Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources

Yoong Keok Lee and **Hwee Tou Ng** and **Tee Kiah Chia**
Department of Computer Science
National University of Singapore
3 Science Drive 2, Singapore 117543
y.k.lee@alumni.nus.edu.sg
nght@comp.nus.edu.sg
chiateek@comp.nus.edu.sg

Abstract

We participated in the SENSEVAL-3 English lexical sample task and multilingual lexical sample task. We adopted a supervised learning approach with Support Vector Machines, using only the official training data provided. No other external resources were used. The knowledge sources used were part-of-speech of neighboring words, single words in the surrounding context, local collocations, and syntactic relations. For the translation and sense subtask of the multilingual lexical sample task, the English sense given for the target word was also used as an additional knowledge source. For the English lexical sample task, we obtained fine-grained and coarse-grained score (for both recall and precision) of 0.724 and 0.788 respectively. For the multilingual lexical sample task, we obtained recall (and precision) of 0.634 for the translation subtask, and 0.673 for the translation and sense subtask.

1 Introduction

This paper describes the approach adopted by our systems which participated in the English lexical sample task and the multilingual lexical sample task of SENSEVAL-3. The goal of the English lexical sample task is to predict the correct sense of an ambiguous English word w, while that of the multilingual lexical sample task is to predict the correct Hindi (target language) translation of an ambiguous English (source language) word w.

The multilingual lexical sample task is further subdivided into two subtasks: the translation subtask, as well as the translation and sense subtask. The distinction is that for the translation and sense subtask, the English sense of the target ambiguous word w is also provided (for both training and test data).

In all, we submitted 3 systems: system nusels for the English lexical sample task, system nusmlst for the translation subtask, and system nusmlsts for the translation and sense subtask.

All systems were based on the supervised word sense disambiguation (WSD) system of Lee and Ng (2002), and used Support Vector Machines (SVM) learning. Only the training examples provided in the official training corpus were used to train the systems, and no other external resources were used. In particular, we did not use any external dictionary or the sample sentences in the provided dictionary.

The knowledge sources used included part-of-speech (POS) of neighboring words, single words in the surrounding context, local collocations, and syntactic relations, as described in Lee and Ng (2002). For the translation and sense subtask of the multilingual lexical sample task, the English sense given for the target word was also used as an additional knowledge source. All features encoding these knowledge sources were used, without any feature selection.

We next describe SVM learning and the combined knowledge sources adopted. Much of the description follows that of Lee and Ng (2002).

2 Support Vector Machines (SVM)

The SVM (Vapnik, 1995) performs optimization to find a hyperplane with the largest margin that separates training examples into two classes. A test example is classified depending on the side of the hyperplane it lies in. Input features can be mapped into high dimensional space before performing the optimization and classification. A kernel function can be used to reduce the computational cost of training and testing in high dimensional space. If the training examples are nonseparable, a regularization parameter C ($= 1$ by default) can be used to control the trade-off between achieving a large margin and a low training error. We used the implementation of SVM in WEKA (Witten and Frank, 2000), where each nominal feature with n possible values is converted into n binary (0 or 1) features. If a nominal feature takes the ith value, then the ith binary feature is set to 1 and all the other binary features are set to 0. The default linear kernel is used. Since SVM only handles binary (2-class) classification, we built one binary classifier for each sense class.

Note that our supervised learning approach made use of a single learning algorithm, without combining multiple learning algorithms as adopted in other research (such as (Florian et al., 2002)).

3 Multiple Knowledge Sources

To disambiguate a word occurrence w, systems nusels and nusmlst used the first four knowledge sources listed below. System nusmlsts used the English sense given for the target ambiguous word w as an additional knowledge source. Previous research (Ng and Lee, 1996; Stevenson and Wilks, 2001; Florian et al., 2002; Lee and Ng, 2002) has shown that a combination of knowledge sources improves WSD accuracy.

Our experiments on the provided training data of the SENSEVAL-3 translation and sense subtask also indicated that the additional knowledge source of the English sense of the target word further improved accuracy (See Section 4.3 for details).

We did not attempt feature selection since our previous research (Lee and Ng, 2002) indicated that SVM performs better without feature selection.

3.1 Part-of-Speech (POS) of Neighboring Words

We use 7 features to encode this knowledge source: $P_{-3}, P_{-2}, P_{-1}, P_0, P_1, P_2, P_3$, where P_{-i} (P_i) is the POS of the ith token to the left (right) of w, and P_0 is the POS of w. A token can be a word or a punctuation symbol, and each of these neighboring tokens must be in the same sentence as w. We use a sentence segmentation program (Reynar and Ratnaparkhi, 1997) and a POS tagger (Ratnaparkhi, 1996) to segment the tokens surrounding w into sentences and assign POS tags to these tokens.

