SemEval-2007 Task 11: English Lexical Sample Task via English-Chinese Parallel Text

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

Abstract

We made use of parallel texts to gather training and test examples for the English lexical sample task. Two tracks were organized for our task. The first track used examples gathered from an LDC corpus, while the second track used examples gathered from a Web corpus. In this paper, we describe the process of gathering examples from the parallel corpora, the differences with similar tasks in previous SENSEVAL evaluations, and present the results of participating systems.

1 Introduction

As part of the SemEval-2007 evaluation exercise, we organized an English lexical sample task for word sense disambiguation (WSD), where the sense-annotated examples were semi-automatically gathered from word-aligned English-Chinese parallel texts. Two tracks were organized for this task, each gathering data from a different corpus. In this paper, we describe our motivation for organizing the task, our task framework, and the results of participants.

Past research has shown that supervised learning is one of the most successful approaches to WSD. However, this approach involves the collection of a large text corpus in which each ambiguous word has been annotated with the correct sense to serve as training data. Due to the expensive annotation process, only a handful of manually sense-tagged corpora are available.

An effort to alleviate the training data bottleneck is the Open Mind Word Expert (OMWE) project (Chklovski and Mihalcea, 2002) to collect sense-tagged data from Internet users. Data gathered through the OMWE project were used in the SENSEVAL-3 English lexical sample task. In that task, WordNet-1.7.1 was used as the sense inventory for nouns and adjectives, while Wordsmyth[1] was used as the sense inventory for verbs.

Another source of potential training data is parallel texts. Our past research in (Ng et al., 2003; Chan and Ng, 2005) has shown that examples gathered from parallel texts are useful for WSD. Briefly, after manually assigning appropriate Chinese translations to each sense of an English word, the English side of a word-aligned parallel text can then serve as the training data, as they are considered to have been disambiguated and "sense-tagged" by the appropriate Chinese translations.

Using the above approach, we gathered the training and test examples for our task from parallel texts. Note that our examples are collected without manually annotating each individual ambiguous word occurrence, allowing us to gather our examples in a much shorter time. This contrasts with the setting of the English lexical sample task in previous SENSEVAL evaluations. In the English lexical sample task of SENSEVAL-2, the sense tagged data were created through manual annotation by trained lexicographers. In SENSEVAL-3, the data were gathered through manual sense annotation by Internet users.

In the next section, we describe in more detail the process of gathering examples from parallel texts and the two different parallel corpora we used. We then give a brief description of each of the partici-

[1] http://www.wordsmyth.net

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 54–58,
Prague, June 2007.

pating systems. In Section 4, we present the results obtained by the participants, before concluding in Section 5.

2 Gathering Examples from Parallel Corpora

To gather examples from 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. The English half of the parallel texts (each ambiguous English word and its 3-sentence context) were used as the training and test material to set up our English lexical sample task.

Note that in our approach, the sense distinction is decided by the different Chinese translations assigned to each sense of a word. This is thus similar to the multilingual lexical sample task in SENSEVAL-3 (Chklovski et al., 2004), except that our training and test examples are collected *without* manually annotating each individual ambiguous word occurrence. The average time needed to assign Chinese translations for one noun and one adjective is 20 minutes and 25 minutes respectively. This is a relatively short time, compared to the effort otherwise needed to manually sense annotate individual word occurrences. Also, once the Chinese translations are assigned, more examples can be automatically gathered as more parallel texts become available.

We note that frequently occurring words are usually highly polysemous and hard to disambiguate. To maximize the benefits of our work, we gathered training data from parallel texts for a set of most frequently occurring noun and adjective types in the Brown Corpus. Also, similar to the SENSEVAL-3 English lexical sample task, we used WordNet-1.7.1 as our sense inventory.

Dataset	Avg. no. of senses	Avg. no. of examples	
		Training	Test
LDC noun	5.2	197.6	98.5
LDC adjective	3.9	125.6	62.9
Web noun	3.5	182.0	91.3
Web adjective	2.8	88.8	44.6

Table 1: Average number of senses, training examples, and test examples per word.

2.1 LDC Corpus

We have two tracks for this task, each track using a different corpus. The first corpus is the Chinese English News Magazine Parallel Text (LDC2005T10), which is an English-Chinese parallel corpus available from the Linguistic Data Consortium (LDC).

