Machine Learning for Domain-Adaptive Word Sense Disambiguation

Georgios Paliouras, Vangelis Karkaletsis, Costantine D. Spyropoulos
Institute of Informatics and Telecommunications,
NCSR "Demokritos",
15310, Aghia Paraskevi, Athens, Greece
Tel: +301-6503197, Fax: +301-6532175
{paliourg, vangelis, costass}@iit.nrcps.ariadne-t.gr

Abstract

This paper investigates the use of machine learning techniques for word sense disambiguation. The aim is to improve on the performance of general-purpose methods, by making the disambiguation method adaptable to new domains. Results are presented here for two different test cases: financial news from the Wall Street Journal, extracted from the SEMCOR corpus, and general-theme news from the same corpus. The two experiments show that the adaptive disambiguation method can achieve high recall and precision; more so in the restricted domain of financial news than in the general-theme case.

Introduction

The aim of the work presented in this paper is to improve word sense disambiguation (WSD) results, by providing a method which is adaptable to specific domains. Adaptivity is achieved through learning from empirical data, i.e., the WSD system is built/modified to perform well on a set of training data that is representative of a particular domain. The training data is extracted from pieces of text, which have been hand-tagged by experts in the domain. Thus, the task is to perform supervised learning and the methods that we examine here belong to this branch of machine learning; namely *symbolic supervised learning* from data.

The task of associating a word in text with one of a number of possible senses, which is the aim of WSD, arises in the context of several natural language processing (NLP) problems, such as machine translation and information extraction. Due to its importance, WSD has attracted the attention of researchers and has almost become a separate NLP task. Despite its practical importance, WSD is a difficult task even for humans to perform. An extreme position presented in (Kilgarriff, 1993) is that WSD is very hard for humans and therefore an overambitious task for machines. The difficulty of the task becomes apparent by the low success rates of most automated WSD systems, with the exception of some which concentrate on a small set of words, e.g. (Schuetze, 1992) and (Yarowsky, 1995). Our position on this issue is that we can improve the performance of WSD by restricting its scope to a particular domain. The underlying assumption is that there is less variability in the use of a word within a domain, rather than in unrestricted text.

Despite its potential desirability, the adaptation of a WSD system to a particular domain is problematic when done manually. It is a knowledge engineering task, involving all the problems associated with the manual acquisition of knowledge. Machine learning addresses exactly this issue and has so far been used successfully in a wide variety of real-world problems. A learning method provides automatic acquisition of knowledge, which in the case of symbolic learning takes the form of a rule base or a similar symbolic representation. Incorporating such a component in a WSD system can provide capabilities of automatic customisation of the system to a particular domain. The particular type of machine learning method that we use in this work performs supervised learning, i.e., some information is needed from the domain expert. However, this information does not take the form of organised knowledge, but hand-tagging of a training set of text pieces. This process is much simpler than the acquisition of a knowledge base and can be made very simple with the use of intelligent interfaces.

The work presented in this paper has been performed in the context of the research project ECRAN,[1] which examines the customisation of language resources for information extraction to a domain or a particular user. The task that we try to solve is the assignment of tags from the *Longman Dictionary of Contemporary English* (LDOCE) to the words in a piece of text. This text has passed through several stages of pre-processing: tokenisation, lemmatisation, sentence-splitting, part-of-speech tagging and named-entity identification. Wilks and Stevenson (1997) report results, in the context of ECRAN, with a general-purpose tagger, which is based on the definition of words in a dictionary. This method measures the overlap in the textual definition of a word with the definitions of collocated words. The senses that best fit to the context are chosen. This approach was introduced in (Lesk, 1986), while Cowie *et al*. (1992) improved its computational performance, using simulated annealing for the search. Wilks and Stevenson (1997) examine disambiguation on two different levels of generality: *homographs* and *senses*. A homograph in LDOCE is a

[1] Extraction of Content: Research at Near-market, LE-2110, Telematics Applications Programme, partially supported by the EU.

group of senses with related meaning for a word. The general-purpose WSD tagger achieved high recall (86%) at the level of homographs, but much lower at the level of individual senses (57%). For this reason, we attempted to improve the results on the latter task by using a WSD method that can adapt to a particular domain. The aim is to disambiguate between individual senses within a homograph, rather than between the homographs of a word.

Section 2 describes briefly the properties of the SEMCOR corpus, the machine learning method and the evaluation criteria that we use in this paper. Section 3 presents the first of the two test cases. In this case, the task is to disambiguate between words in news articles from the SEMCOR corpus. The disambiguator is a *decision tree* constructed by the machine learning algorithm C4.5 (Quinlan, 1993), which is the most widely used algorithm in practical applications. In section 4 we restrict the application domain, by examining only financial news articles from the SEMCOR corpus. Section 5 summarises the results achieved in the study, concluding with the questions which remain unanswered.

