lemmatization and Erik Tjong Kim Sang for programming support.

References

D.W. Aha and R.L. Bankert. 1994. Feature selection for case-based classification of cloud types: An empirical comparison. In *Proceedings of the 1994 AAAI Workshop on Case-Based Reasoning*, pages 106–112. AAAI Press.

W.W. Cohen. 1995. Fast effective rule induction. In *Proc. 12th International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann.

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. 2000. Timbl: Tilburg memory based learner, version 3.0, reference guide.

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

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

A. Killgarriff and J. Rosenzweig. 2000. English senseval: Report and results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 1239–1243.

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

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.