Sometimes Less Is More: Romanian Word Sense Disambiguation Revisited

Georgiana Dinu
University of Tübingen
dinu@sfs.uni-tuebingen.de

Sandra Kübler
Indiana University
skuebler@indiana.edu

Abstract

Recent approaches to Word Sense Disambiguation (WSD) generally fall into two classes: (1) information-intensive approaches and (2) information-poor approaches. Our hypothesis is that for memory-based learning (MBL), a reduced amount of data is more beneficial than the full range of features used in the past. Our experiments show that MBL combined with a restricted set of features and a feature selection method that minimizes the feature set leads to competitive results, outperforming all systems that participated in the SENSEVAL-3 competition on the Romanian data. Thus, with this specific method, a tightly controlled feature set improves the accuracy of the classifier, reaching 74.0% in the fine-grained and 78.7% in the coarse-grained evaluation.

Keywords

Word Sense Disambiguation, Romanian, memory-based learning

1 Introduction

Recent approaches to Word Sense Disambiguation (WSD) generally fall into two classes: (1) information-intensive approaches and (2) information-poor approaches. The typical features that are used in information-intensive approaches are the part-of-speech tags of the ambiguous word, the surrounding words with their part-of-speech (POS) tags, as well as collocational features from a larger context. If available, additional information such as document type, named entity information, or syntactic information are added. These approaches use supervised learning with a separate classifier for each ambiguous word. Additionally, the best results are achieved by combining different classifiers into an ensemble, in which the final decision is based on the votes of the different classifiers. The other end of the spectrum generally restricts the available information to types of information that can be extracted from small amounts of text without running into data-sparseness problems. Generally, such systems use a combined approach for all words and only a single algorithm to solve the problem.

From the description above, it is already clear which category is expected to perform better: supervised learning with the maximum of information and with an ensemble classifier. In contrast, our hypothesis is that for some classifiers, a reduced amount of data is more beneficial than the full range of features that have been used in the past.

As data set, we chose the Romanian data from the SENSEVAL-3 competition [7]. This data set is rather small with regard to both the number of ambiguous words (39) and the number of instances for each word (between 19 and 266). Such a size can be expected for many languages for which there is no financial interest. The supervised learning method we chose is memory-based learning (MBL), a k-nearest neighbor approach. It bases the classification of a new instance on the k most similar instances found in the training data. This approach has been shown to be successful for a range of problems in NLP [1, 2] Daelemans et al. argue that MBL has a suitable bias for such problems because it allows learning from atypical and low-frequency events, thus enabling a principled approach to the treatment of exceptions and sub-regularities in language. Another advantage of MBL lies in the fact that it can work with complete words as feature values. As a consequence, however, MBL is also sensitive to large numbers of features that are only relevant for the classification of specific instances but not for all instances. This is the case even when features are weighted. This last characteristic of MBL suggests that a good balance between too much and too little information must be found, which in turn makes it a good candidate for our approach.

In the following sections, we show that MBL combined with a restricted set of features of three context words to each side of the ambiguous word, their POS tags, the closest verbs, nouns, and prepositions on both sides, lead to competitive results. We then employ two feature selection methods to further optimize the feature set. The results show that forward selection, which selects a smaller feature set, leads to optimal results, reaching an accuracy of 74.0% in the fine-grained and 78.7% in the coarse-grained evaluation, outperforming all systems that participated in the SENSEVAL-3 competition on the Romanian data.

2 Related Work

In building a supervised WSD system, one of the main decisions is the choice of a classifier. Memory-based learning (MBL) is a supervised learning method that has been successfully used in WSD after a difficult start: Mooney [8] reports the first experiment using a simple nearest neighbor method in a comparison of different machine learning methods for disambiguating the word *line*. He attributes the low performance

CT_k	the token at position k [-3..3] relative to the target word; CT_0: target word
CP_k	the POS tag of the token at position k
VA	the first verb found after the target word
VB	the first verb found before the target word
NA	the first noun found after the ambiguous word
NB	the first noun found before the ambiguous word
PA	the first preposition found after the ambiguous word
PB	the first preposition found before the ambiguous word