Experimental Setting

The data used in the following experiments are extracted from the SEMCOR corpus, which is a 200,000-word selection of news articles from the Wall-Street Journal. The important feature of this corpus is that the content words, i.e., nouns, verbs, adjectives and adverbs, have been hand-tagged with semantic information, as part of the WordNet project. The fact that the data are taken from news articles, and in particular from the Wall-Street Journal, already restricts significantly the domain for word-sense disambiguation. However, in our experiments we have gone a step further to select one part of the SEMCOR corpus: financial news articles. The aim was to assess the benefits of restricting the scope of WSD.

The SEMCOR corpus is tagged with WordNet sense-numbers. However, the dictionary used in ECRAN is the Longman Dictionary of Contemporary English (LDOCE). For this reason we translated the WordNet tags into their equivalent in LDOCE. This translation was supported by a resource that was constructed in the WordNet project: a mapping between the senses in the two dictionaries (Bruce and Guthrie, 1992). The mapping between WordNet and LDOCE senses suffers in several respects:

- there is a large number of senses on both dictionaries that have not been mapped onto senses in the other dictionary;
- the mapping between senses is hardly ever one-to-one, e.g. seven different Wordnet senses for the verb 'absorb' are mapped on the same LDOCE sense, while the word has four LDOCE senses;
- there are mistakes in the mapping, i.e., WordNet senses are mapped to irrelevant LDOCE senses.

Due to these problems, there is a loss of information in the translation of the data from WordNet to LDOCE tags. In average, only a quarter of the words in the corpus were assigned LDOCE senses, in our experiments. An additional problem is the assignment of WordNet senses in SEMCOR. Despite the thorough consistency checks that have been performed on the data, there seem to be some mistakes. In our experiments, we do not make use of word-meaning information in either of the two dictionaries and therefore such mistakes do not affect our results. However, since these resources are invaluable in NLP research, we believe that these issues are worth-raising.

One final stage of pre-processing for the data that is used here is the translation of words in the feature-vector representation, commonly used in machine learning. For each word, each LDOCE sense which could be used instead of the correct sense (i.e., all senses in the same homograph), is represented as a separate example case for learning. The correct sense is labelled as a positive example and all other senses as negative.[2] Each example case contains the following characteristic information about the word and the context in which it appears: the lemma of the word, the rank of the sense in LDOCE,[3] the part-of-speech tag for the word and the ten collocates (first noun/verb/preposition to the left/right and first/second word to the left).

Given example cases in the above-described format, the machine learning algorithm C4.5 constructs a decision tree, which can then be used to assign sense tags to unseen data. C4.5 generates decision trees, the nodes of which, for WSD, evaluate the descriptive features of words, i.e., the lemma, sense-rank, part-of-speech tag and the values of collocates. Following a path from the root to the leaves of the tree a sequence of such tests is performed, resulting in a decision about the appropriate sense for the word. Thus, each path from root to leaves is a conjunctive rule, the conditions of which are the individual nodes. Alternative paths are combined disjunctively. For instance, the following simple rule could appear in the decision tree:

IF lemma=bank **AND** sense-rank=1 **THEN** *true*
ELSEIF lemma=bank **THEN** *false*

where *true* and *false* signify whether the sense is appropriate or not.

The measures that we chose for the evaluation of our methods are those typically used in the language engineering and machine learning literature: *recall*, *precision* and *accuracy*. The recall measure counts the number of words that are assigned the correct sense, out of the total number of words to be assigned a sense. This corresponds to the ratio of true positive examples to the total number of positives in the test data. On the other hand, precision counts the number of words assigned the correct sense, out of the number of word-senses considered positive by the decision tree, i.e., the ratio of true positive to true and false positive examples. In addition to these two measures the percentage correct classification (accuracy), which is a standard measure for machine learning methods is used. In summary the three ratios:

recall = TP/P,
precision = TP/(TP+FP),
accuracy = (TP+TN)/(P+N),

where TP/FP and TN/FN stand for True/False Positive and True/False Negative and P/N for Positive/Negative examples.

[2] Due to the one-to-many mapping, more than one senses could be considered positive.

[3] Senses for each word in LDOCE are ordered according to the frequency in which they occur.

The performance of the system is always measured on unseen data. In order to arrive at a robust estimate of the method's performance, we use *10-fold cross-validation* at each individual experiment. According to this evaluation method, the dataset is split into ten, equally-sized subsets and the final result is the average over ten runs. In each run nine of the ten subsets of the data are used to construct the decision tree and the tenth is held out for the evaluation. Thus, each recall, precision and accuracy figure presented in the following section is an average over ten runs, rather than a single train-and-test result, which can often be accidentally high or low. The computational efficiency of C4.5 allows the use of 10-fold cross validation, without a significant effect on the progress of the experiments.