From this parallel corpus, we gathered examples for 50 English words (25 nouns and 25 adjectives) using the method described above. From the gathered examples of each word, we randomly selected training and test examples, where the number of training examples is about twice the number of test examples.

The rows *LDC noun* and *LDC adjective* in Table 1 give some statistics about the examples. For instance, each noun has an average of 197.6 training and 98.5 test examples and these examples represent an average of 5.2 senses per noun.[2] Participants taking part in this track need to have access to this LDC corpus in order to access the training and test material in this track.

2.2 Web Corpus

Since not all interested participants may have access to the LDC corpus described in the previous subsection, the second track of this task makes use of English-Chinese documents gathered from the URL pairs given by the STRAND Bilingual Databases.[3] STRAND (Resnik and Smith, 2003) is a system that acquires document pairs in parallel translation automatically from the Web. Using this corpus, we gathered examples for 40 English words (20 nouns and

[2] Only senses present in the examples are counted.

[3] http://www.umiacs.umd.edu/∼resnik/strand

20 adjectives).

The rows *Web noun* and *Web adjective* in Table 1 show that we selected an average of 182.0 training and 91.3 test examples for each noun and these examples represent an average of 3.5 senses per noun. We note that the average number of senses per word for the Web corpus is slightly lower than that of the LDC corpus.

2.3 Annotation Accuracy

To measure the annotation accuracy of examples gathered from the LDC corpus, we examined a random selection of 100 examples each from 5 nouns and 5 adjectives. From these 1,000 examples, we measured a sense annotation accuracy of 84.7%. These 10 words have an average of 8.6 senses per word in the WordNet-1.7.1 sense inventory. As described in (Ng et al., 2003), when several senses of an English word are translated by the same Chinese word, we can collapse these senses to obtain a coarser-grained, lumped sense inventory. If we do this and measure the sense annotation accuracy with respect to a coarser-grained, lumped sense inventory, these 10 words will have an average of 6.5 senses per word and an annotation accuracy of 94.7%.

For the Web corpus, we similarly examined a random selection of 100 examples each from 5 nouns and 5 adjectives. These 10 words have an average of 6.5 senses per word in WordNet-1.7.1 and the 1,000 examples have an average sense annotation accuracy of 85.0%. After sense collapsing, annotation accuracy is 95.3% with an average of 4.8 senses per word.

2.4 Training and Test Data from Different Documents

In our previous work (Ng et al., 2003), we conducted experiments on the nouns of SENSEVAL-2 English lexical sample task. We found that there were cases where the same document contributed both training and test examples and this inflated the WSD accuracy figures. To avoid this, during our preparation of the LDC and Web data, we made sure that a document contributed only either training or test examples, but not both.

3 Participating Systems

Three teams participated in the Web corpus track of our task, with each team employing one system. There were no participants in the LDC corpus track, possibly due to the licensing issues involved. All participating systems employed supervised learning and only used the training examples provided by us.

3.1 CITYU-HIF

The CITYU-HIF team from the City University of Hong Kong trained a naive Bayes (NB) classifier for each target word to be disambiguated, using knowledge sources such as parts-of-speech (POS) of neighboring words and single words in the surrounding context. They also experimented with using different sets of features for each target word.

3.2 HIT-IR-WSD

The system submitted by the HIT-IR-WSD team from Harbin Institute of Technology used Support Vector Machines (SVM) with a linear kernel function as the learning algorithm. Knowledge sources used included POS of surrounding words, local collocations, single words in the surrounding context, and syntactic relations.

3.3 PKU

The system submitted by the PKU team from Peking University used a combination of SVM and maximum entropy classifiers. Knowledge sources used included POS of surrounding words, local collocations, and single words in the surrounding context. Feature selection was done by ignoring word features with certain associated POS tags and by selecting the subset of features based on their entropy values.

4 Results

As all participating systems gave only one answer for each test example, recall equals precision and we will only report micro-average recall on the Web corpus track in this section.