Table 1: *The complete list of features used in the experiments*

of this approach to the fact that it did not use feature weighting. Escudero et al. [4] show later that one of the problems for the nearest neighbor approach was the high number of context features, which resulted in a very sparse feature matrix. After they introduced feature weighting and collapsed the context features into one set-valued feature (and modified the similarity metric accordingly to calculate a set-based similarity), they showed that the nearest neighbor method outperforms the naive Bayes model, Mooney's best performing model. Veenstra et al. [10] present a system that competed successfully in SENSEVAL-1. They use context features (word form and POS tag of the ambiguous word and 2 words on either side) as well as keyword features and definition features. For keyword features, the most informative words from the context are used. Veenstra et al.'s results show that the optimal settings depend on the individual ambiguous words. There is no optimal setting that works equally well for all words. Mihalcea [6] shows that even if feature weighting methods are used, memory-based learning is susceptible to irrelevant or redundant features. She improved her results for the SENSEVAL-2 English lexical sample task by using forward selection. This method reduces the number of features on average to 3.7 for nouns, 4.4 for adjectives, and to 4.5 for verbs.

Lee and Ng [5] thoroughly investigate which knowledge sources are relevant for WSD. They used four different classifiers and the SENSEVAL-1 and SENSEVAL-2 English data. Their findings show a trend that classifiers perform best when all features are offered to the systems. The Support Vector Machines classifier and AdaBoost perform best without feature selection while the naive Bayes and the decision tree classifier profit from feature selection. The only experiment in which a classifier performs best on a restricted set of features is the combination of the decision tree classifier with SENSEVAL-2 data, but the difference to the results on all features is rather small (57.2% for only collocational features versus 56.8% for all features). These findings suggest that a complete feature set provides an optimal setting for WSD.

WSD for Romanian was one of the tasks in SENSEVAL-3. In the competition, seven systems were evaluated. We will concentrate on the three best performing systems here: SWAT-HK-boost, SWAT-HK [11] and the Duluth system [9]. SWAT-HK-boost is a boosting approach that used context features and bigrams and trigrams of words and parts of speech. SWAT-HK is an ensemble voting approach based on SWAT-HK-boost and four other classifiers, using the same feature set as SWAT-HK-boost. The Duluth system uses an ensemble of three decision trees, each trained on a different set of features, word bigrams, word unigrams, and word co-occurrence features. Note that all three best-performing systems use a combination of simpler classifiers. SWAT-HK-boost reaches a fine-grained accuracy of 72.7%, SWAT-HK 72.4%, and the Duluth system 71.4%. Since all systems described here attempted all words, precision and recall are identical, and we only report accuracy.

3 Experiments

For all experiments reported here, we used the SENSEVAL-3 Romanian lexical sample data [7], which consists of labeled examples for 39 ambiguous words: 25 nouns, 9 verbs, and 5 adjectives[1]. In order to allow a comparison of our experiments to systems that participated in SENSEVAL, we used the designate3d training and test sets. The senses, with an average of 8.8 fine-grained senses per word (4.7 coarse-grained), are manually extracted from a Romanian dictionary.

The experiments reported here were conducted with TiMBL [3], a memory-based learning system. TiMBL was used with the following settings: the IB1 algorithm, Gain Ratio for feature weighting, and $k = 1$. For evaluation, a leave-one-out cross-validation was performed.

As reported above, the experiments were conducted with a rather restricted feature set: We used lexical and POS information of the ambiguous word and of a context of three words on both sides, as well as information concerning the closest verbs, nouns, and prepositions in the sentence. Table 1 lists the complete set of features.

For each word, an optimal set of features is determined. We performed experiments with **forward** and **backward selection**. Initially, a pool of features containing all the features is generated. Forward selection starts the selection process by selecting a single feature from the pool, running the classifier with this single feature. Then the feature with the highest accuracy is selected. In the next step, the second feature is selected based on combinations of the selected feature and the remaining features in the pool. Features are added as long as accuracy improves. Backward selection starts with the complete pool of features. In the first step, experiments are conducted removing one of the features. Then the feature whose absence results in the highest improvement in accuracy is removed permanently. The process of removing features continues as long as accuracy improves or remains stable.