For example, to disambiguate the word *bars* in the POS-tagged sentence *"Reid/NNP saw/VBD me/PRP looking/VBG at/IN the/DT iron/NN bars/NNS ./."*, the POS feature vector is $< IN, DT, NN, NNS, ., \epsilon, \epsilon >$ where ϵ denotes the POS tag of a null token.

3.2 Single Words in the Surrounding Context

For this knowledge source, we consider all single words (unigrams) in the surrounding context of w, and these words can be in a different sentence from w. For each training or test example, the SENSEVAL-3 official data set provides a few sentences as the surrounding context. In the results reported here, we consider all words in the provided context.

Specifically, all tokens in the surrounding context of w are converted to lower case and replaced by their morphological root forms. Tokens present in a list of stop words or tokens that do not contain at least an alphabet character (such as numbers and punctuation symbols) are removed. All remaining tokens from all training contexts provided for w are gathered. Each remaining token t contributes one feature. In a training (or test) example, the feature corresponding to t is set to 1 iff the context of w in that training (or test) example contains t.

For example, if w is the word *bars* and the set of selected unigrams is $\{$*chocolate, iron, beer*$\}$, the feature vector for the sentence *"Reid saw me looking at the iron bars ."* is <0, 1, 0>.

3.3 Local Collocations

A local collocation $C_{i,j}$ refers to the ordered sequence of tokens in the local, narrow context of w. Offsets i and j denote the starting and ending position (relative to w) of the sequence, where a negative (positive) offset refers to a token to its left (right). For example, let w be the word *bars* in the sentence *"Reid saw me looking at the iron bars ."* Then $C_{-2,-1}$ is *the_iron* and $C_{-1,2}$ is *iron_._*ϵ, where ϵ denotes a null token. Like POS, a collocation does not cross sentence boundary. To represent this knowledge source of local collocations, we extracted 11 features corresponding to the following collocations: $C_{-1,-1}$, $C_{1,1}$, $C_{-2,-2}$, $C_{2,2}$, $C_{-2,-1}$, $C_{-1,1}$, $C_{1,2}$, $C_{-3,-1}$, $C_{-2,1}$, $C_{-1,2}$, and $C_{1,3}$. This set of 11 features is the union of the collocation features used in Ng and Lee (1996) and Ng (1997).

Note that each collocation $C_{i,j}$ is represented by one feature that can have many possible feature values (the local collocation strings), whereas each distinct surrounding word is represented by one feature that takes binary values (indicating presence or absence of that word). For example, if w is the word *bars* and suppose the set of collocations for $C_{-2,-1}$ is $\{$*a_chocolate, the_wine, the_iron*$\}$, then the feature value for collocation $C_{-2,-1}$ in the sentence *"Reid saw me looking at the iron bars ."* is *the_iron*.

3.4 Syntactic Relations

We first parse the sentence containing w with a statistical parser (Charniak, 2000). The constituent tree structure generated by Charniak's parser is then converted into a dependency tree in which every word points to a parent headword. For example, in the sentence *"Reid saw me looking at the iron bars ."*, the word *Reid* points to the parent headword *saw*. Similarly, the word *me* also points to the parent headword *saw*.

We use different types of syntactic relations, depending on the POS of w. If w is a noun, we use four features: its parent headword h, the POS of h, the voice of h (*active*, *passive*, or $\emptyset$ if h is not a verb),

1(a) *attention* (noun) 1(b) He turned his *attention* to the workbench . 1(c) <turned, VBD, active, left>
2(a) *turned* (verb) 2(b) He *turned* his attention to the workbench . 2(c) <he, attention, PRP, NN, VBD, active>
3(a) *green* (adj) 3(b) The modern tram is a *green* machine . 3(c) <machine, NN>

Table 1: Examples of syntactic relations

and the relative position of h from w (whether h is to the *left* or *right* of w). If w is a verb, we use six features: the nearest word l to the left of w such that w is the parent headword of l, the nearest word r to the right of w such that w is the parent headword of r, the POS of l, the POS of r, the POS of w, and the voice of w. If w is an adjective, we use two features: its parent headword h and the POS of h.

Headwords are obtained from a parse tree with the script used for the CoNLL-2000 shared task (Tjong Kim Sang and Buchholz, 2000).[1]

Some examples are shown in Table 1. Each POS noun, verb, or adjective is illustrated by one example. For each example, (a) shows w and its POS; (b) shows the sentence where w occurs; and (c) shows the feature vector corresponding to syntactic relations.