Table 2 gives the overall results obtained by each of the systems when evaluated on all the test examples of the Web corpus. We note that all the participants obtained scores which exceed the baseline heuristic of tagging all test examples with the most

System ID	Contact author	Learning algorithm	Score
HIT-IR-WSD	Yuhang Guo, <astronaut@ir.hit.edu.cn>	SVM	0.819
PKU	Peng Jin, <jandp@pku.edu.cn>	SVM and maximum entropy	0.815
CITYU-HIF	Oi Yee Kwong, <rlolivia@cityu.edu.hk>	NB	0.753
MFS	–	Most frequent sense baseline	0.689

Table 2: Overall micro-average scores of the participants and the most frequent sense (MFS) baseline.

Noun	MFS	CITYU-HIF	HIT-IR-WSD	PKU
age	0.486	0.643	0.743	0.700
area	0.480	0.693	0.773	0.773
body	0.872	0.897	0.910	0.923
change	0.411	0.400	0.578	0.611
director	0.580	0.890	0.960	0.960
experience	0.830	0.830	0.880	0.840
future	0.889	0.889	0.990	0.990
interest	0.308	0.165	0.813	0.780
issue	0.651	0.711	0.892	0.855
life	0.820	0.830	0.860	0.740
material	0.719	0.719	0.781	0.641
need	0.907	0.907	0.918	0.918
performance	0.410	0.570	0.690	0.700
program	0.590	0.590	0.730	0.690
report	0.870	0.840	0.880	0.870
system	0.510	0.700	0.610	0.730
time	0.455	0.673	0.733	0.693
today	0.800	0.750	0.800	0.780
water	0.882	0.921	0.868	0.895
work	0.644	0.743	0.842	0.891
Micro-avg	0.656	0.719	0.813	0.802

Table 3: Micro-average scores of the most frequent sense baseline and the various participants on each noun.

Adjective	MFS	CITYU-HIF	HIT-IR-WSD	PKU
ancient	0.778	0.667	0.778	0.741
bad	0.857	0.857	0.905	0.905
common	0.533	0.567	0.533	0.633
early	0.769	0.846	0.769	0.769
educational	0.911	0.911	0.911	0.911
free	0.760	0.792	0.854	0.917
high	0.630	0.926	0.815	0.852
human	0.872	0.987	0.962	0.962
little	0.450	0.750	0.650	0.650
long	0.667	0.690	0.786	0.714
major	0.870	0.902	0.880	0.913
medical	0.738	0.787	0.800	0.725
national	0.267	0.467	0.667	0.700
new	0.441	0.441	0.529	0.559
present	0.875	0.917	0.875	0.875
rare	0.727	0.818	0.727	0.909
serious	0.879	0.879	0.879	0.879
simple	0.795	0.818	0.864	0.864
small	0.714	0.929	0.893	0.929
third	0.888	0.988	0.963	0.963
Micro-avg	0.757	0.823	0.831	0.842

Table 4: Micro-average scores of the most frequent sense baseline and the various participants on each adjective.

frequent sense (MFS) in the training data. This suggests that the Chinese translations assigned to senses of the ambiguous words are appropriate and provide sense distinctions which are clear enough for effective classifiers to be learned.

In Table 3 and Table 4, we show the scores obtained by each system on each of the 20 nouns and 20 adjectives. For comparison purposes, we also show the corresponding MFS score of each word. Paired t-test on the results of the top two systems show no significant difference between them.

5 Conclusion

We organized an English lexical sample task using examples gathered from parallel texts. Unlike the English lexical task of previous SENSEVAL evaluations where each example is manually annotated, we only need to assign appropriate Chinese translations to each sense of a word. Once this is done, we automatically gather training and test examples from the parallel texts. All the participating systems of our task obtain results that are significantly better than the most frequent sense baseline.

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 *Proceedings of AAAI05*, pages 1037–1042, Pittsburgh, Pennsylvania, USA.

Timothy Chklovski and Rada Mihalcea. 2002. Building a sense tagged corpus with Open Mind Word Expert. In *Proceedings of ACL02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 116–122, Philadelphia, USA.

Timothy Chklovski, Rada Mihalcea, Ted Pedersen, and Amruta Purandare. 2004. The SENSEVAL-3 multilingual English-Hindi lexical sample task. In *Proceedings of SENSEVAL-3*, pages 5–8, Barcelona, Spain.

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

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

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

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.