

Scaling Up Word Sense Disambiguation via Parallel Texts

Yee Seng Chan and Hwee Tou Ng

Department of Computer Science
National University of Singapore
3 Science Drive 2, Singapore 117543
{chanys, nght}@comp.nus.edu.sg

Abstract

A critical problem faced by current supervised WSD systems is the lack of manually annotated training data. Tackling this data acquisition bottleneck is crucial, in order to build high-accuracy and wide-coverage WSD systems. In this paper, we show that the approach of automatically gathering training examples from parallel texts is scalable to a large set of nouns. We conducted evaluation on the nouns of SENSEVAL-2 English all-words task, using fine-grained sense scoring. Our evaluation shows that training on examples gathered from 680MB of parallel texts achieves accuracy comparable to the best system of SENSEVAL-2 English all-words task, and significantly outperforms the baseline of always choosing sense 1 of WordNet.

Introduction

In human languages, many words have multiple meanings, or senses. Word sense disambiguation (WSD) is the process of determining the correct sense of a word in context. WSD is a fundamental problem in natural language processing (NLP), and is important for applications such as machine translation and information retrieval.

One of the most successful approaches to WSD is the use of supervised machine learning. 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 sense-tagged corpora are publicly available.

Among the existing sense-tagged corpora, the SEMCOR corpus (Miller *et al.* 1994) is one of the most widely used. In SEMCOR, content words have been manually tagged with word senses from the WordNet (Miller 1990) sense inventory. Current supervised WSD systems usually rely on this relatively small manually annotated corpus for training examples. However, this has affected the scalability of these systems.

In order to build wide-coverage and scalable WSD systems, tackling this data acquisition bottleneck for WSD is crucial. In an attempt to do this, the DSO corpus (Ng & Lee 1996) was manually annotated, consisting of 192,800 word occurrences of 121 nouns and 70 verbs, which are

Copyright © 2005, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

the most frequently occurring nouns and verbs in English. Chklovski & Mihalcea initiated the Open Mind Word Expert (OMWE) project (Chklovski & Mihalcea 2002) to collect sense-tagged data from Internet users.

Another source of potential training data is parallel texts, as proposed by Resnik & Yarowsky (1997). To explore the potential of this approach, our prior work (Ng, Wang, & Chan 2003) exploited English-Chinese parallel texts for WSD. For each noun of SENSEVAL-2 English lexical sample task (Kilgarriff 2001), we provided some Chinese translations for each of the senses. Senses were lumped together if they were translated in the same way in Chinese. Given a word-aligned English-Chinese parallel corpus, these different Chinese translations will then serve as the “sense-tags” of the corresponding English noun. Encouraging results were obtained in our evaluation on the nouns of SENSEVAL-2 English lexical sample task.

Although parallel text was shown to be a potential source of training data for WSD, some issues remain:

- Evaluation carried out in (Ng, Wang, & Chan 2003) was on a relatively small set of 22 nouns. Can the approach scale up to achieve accurate disambiguation over a large set of nouns?
- Our previous evaluation was on lumped senses. However, evaluation in SENSEVAL-2, such as the English all-words task (Palmer *et al.* 2001), primarily focused on fine-grained scoring. How would the data drawn from parallel texts perform in a fine-grained disambiguation setting?
- The approach requires assigning some possible Chinese translations to each of the senses. Examples are then gathered from parallel texts according to these translations. In practice, would any difficulties arise in the gathering of training examples from parallel texts?

Research needs to be done on the above important issues in order to fully explore the approach of gathering training examples from parallel texts for WSD. In this work, we aim to investigate the issues raised above.

Training Data from Parallel Texts

In this section, we describe the parallel texts used in our experiments, and the process of gathering training data from them.

Parallel corpora	Size of English texts (in million words (MB))	Size of Chinese texts (in million characters (MB))
Hong Kong Hansards	39.9 (223.2)	35.4 (146.8)
Hong Kong News	16.8 (96.4)	15.3 (67.6)
Hong Kong Laws	9.9 (53.7)	9.2 (37.5)
Sinorama	3.8 (20.5)	3.3 (13.5)
Xinhua News	2.1 (11.9)	2.1 (8.9)
English translation of Chinese Treebank	0.1 (0.7)	0.1 (0.4)
Total	72.6 (406.4)	65.4 (274.7)

Table 1: Size of English-Chinese parallel corpora

WordNet sense id	WordNet English sense description	Chinese translations
1	A path over which electrical signals can pass	频道
2	A passage for water	水道, 水渠, 排水渠
3	A long narrow furrow	沟
4	A relatively narrow body of water	海峡
5	A means of communication or access	途径
6	A bodily passage or tube	导管
7	A television station and its programs	频道

