NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks

Yee Seng Chan and **Hwee Tou Ng** and **Zhi Zhong**
Department of Computer Science, National University of Singapore
3 Science Drive 2, Singapore 117543
{chanys, nght, zhongzhi}@comp.nus.edu.sg

Abstract

We participated in the SemEval-2007 coarse-grained English all-words task and fine-grained English all-words task. We used a supervised learning approach with SVM as the learning algorithm. The knowledge sources used include local collocations, parts-of-speech, and surrounding words. We gathered training examples from English-Chinese parallel corpora, SEMCOR, and DSO corpus. While the fine-grained sense inventory of WordNet was used to train our system employed for the fine-grained English all-words task, our system employed for the coarse-grained English all-words task was trained with the coarse-grained sense inventory released by the task organizers. Our scores (for both recall and precision) are 0.825 and 0.587 for the coarse-grained English all-words task and fine-grained English all-words task respectively. These scores put our systems in the first place for the coarse-grained English all-words task[1] and the second place for the fine-grained English all-words task.

1 Introduction

In this paper, we describe the systems we developed for the coarse-grained English all-words task and fine-grained English all-words task of SemEval-2007. In the coarse-grained English all-words task, systems have to perform word sense disambiguation (WSD) of all content words (noun, adjective, verb, and adverb) occurring in five documents, using a coarse-grained version of the WordNet sense inventory. In the fine-grained English all-words task, systems have to predict the correct sense of verbs and head nouns of the verb arguments occurring in three documents, according to the fine-grained sense inventory of WordNet.

Results from previous SENSEVAL English all-words task have shown that supervised learning gives the best performance. Further, the best performing system in SENSEVAL-3 English all-words task (Decadt et al., 2004) used training data gathered from multiple sources, highlighting the importance of having a large amount of training data. Hence, besides gathering examples from the widely used SEMCOR corpus, we also gathered training examples from 6 English-Chinese parallel corpora and the DSO corpus (Ng and Lee, 1996).

We developed 2 separate systems; one for each task. For both systems, we performed supervised word sense disambiguation based on the approach of (Lee and Ng, 2002) and using Support Vector Machines (SVM) as our learning algorithm. The knowledge sources used include local collocations, parts-of-speech (POS), and surrounding words. Our system employed for the coarse-grained English all-words task was trained with the coarse-grained sense inventory released by the task organizers, while our system employed for the fine-grained English all-words task was trained with the fine-grained sense

[1] A system developed by one of the task organizers of the coarse-grained English all-words task gave the highest overall score for the coarse-grained English all-words task, but this score is not considered part of the official scores.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 253–256,
Prague, June 2007. ©2007 Association for Computational Linguistics

inventory of WordNet.

In the next section, we describe the different sources of training data used. In Section 3, we describe the knowledge sources used by the learning algorithm. In Section 4, we present our official evaluation results, before concluding in Section 5.

2 Training Corpora

We gathered training examples from parallel corpora, SEMCOR (Miller et al., 1994), and the DSO corpus. In this section, we describe these corpora and how examples gathered from them are combined to form the training data used by our systems. As these data sources use an earlier version of the WordNet sense inventory as compared to the test data of the two tasks we participated in, we also discuss the need to map between different versions of WordNet.

2.1 Parallel Text

Research in (Ng et al., 2003; Chan and Ng, 2005) has shown that examples gathered from parallel texts are useful for WSD. In this evaluation, we gathered training data from 6 English-Chinese parallel corpora (Hong Kong Hansards, Hong Kong News, Hong Kong Laws, Sinorama, Xinhua News, and English translation of Chinese Treebank), available from the Linguistic Data Consortium (LDC). To gather examples from these parallel corpora, we followed the approach in (Ng et al., 2003). Briefly, after ensuring the corpora were sentence-aligned, we tokenized the English texts and performed word segmentation on the Chinese texts (Low et al., 2005). We then made use of the GIZA++ software (Och and Ney, 2000) to perform word alignment on the parallel corpora. Then, we assigned some possible Chinese translations to each sense of an English word *w*. From the word alignment output of GIZA++, we selected those occurrences of *w* which were aligned to one of the Chinese translations chosen. The English side of these occurrences served as training data for *w*, as they were considered to have been disambiguated and "sense-tagged" by the appropriate Chinese translations.