The forward selection experiment is similar to the experiment that Mihalcea [6] performed for the

[1] For the list of words and characteristics, cf. Tables 5 to 7.

forward selection										
feature	NA	CT_0	NB	CT_1	CT_{-1}	CP_0	CT_2	CP_{-1}	CT_{-2}	VB
# words	28	25	24	19	18	18	14	15	13	12
backward selection										
feature	CP_1	CP_{-1}	CT_1	CP_{-2}	PB	NA	VB	CP_2	CT_{-1}	CP_0
# words	28	27	25	23	23	22	22	21	20	19

Table 2: *The most commonly selected features in per-word feature selection*

	fine	coarse
baseline (MFS)	58.5	62.8
all features	71.2†	76.4†
backward selection	72.7†	77.4*
forward selection	**74.0***	**78.7***

Table 3: *Results for the feature-selection experiments; all differences are significant on the 0.01 (*) / 0.001 (†) level, McNemar*

POS	forward	backward
nouns	7.4	9.9
verbs	5.0	11.0
adjectives	6.8	7.2
overall	6.8	9.8

Table 4: *Feature selection and number of features*

SENSEVAL-2 English lexical sample task. Note, however, that Mihalcea used a larger feature pool including collocation information, sense specific keywords, named entity information, and syntactic information.

4 Results

The evaluation of the experiments was performed with the SENSEVAL scoring software, which provides coarse-grained and fine-grained accuracies.

4.1 Feature Selection

Table 3 gives the results of the selection process. The baseline reported here is computed by assigning the most frequent sense (MFS), as computed from the training data, to the test instances. It is evident that TiMBL, even without much optimization of the parameter settings, outperforms the baseline significantly. Classification accuracy can be further improved when these system parameters are optimized. However, this is irrelevant for the experiments reported here.

The results also show that WSD for Romanian profits from both feature selection methods, with forward selection outperforming backward selection. Our starting hypothesis was that irrelevant or redundant features harm TiMBL's performance. A look at the average number of features after feature selection shows that this is true. Table 4 reports the average number of features used for the different selection algorithms and POS categories. From a total of 20 features, forward selection uses only approximately 7 features and backward selection 10. From these results, we can conclude that not all of the features of the original set are helpful for the task and that TiMBL suffers from irrelevant or redundant features despite the use of a feature weighting mechanism. Additionally, backward selection does not restrict the number of features as much as forward selection does. The forward selection results are comparable to the findings of Mihalcea [6], where a similar selection algorithm on SENSEVAL-2 English data improves the average performance by 3.9% in nouns and verbs, and 5.4% in adjectives.

The selection experiments can also be used to answer a linguistically relevant question: Which features provide the best information for WSD? Table 2 reports the features used in classifying the most words and the number of words for which the feature was used (out of a total of 39 words). It is surprising to see that the two selection methods prefer different types of features: While forward selection prefers word forms over POS information, backward selection has a more balanced distribution, favoring POS tags as the most often used features.

As reported before, the near context is a very good indicator for a word's sense. The words surrounding the target word seem to be most helpful for disambiguation, and their relevance decreases with an increasing distance from the target word. The nouns preceding and following the ambiguous word as well as the word form of the ambiguous word itself play a very important role. This is a general trend for all words, irrespective of their parts of speech. The last feature may be surprising since one could assume that the forms would be very similar considering that there is a separate classifier for each ambiguous word. However, Romanian is an inflected language, so that the word form can provide information on some morphological and syntactic features, especially in the absence of further linguistic analysis.

Adjectives are special in that they are biased towards choosing features extracted from preceding context (preceding noun, preceding tokens), unlike verbs or nouns, which prefer an extraction window centered around the target word. On average, a noun chooses 3 features from the left context and 3 from the right. For verbs, its on average 2 words on each side while an adjective chooses 3.4 features from the left context and 2.2 from the right. Part of the explanation for the last number can be found in the fact that in Romanian, both predicative and attributive adjectives follow the constituents they modify, which presumably are important indicators for the sense of the adjective.