3.5 Source Language (English) Sense

For the translation and sense subtask of the multilingual lexical sample task, the sense of an ambiguous word w in the source language (English) is provided for most of the training and test examples. An example with unknown English sense is denoted with question mark ("?") in the corpus. We treat "?" as another "sense" of w (just like any other valid sense of w).

We compile the set of English senses of a word w encountered in the whole training corpus. For each sense s in this set, a binary feature is generated for each training and test example. If an example has s as the English sense of w, this binary feature (corresponding to s) is set to 1, otherwise it is set to 0.

4 Evaluation

Since our WSD system always outputs exactly one prediction for each test example, its recall is always the same as precision. We report below the micro-averaged recall over all test words.

Evaluation data	Recall
SE-2	0.656
SE-1 (with dictionary examples)	0.796
SE-1 (without dictionary examples)	0.776

Table 2: Micro-averaged, fine-grained recall on SENSEVAL-2 and SENSEVAL-1 test data

System	Recall
nusels	0.724 (fine-grained)
	0.788 (coarse-grained)
nusmlst	0.634
nusmlsts	0.673

Table 3: Micro-averaged recall on SENSEVAL-3 test data

4.1 Evaluation on SENSEVAL-2 and SENSEVAL-1 Data

Before participating in SENSEVAL-3, we evaluated our WSD system on the English lexical sample task of SENSEVAL-2 and SENSEVAL-1. The micro-averaged, fine-grained recall over all SENSEVAL-2 test words and all SENSEVAL-1 test words are given in Table 2.

In SENSEVAL-1, some example sentences are provided with the dictionary entries of the words used in the evaluation. We provide the recall on SENSEVAL-1 test data with and without the use of such additional dictionary examples in training.

On both SENSEVAL-2 and SENSEVAL-1 test data, the accuracy figures we obtained, as reported in Table 2, are higher than the best official test scores reported on both evaluation data sets.

4.2 Official SENSEVAL-3 Scores

We participated in the SENSEVAL-3 English lexical sample task, and both subtasks of the multilingual lexical sample task. The official SENSEVAL-3 scores are shown in Table 3. Each score is the micro-averaged recall (which is the same as precision) over all test words.

According to the task organizers, the fine-grained (coarse-grained) recall of the best participating system in the English lexical sample task is 0.729 (0.795). As such, the performance of our system nusels compares favorably with the best participating system.

We are not able to fully assess the performance of our multilingual lexical sample task systems nusmlst and nusmlsts at the time of writing this paper, since performance figures of the best participating system in this task have not been released by

[1] Available at http://ilk.uvt.nl/∼sabine/chunklink/chunklink_2-2-2000_for_conll.pl

the task organizers.

4.3 Utility of English Sense as an Additional Knowledge Source

To determine if using the English sense as an additional knowledge source improved the accuracy of the translation and sense subtask, we conducted a five-fold cross validation experiment. We randomly divided the training data of the translation and sense subtask for each word into 5 portions, using 4 portions for training and 1 portion for test. We then repeated the process by selecting a different portion as the test data each time and training on the remaining portions.

Our investigation revealed that adding the English sense to the four existing knowledge sources improved the micro-averaged recall from 0.628 to 0.638 on the training data. As such, we decided to use the English sense as an additional knowledge source for our system nusmlsts.

After the official SENSEVAL-3 evaluation ended, we evaluated a variant of our system nusmlsts *without* using the English sense as an additional knowledge source. Based on the official test keys released, the micro-averaged recall drops to 0.643, which seems to suggest that the English sense is a helpful knowledge source for the translation and sense subtask.

5 Conclusion

In this paper, we described our participating systems in the SENSEVAL-3 English lexical sample task and multilingual lexical sample task. Our WSD systems used SVM learning and multiple knowledge sources. Evaluation results on the English lexical sample task indicate that our method achieves good accuracy on this task.

6 Acknowledgements

This research is partially supported by a research grant R252-000-125-112 from National University of Singapore Academic Research Fund.

References

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 132–139.

Radu Florian, Silviu Cucerzan, Charles Schafer, and David Yarowsky. 2002. Combining classifiers for word sense disambiguation. *Natural Language Engineering*, 8(4):327–341.

Yoong Keok Lee and Hwee Tou Ng. 2002. 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)*, pages 41–48.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 40–47.

Hwee Tou Ng. 1997. Exemplar-based word sense disambiguation: Some recent improvements. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 208–213.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142.

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 16–19.

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

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of the CoNLL-2000 and LLL-2000*, pages 127–132.

Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Ian H. Witten and Eibe Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.