We note that frequently occurring words are usually highly polysemous and hard to disambiguate. To maximize the benefits of using parallel texts, we gathered training data from parallel texts for the set of most frequently occurring noun, adjective, and verb types in the Brown Corpus (BC). These word types (730 nouns, 326 adjectives, and 190 verbs) represent 60% of the noun, adjective, and verb tokens in BC.

2.2 SEMCOR

The SEMCOR corpus (Miller et al., 1994) is one of the few currently available, manually sense-annotated corpora for WSD. It is widely used by various systems which participated in the English all-words task of SENSEVAL-2 and SENSEVAL-3, including one of the top performing teams (Hoste et al., 2001; Decadt et al., 2004) which had performed consistently well in both SENSEVAL all-words tasks. Hence, we also gathered examples from SEMCOR as part of our training data.

2.3 DSO Corpus

Besides SEMCOR, the DSO corpus (Ng and Lee, 1996) also contains manually annotated examples for WSD. As part of our training data, we gathered training examples for each of the 70 verb types present in the DSO corpus.

2.4 Combination of Training Data

Similar to the top performing supervised systems of previous SENSEVAL all-words tasks, we used the annotated examples available from the SEMCOR corpus as part of our training data. In gathering examples from parallel texts, a maximum of 1,000 examples were gathered for each of the frequently occurring noun and adjective types, while a maximum of 500 examples were gathered for each of the frequently occurring verb types. In addition, a maximum of 500 examples were gathered for each of the verb types present in the DSO corpus. For each word, the examples from the parallel corpora and DSO corpus were randomly chosen but adhering to the sense distribution (proportion of each sense) of that word in the SEMCOR corpus.

2.5 Sense Inventory

The test data of the two SemEval-2007 tasks we participated in are based on the WordNet-2.1 sense inventory. However, the examples we gathered from the parallel texts and the SEMCOR corpus are based on the WordNet-1.7.1 sense inventory. Hence, there

is a need to map these examples from WordNet-1.7.1 to WordNet-2.1 sense inventory. For this, we rely primarily on the WordNet sense mappings automatically generated by the work of (Daude et al., 2000). To ensure the accuracy of the mappings, we performed some manual corrections of our own, focusing on the set of most frequently occurring nouns, adjectives, and verbs. For the verb examples from the DSO corpus which are based on the WordNet-1.5 sense inventory, we manually mapped them to WordNet-2.1 senses.

3 WSD System

Following the approach of (Lee and Ng, 2002), we train an SVM classifier for each word using the knowledge sources of local collocations, parts-of-speech (POS), and surrounding words. We omit the syntactic relation features for efficiency reasons. For local collocations, we use 11 features: $C_{-1,-1}$, $C_{1,1}$, $C_{-2,-2}$, $C_{2,2}$, $C_{-2,-1}$, $C_{-1,1}$, $C_{1,2}$, $C_{-3,-1}$, $C_{-2,1}$, $C_{-1,2}$, and $C_{1,3}$, where $C_{i,j}$ refers to the ordered sequence of tokens in the local context of an ambiguous word *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 parts-of-speech, we use 7 features: P_{-3}, P_{-2}, P_{-1}, P_0, P_1, P_2, P_3, where P_0 is the POS of *w*, and P_{-i} (P_i) is the POS of the *i*th token to the left (right) of *w*. For surrounding words, we consider all unigrams (single words) in the surrounding context of *w*. These words can be in a different sentence from *w*.

4 Evaluation

We participated in two tasks of SemEval-2007: the coarse-grained English all-words task and the fine-grained English all-words task. In both tasks, when there is no training data at all for a particular word, we tag all test examples of the word with its first sense in WordNet. Since our systems give exactly one answer for each test example, recall is the same as precision. Hence we will just report the micro-average recall in this section.