One of the extreme examples of words that were dis-

word	translation	no. senses (f/c)	size	MFS (f)	MFS (c)	acc. (f)	acc. (c)
ac	needle	16/7	127	50.8	50.8	73.8	75.4
accent	accent	5/3	172	73.6	77.0	89.7	93.1
actiune	action	10/7	261	39.8	39.8	61.7	85.2
canal	channel	6/5	134	68.2	68.2	69.7	75.8
circuit	circuit	7/5	200	49.5	50.5	59.4	65.3
circulatie	circulation	9/3	221	45.6	45.6	59.4	68.4
coroana	crown	15/11	252	58.7	61.9	77.0	77.8
delfin	dolphin	5/4	31	100	100	80.0	80.0
demonstratie	demonstration	6/3	229	64.3	64.3	73.0	73.0
eruptie	eruption	2/2	54	40.7	40.7	81.5	81.5
geniu	genius	5/3	106	72.2	77.8	64.8	70.4
nucleu	nucleus	7/5	64	78.8	78.8	81.8	81.8
opozitie	opposition	12/7	266	96.3	96.3	95.5	95.5
perie	brush	5/3	46	79.2	95.8	75.0	95.8
pictura	painting	5/2	221	47.7	47.7	75.7	81.1
platforma	platform	11/8	226	38.8	38.8	58.6	58.6
port	port	7/3	219	51.9	51.9	81.5	83.3
problema	problem	6/4	262	44.3	44.3	69.5	69.5
proces	process	11/3	166	62.2	64.6	81.7	82.9
reactie	reaction	7/6	261	83.2	83.2	83.2	83.2
stil	style	14/4	199	60.4	80.2	62.4	76.2
timbru	stamp	7/3	231	94.0	99.1	94.8	98.3
tip	type	7/4	263	76.3	76.3	87.8	89.3
val	wave	15/9	242	85.1	85.1	87.6	88.4
valoare	value	23/9	251	63.2	75.2	72.8	85.6
total	-	8.9/4.9	-	63.8	66.2	75.9	80.6

Table 5: *MBL with per-word forward-feature selection: nouns*

ambiguated using a very small number of features is the verb *câştiga* (to win). By only using the word form of the verb and the word form of the following noun, disambiguation accuracy increases from 52.2% (MFS) to 72.2%. An examination of the training data provides an explanation for this extreme behavior: This word has five senses but the predominant two senses are to gain material benefits, and to win a sports competition (or a contest, a trial). The following noun (NA) is a very good sense indicator in this case since in most cases this feature contains the object of the verb: *bani*, *dolari*, *mărci* or *lei* (money or various currencies) for the first sense, and *partida*, *derby*, *meci* (sport competitions) for the second sense. Thus, this single feature increases accuracy from 50.2% (MFS) to 66.9% on the training data. Additionally, the word form (CT_0) of the ambiguous word helps to distinguish the two senses. For example, the third person plural form *caştigăm* is predominantly used within the winning a sport competition sense, as 'our team (we) won the game'. CT_0 is thus the feature that brings the second best improvement, increasing accuracy from 66.9% to 71.8%. Adding any of the other features results in accuracy drops varying between 0.5% and 12%, suggesting that for this word, all these features provide irrelevant information.

4.2 Comparison with SENSEVAL-3 Participants

In contrast to most state-of-the-art WSD systems, our approach uses a rather impoverished feature set. It contains neither collocational features nor syntactic or global features. Thus, the conjecture is that the system should be at a disadvantage when compared to systems that had access to such data sources. A comparison with two of the best 3 systems in the SENSEVAL-3 competition, the SWAT-HK-boost system [11], and the Duluth system [9], shows that this is not the case (cf. Table 6). On the contrary, our memory-based system (with default parameter settings) outperforms both systems on this task[2]. The difference to the SWAT-HK-boost system is statistically significant (McNemar), on the 0.05 level.

system	fine	coarse
feature selection MBL	**74.0**	**78.7**
SWAT-HK-boost [11]	72.7	77.1
Duluth [9]	71.4	75.2

Table 6: *System comparison*

One reason why we did not use collocational features is that collocations tend to increase the number of features by at least an order of magnitude, with most of the features having zero values for each example. Escudero et al. [4] show that such a selection of features harms the performance of k-nearest neighbor approaches. Since their suggested solution, a set-based approach in calculating the similarity of feature values, is not available in TiMBL, we decided not to use this type of information.

4.3 Results for Individual Words

Table 5 gives the results of the forward selection experiment for the individual nouns and Table 7 for verbs

[2] Wicentowski et al. [11] report a fine-grained accuracy of 73.3% for SWAT-HK-boost after an error was corrected.

word	translation	no. senses (f/c)	size	MFS (f)	MFS (c)	acc. (f)	acc. (c)
			Verbs				
castiga	win	5/4	227	52.2	52.2	72.2	72.2
citi	read	10/4	259	82.3	90.8	82.3	89.2
cobori	descend	11/6	252	47.7	75.8	68.0	85.2
conduce	drive	7/6	265	55.2	56.0	81.3	82.1
creste	grow	14/6	209	43.7	43.7	72.8	74.8
desena	draw	3/3	54	81.5	81.5	81.5	81.5
desface	untie	11/5	115	27.6	32.8	53.4	56.9
fierbe	boil	11/4	83	32.6	37.2	48.8	58.1
indulci	sweeten	7/4	19	40.0	80.0	60.0	80.0
total	-	8.7/4.6	-	53.9	61.5	72.3	77.9
			Adjectives				
incet	slow	6/3	224	41.6	41.6	79.6	79.6
natural	natural	12/5	242	23.6	51.2	67.5	74.8
neted	smooth	7/3	34	41.2	52.9	41.2	41.2
oficial	official	5/3	185	53.1	53.1	72.9	72.9
simplu	simple	15/6	153	36.6	36.6	46.3	48.8
total	-	9/4	-	38.1	46.4	66.8	69.4