Table 2: WordNet sense descriptions and assigned Chinese translations of the noun *channel*

Parallel Text Alignment

We list in Table 1 the 6 English-Chinese parallel corpora (available from Linguistic Data Consortium) used in our experiments. To make use of the parallel corpora, they have to be sentence and word aligned. The sentences of the 6 corpora were already pre-aligned, either manually or automatically, when they were prepared. After ensuring the corpora were sentence aligned, we tokenized the English texts, and performed word segmentation on the Chinese texts. We then made use of the GIZA++ software (Och & Ney 2000) to perform word alignment on the parallel corpora.

Selection of Target Translations

We will assign some possible Chinese translations to each sense of an English word w . In our work, WordNet 1.7 is used as the sense inventory, and the Chinese translations are obtained from 2 sources: Oxford Advanced Learner’s English-Chinese dictionary and Kingsoft Powerword 2003 (Powerword 2003 contains English to Chinese translations of entries in the American Heritage Dictionary).

As an example, WordNet 1.7 lists 7 senses for the noun *channel*, as shown in Table 2. We mapped each of these senses to similar definition entries in the Oxford dictionary and Powerword 2003. The Chinese translations are then gathered from the relevant definition entries of the Oxford dictionary and Powerword 2003. If the same Chinese translation is assigned to several senses, only the least numbered sense will have a valid translation. Note that this is different from the *lumped-sense* approach taken in (Ng, Wang, & Chan 2003). For instance, as shown in Table 2, the same Chinese translation “频道” is assigned to senses 1 and 7 of the noun *channel*. From the word alignment output of GIZA++, all occurrences of the noun *channel* which have been aligned to “频道” will be gathered as training exam-

ples for sense 1 of *channel*. Consequently, there will be no parallel text examples gathered for sense 7 of *channel*. Our work is thus evaluated based on fine-grained scoring.

From the word alignment output of GIZA++, we select those occurrences of the noun *channel* which have been aligned to one of the Chinese translations chosen. The English side of these occurrences will then serve as training data for the noun *channel*, as they are considered to have been disambiguated and “sense-tagged” by the appropriate Chinese translations.

The average time needed to assign target Chinese translations for one noun is 15 minutes. This is a relatively short time, compared to the effort otherwise needed to manually sense annotate training examples. Also, this step could potentially be automated with a suitable bilingual translation lexicon. Once the Chinese translations are assigned, the number of examples gathered will depend only on the size of the parallel texts. More examples could be automatically gathered as more parallel texts become available.

Training and Evaluation

In our experiments, we use the supervised learning approach of Lee & Ng (2002), which achieves state-of-the-art WSD accuracy. The knowledge sources used include parts-of-speech, surrounding words, and local collocations. A naive Bayes classifier is built for each noun.

In the following sections, we first discuss the set of nouns for which we provided Chinese translations, before presenting the evaluation results.

Selection of Nouns Based on Brown Corpus

One of the widely used corpora in NLP research is the Brown corpus (BC). As the BC is built as a balanced cor-

Noun set	No. of noun types	No. of noun tokens	No. of WN _s 1 tokens	WN _s 1 accuracy (%)	Average no. of senses
All nouns	437	1067	767	71.9	4.23
MFS _{set}	212	494	302	61.1	5.89
All – MFS _{set}	225	573	465	81.2	2.67

Table 3: Summary figures on the SENSEVAL-2 English all-words task evaluation data

pus, containing texts in various genres, it is representative of a general text article.

One of the aims of this work is to determine the scalability of parallel text as a source of training examples for WSD. In particular, we are interested in gathering training examples for a large set of nouns. We note that frequently occurring nouns are usually highly polysemous. To maximize the potential benefits from our work, we selected the set of 800 most frequently occurring noun types in the BC, which represents 60% of the noun tokens in the BC. We assigned Chinese translations to this set of 800 nouns and subsequently gathered examples from the parallel texts for each of these 800 nouns.