Finally, it should be noted that, unlike general-purpose WSD methods, the constructed decision tree makes no use of external resources for disambiguation.

Sense Disambiguation in General-Theme News Articles

In the first test case, we used a subset of the SEMCOR news articles, the subject of which varied. SEMCOR is organised in 103 files. We chose the first two sentences from each of the first 72 of the 103 files. The size of the final dataset was dictated by the size of the set in the second experiment, i.e., the financial news articles. The two should be comparable, so that valid conclusions on the results can be drawn. The 144 sentences of the SEMCOR data consisted of 4,262 word occurrences, of which 2,359 were tagged with WordNet senses. The translation to LDOCE resulted to 600 word occurrences with LDOCE tags, corresponding to 448 distinct words and 3,541 example cases for C4.5. Thus, the set had an average LDOCE polysemy of 3,541/600=5.9.

C4.5 facilitates pre- and post-pruning of the decision tree using a significance statistic. We have set the parameters of C4.5 so as to prevent any pruning and evaluated the full, unpruned tree. Table 1 presents the 10-fold cross-validation results obtained in this experiment.

Recall	*Precision*	*Accuracy*
65.2%	84.9%	88.9%

Table 1: Results on the general-theme data using C4.5.

Qualitatively, there seems to be an improvement over the general-purpose sense tagger. The method seems to be doing particularly well in terms of precision, but not so in terms of recall. This is an indication that the decision tree is conservative in labelling example cases as positive. Thus, it misses a large proportion of positive examples, but does not misclassify many negatives.

In order to set these results in context we present the results of two base cases. The first is a naïve rule choosing always the majority class in the data set, which in this case means that all cases are considered negative. Recall and precision are both zero in this case, but the accuracy is 76.1%, safely below the accuracy of the decision tree. The second base case is more interesting: we consider as appropriate the first sense of each word in LDOCE, i.e., the most frequently used sense. This rule gives higher accuracy, but quite low recall and precision values. Table 2 presents these results. Clearly, any results close or below these values are not acceptable as a solution to the problem. The performance of C4.5 is much better than the base case.

Recall	*Precision*	*Accuracy*
50.0%	65.6%	81.8%

Table 2: The base case of the most frequent sense in the general-theme data.

The results acquired in this broad-domain experiment are encouraging. The decision tree seems to improve on the results of the general-purpose tagger. However, one of the properties of the constructed decision tree is unusual for machine learning work: it is very large. The average size of the ten decision trees that gave the results of Tab. 1 was 8,339.6, which is much larger than the number of distinct words in the training set (448). This is usually an indication of a low level of generalisation. Low generalisation is usually undesirable and followed by low performance on unseen data. However, the results that we acquired on unseen data are high. In order to explain this phenomenon, we used the post-pruning mechanisms of C4.5 to reduce the size of the tree and measured is performance at different sizes.

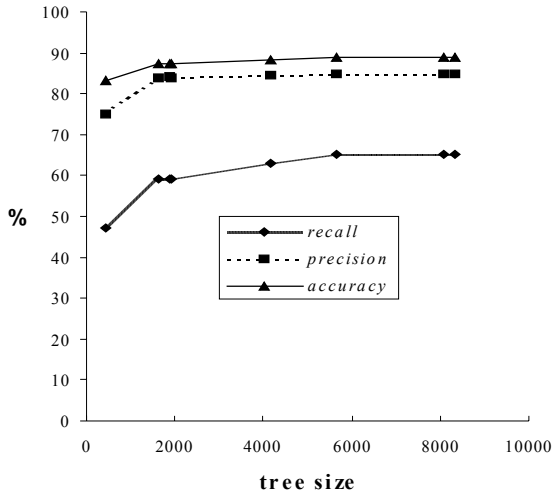

Figure 1: Performance on the general-theme data for different sizes of the decision tree.

Figure 1 plots the three performance measures for different sizes of the tree. Performance continues to increase up to the size of 8,339.6 nodes, which is the unpruned tree. Thus, there does not seem to be overspecialisation on the training data. The large size of the tree is due to the value set of the features: many of the features, such as the collocates take as values all possible collocated words in the text. The result is that when such a feature is used in the tree, the branching factor is very large. Thus, although the tree is of average depth, it is very wide. We are currently investigating the possibility of compressing the tree into a concise set of rules without loss of information.