Table 7: *MBL with per-word forward-feature selection: verbs and adjectives*

and adjectives. Compared to the MFS baseline, nouns achieve a net gain of 12.1% (14.4% coarse-grained) and verbs 18.4% (16.4% coarse). Adjectives are disambiguated best for the Romanian task, achieving an accuracy gain of 28.7% (23% coarse). The error reduction rates for fine-grained scores are 33.4% for nouns, 40% verbs and 46.3% for adjectives.

5 Conclusion and Future Work

We have shown that when using a memory-based classifier for WSD, the feature set needs to be tightly controlled. In contrast to other experiments, the MBL classifier achieved optimal results with on average seven features per word. The most important features are the noun following the ambiguous word, the word form of the ambiguous word, and the preceding noun: features that can easily be retrieved. The experiments show that forward selection allows a greater reduction of features: on average seven features as compared to an average of ten features for backward selection. This is another indication that MBL suffers from irrelevant or redundant features. These findings are partially in line with the findings of Lee and Ng [5] for a naive Bayes and a decision tree classifier: both show an increased performance when feature selection is performed. However, the initial feature set that Lee and Ng used was much larger than the one used in the present study. A logical explanation for the differences between the results reported here and Lee and Ng's findings can be found in the differences of the classifiers used in the experiments. All classifiers Lee and Ng used in their experiments are based on greedy learning approaches while MBL is a lazy learning approach. There is a slight chance, however, that the results reported here are due to idiosyncrasies in the Romanian data set. For this reason, the next step is to test the same combination of classifier and features on data sets for different languages. Another reason for the success of this combination may be a consequence of the rather limited size of the training data. Therefore, the combination suggested here needs to be tested on larger data sets with controlled data sizes.

References

[1] W. Daelemans and A. van den Bosch. *Memory Based Language Processing*. Cambridge University Press, 2005.

[2] W. Daelemans, A. van den Bosch, and J. Zavrel. Forgetting exceptions is harmful in language learning. *Machine Learning*, 34:11–43, 1999. Special Issue on Natural Language Learning.

[3] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. TiMBL: Tilburg memory based learner – version 5.1 – reference guide. Technical Report ILK 04-02, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University, 2004.

[4] G. Escudero, L. Márquez, and G. Rigau. Naive Bayes and exemplar-based approaches to Word Sense Disambiguation revisited. In *Proceedings of the 14th European Conference on Artificial Intelligence, ECAI'2000*, pages 421–425, Berlin, Germany, 2000.

[5] Y. K. Lee and H. T. Ng. An empirical evaluation of knowledge sources and learning algorithms for Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, PA, 2002.

[6] R. Mihalcea. Instance based learning with automatic feature selection applied to Word Sense Disambiguation. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipeh, Taiwan, 2002.

[7] R. Mihalcea, V. Năstase, T. Chklovski, D. Tătar, D. Tufiş, and F. Hristea. An evaluation exercise for Romanian Word Sense Disambiguation. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 29–32, Barcelona, Spain, 2004.

[8] R. J. Mooney. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing EMNLP*, pages 82–91, Philadelphia, PA, 1996.

[9] T. Pedersen. The Duluth lexical sample systems in SENSEVAL-3. In *SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs*, pages 203–08, Barcelona, Spain, 2004.

[10] J. Veenstra, A. van den Bosch, S. Buchholz, W. Daelemans, and J. Zavrel. Memory-based Word Sense Disambiguation. *Computers and the Humanities, Special Issue on Senseval, Word Sense Disambiguation*, 34(1/2):171–177, 2000.

[11] R. Wicentowski, G. Ngai, D. Wu, M. Carpuat, E. Thomforde, and A. Packel. Joining forces to resolve lexical ambiguity: East meets West in Barcelona. In *SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs*, pages 262–264, Barcelona, Spain, 2004.