4.1 Coarse-Grained English All-Words Task

Our system employed for the coarse-grained English all-words task was trained with the coarse-grained WordNet-2.1 sense inventory released by the task organizers. We obtained a score of 0.825 in this task, as shown in Table 1 under the column $SC + DSO + PT$. It turns out that among the 16 participants of this task, the system which returned the best score was developed by one of the task organizers. Since the score of this system is not considered part of the official scores, our score puts our system in the first position among the participants of this task. For comparison, the WordNet first sense baseline score as calculated by the task organizers is 0.789. To gauge the contribution of parallel text examples, we retrained our system using only examples gathered from the SEMCOR and DSO corpus. As shown in Table 1 under the column $SC + DSO$, this gives a score of 0.817 when scored against the answer keys released by the task organizers. Although adding examples from parallel texts gives only a modest improvement in the scores, we note that this improvement is achieved from a relatively small set of word types which are found to be frequently occurring in BC. Future work can explore expanding the set of word types by automating the process of assigning Chinese translations to each sense of an English word, with the use of suit-

English all-words task	Training data	
	SC+DSO	SC+DSO+PT
Coarse-grained	0.817	0.825
Fine-grained	0.578	0.587

Table 1: Scores for the coarse-grained English all-words task and fine-grained English all-words task, using different sets of training data. SC+DSO refers to using examples gathered from SEMCOR and DSO corpus. Similarly, SC+DSO+PT refers to using examples gathered from SEMCOR, DSO corpus, and parallel texts.

Doc-ID	Recall	No. of test instances
d001	0.883	368
d002	0.881	379
d003	0.834	500
d004	0.761	677
d005	0.814	345

Table 2: Score of each individual test document, for the coarse-grained English all-words task.

able bilingual lexicons.

As part of the evaluation results, the task organizers also released the scores of our system on each of the 5 test documents. We show in Table 2 the score we obtained for each document, along with the total number of test instances in each document. We note that our system obtained a relatively low score on the fourth document, which is a Wikipedia entry on computer programming. To determine the reason for the low score, we looked through the list of test words in that document. We noticed that the noun *program* has 20 test instances occurring in that fourth document. From the answer keys released by the task organizers, all 20 test instances belong to the sense of "a sequence of instructions that a computer can interpret and execute", which we do not have any training examples for. Similarly, we noticed that another noun *programming* has 27 test instances occurring in the fourth document which belong to the sense of "creating a sequence of instructions to enable the computer to do something", which we do not have any training examples for. Thus, these two words alone account for 47 of the errors made by our system in this task, representing 2.1% of the 2,269 test instances of this task.

4.2 Fine-Grained English All-Words Task

Our system employed for the fine-grained English all-words task was trained on examples tagged with fine-grained WordNet-2.1 senses (mapped from WordNet-1.7.1 senses and 1.5 senses as described earlier). Unlike the coarse-grained English all-words task, the correct POS tag and lemma of each test instance are not given in the fine-grained task. Hence, we used the POS tag from the mrg parse files released as part of the test data and performed lemmatization using WordNet. We obtained a score of 0.587 in this task, as shown in Table 1. This ranks our system in second position among the 14 participants of this task. If we exclude parallel text examples and train only on examples gathered from the SEMCOR and DSO corpus, we obtain a score of 0.578.

5 Conclusion

In this paper, we describe the approach taken by our systems which participated in the coarse-grained English all-words task and fine-grained English all-words task of SemEval-2007. Using training examples gathered from parallel texts, SEMCOR, and the DSO corpus, we trained supervised WSD systems with SVM as the learning algorithm. Evaluation results show that this approach achieves good performance in both tasks.

6 Acknowledgements

Yee Seng Chan is supported by a Singapore Millennium Foundation Scholarship (ref no. SMF-2004-1076).

References

Yee Seng Chan and Hwee Tou Ng. 2005. Scaling up word sense disambiguation via parallel texts. In *Proc. of AAAI05*, pages 1037–1042.

Jordi Daude, Lluis Padro, and German Rigau. 2000. Mapping WordNets using structural information. In *Proc. of ACL00*, pages 504–511.

Bart Decadt, Veronique Hoste, Walter Daelemans, and Antal van den Bosch. 2004. GAMBL, genetic algorithm optimization of memory-based WSD. In *Proc. of SENSEVAL-3*, pages 108–112.

Veronique Hoste, Anne Kool, and Walter Daelemans. 2001. Classifi er optimization and combination in the English all words task. In *Proc. of SENSEVAL-2*, pages 83–86.

Yoong Keok Lee and Hwee Tou Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proc. of EMNLP02*, pages 41–48.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to Chinese word segmentation. In *Proc. of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 161–164.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identifi cation. In *Proc. of HLT94 Workshop on Human Language Technology*, pages 240–243.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proc. of ACL96*, pages 40–47.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proc. of ACL03*, pages 455–462.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proc. of ACL00*, pages 440–447.