Sense Disambiguation in Financial News

In the second experiment we restricted the application domain, by choosing financial news articles from the SEMCOR data. The dataset for this experiment consisted of 3,613 word occurrences, of which 1,987 were tagged with WordNet senses, resulting in 753 word occurrences with LDOCE senses and 355 distinct words. The LDOCE polysemy of the dataset is 3,516/753=4.67, which is lower than the general-theme dataset. Another notable difference of the two datasets is in the ratio of word occurrences to distinct words, i.e., the average word repetition. In the general-theme case, this ratio was 600/448=1.34, while in the financial news dataset it is 753/355=2.12. Word repetition is one indication of the richness of the vocabulary in the text. The closer the ratio is to 1, the richer the vocabulary. As expected, the restricted-domain text is poorer in this respect.

Table 3 presents the performance results for the unpruned tree, which now has an average size of 4,342.5 nodes.

recall	*precision*	*accuracy*
72.7%	96.5%	91.5%

Table 3: Results on the financial news data.

The improvement in performance is impressive: 7.5 pp. in recall, 11.6 pp. in precision and 2.6 pp. in accuracy. Furthermore, the size of the unpruned tree is almost half of that in the general-theme case. This is an indication that the number of identifiable repeating patterns has increased by the restriction of the domain. This is intuitive and agrees with the higher word repetition ratio.

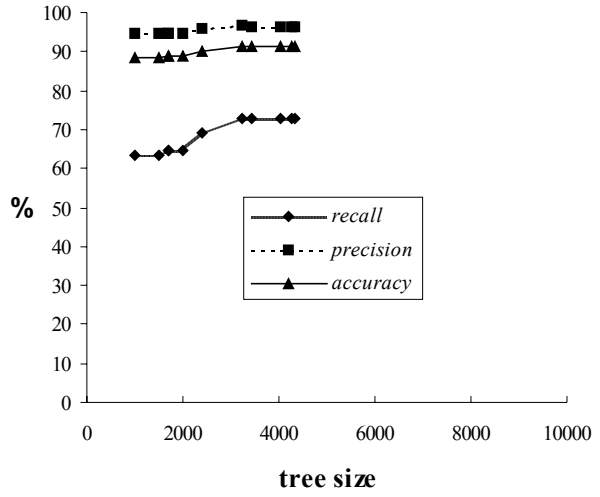

Figure 2: Performance on the financial news data for different sizes of the decision tree.

Figure 2 plots the effect of tree size in this experiment. The curves for precision and accuracy are nearly flat, as in the general-theme experiment. However, recall starts with low values and increases significantly for tree sizes between 2,500 and 3,500 nodes. Once again the size of the tree seems to play an important role in the performance of the method.

Conclusions

In this paper we propose the use of machine learning for the automatic construction of a sense tagger for a particular domain. We used the machine learning algorithm C4.5 to construct a decision-tree tagger for two different test cases. In both cases, we achieve higher results than those previously reported. Furthermore, a significant improvement in performance is observed, when the application domain is restricted. This is an encouraging result that motivates further work in domain-specific WSD with the use of machine learning techniques. A number of issues have arisen in the course of the work presented here and we are currently concentrating our efforts on these issues. First we are examining ways of reducing the size of the decision tree down to a set of concise rules, without any loss in performance. Furthermore, we are looking at the effect that the size of the dataset and the richness of the vocabulary have on performance. Finally, we are planning to test our method on a larger collection of texts and more domains, in order to verify the results that we report here.

Acknowledgements

This research was supported by the European Union Language Engineering project ECRAN (LE2110). We would like to thank Yorick Wilks and Mark Stevenson for the constructive discussions we had, as well as Paola Velardi and Alessandro Cucchiarelli for the provision of the financial news articles from the SEMCOR corpus.

References

R. Bruce and L. Guthrie, 1992. Genus disambiguation: A study in weighted preference. In *Proceedings of COLING-92*, pp. 1187-1191, Nantes, France.

J. Cowie, L. Guthrie and J. Guthrie, 1992. Lexical disambiguation using simulated annealing. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 233-237, Harriman, NY.

A. Kilgarriff, 1993. Dictionary word sense distinctions: An enquiry into their nature. *Computers and the Humanities*, 26:356-387.

M. Lesk, 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of ACM SIGDOC Conference*, pp. 24-26, Toronto, Canada.

J. R. Quinlan, 1993. C4.5: Programs for machine learning, Morgan-Kaufmann, San Mateo, CA.

H. Schuetze, 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92*, pp. 787-796, Minneapolis, MN.

Y. Wilks and M. Stevenson, 1997. Sense tagging: Semantic tagging with a lexicon. In *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics: What, why and how?,"* Washington D.C., April.

D. Yarowsky, 1995. Unsupervised word sense disambiguation rivalling supervised methods. In *Proceedings of ACL-95*, pp. 189-196, Cambridge, MA.