Evaluation

As discussed, we are interested in performing evaluation using the fine-grained senses of WordNet. After having gathered training examples for a large set of nouns, a suitable evaluation data set would be the set of nouns in the SENSEVAL-2 English all-words task.¹

As shown in Table 3, there are 1067 noun tokens, representing 437 noun types, in the evaluation data of SENSEVAL-2 English all-words task. These 437 nouns have on average 4.23 senses, and a simple strategy of always assigning to each noun its first sense in WordNet (we will subsequently refer to this strategy as *WN_s1*) will give an accuracy of 71.9%. Among the set of 800 nouns for which examples were gathered from parallel texts, 212 of them occurred in the evaluation data of SENSEVAL-2 English all-words task. We will refer to this set of 212 nouns as *MFS_{set}*. As shown in Table 3, this set of nouns is highly polysemous. Their average number of senses is 5.89 and the *WN_s1* strategy gives an accuracy of only 61.1%. In contrast, the average number of senses for the remaining set of 225 nouns not covered by *MFS_{set}* is 2.67 and the corresponding *WN_s1* strategy already gives a high accuracy of 81.2%. These figures conform well with our previous discussion that focusing on gathering training examples for frequently occurring nouns to perform subsequent WSD would be most beneficial.

Having gathered training examples from parallel texts for the *MFS_{set}* set of nouns, we want to measure and compare the performance of a WSD system trained on these parallel text examples, versus training on manually annotated examples.

As mentioned in the introduction section, one of the most widely used manually sense-tagged corpora is the SEMCOR

¹Current evaluation is limited to SENSEVAL-2, for which the answers of participants are publicly available. We are not able to perform evaluation on SENSEVAL-3 English all-words task at this time, since the answers of participating systems are not publicly available.

System	Evaluation set	
	MFS _{set}	All nouns
S1	72.9	78.0
S2	65.4	74.5
S3	64.4	70.0
WN _s 1	61.1	71.9
SC	67.8	76.2
SC+OM	68.4	76.5
P1	69.6	75.8
P2	70.7	76.3
P2jcn	72.7	77.2

Table 4: Evaluation results (accuracy in %)

corpus. As a baseline for comparison, we would like to measure the accuracy of a classifier relying on examples drawn from SEMCOR. We gathered training examples from SEMCOR for the *MFS_{set}* set of nouns. As show in the row SC of Table 4, a classifier trained on examples gathered from SEMCOR and evaluated over the *MFS_{set}* set of nouns achieved an accuracy of 67.8%.

We mentioned that the OMWE project was started in an effort to tackle the data acquisition bottleneck for WSD. Since the motivations behind both the OMWE project and our work are the same, it would be informative to compare the quality of the data collected by the two approaches. However, training examples from OMWE² are only available for a subset of the *MFS_{set}* set of nouns. Thus, we combined training examples from SEMCOR and OMWE as an aggregate set of manually annotated examples. As shown in the row SC+OM of Table 4, a classifier trained on this aggregate set of examples and evaluated over the *MFS_{set}* set of nouns achieved an accuracy of 68.4%.

Next, we measure the accuracy of a classifier trained on examples gathered from parallel texts. For each noun in *MFS_{set}*, we randomly selected a maximum of 500 parallel text examples as training data. A classifier trained and evaluated on this set of *MFS_{set}* nouns achieved an accuracy of 69.6%, as indicated in the P1 row of Table 4.

Although examples are currently gathered from parallel texts for only the *MFS_{set}* set of nouns, we would like an indication of how well our current P1 system trained on these parallel text examples would perform on disambiguating all nouns. Thus, we expanded our evaluation set from *MFS_{set}* to all the nouns in the SENSEVAL-2 English all-words task. However, note that we have only gathered training examples

²We used examples from OMWE 1.0 for our experiments.

from parallel texts for the *MFSet* set of nouns. Therefore, we used the *WNsI* strategy for the set of nouns where parallel text examples are not available, although this presents a disadvantage to our P1 system. As shown under the column *All nouns* of Table 4, the P1 system achieves an accuracy of 75.8% when evaluated over all the nouns in SENSEVAL-2 English all-words task.

We similarly expanded the evaluation set of the SC system from *MFSet* to all the nouns in SENSEVAL-2 English all-words task. We gathered training examples from SEMCOR for all the nouns in the all-words task. For the few nouns where examples are not available from SEMCOR, we used the *WNsI* strategy. As shown in row SC under the column *All nouns* of Table 4, a classifier trained on these examples gathered from SEMCOR obtained an accuracy of 76.2% when evaluated on all the nouns in SENSEVAL-2 English all-words task. Next, the relevant training examples from OMWE are added. A classifier trained on this aggregate set of examples achieved an accuracy of 76.5% when evaluated on all the nouns in the all-words task.

As a comparison, we also tabulated the accuracy figures for the top 3 participating systems in the SENSEVAL-2 English all-words task, from the publicly available set of answers of SENSEVAL-2 participants. The accuracies of these 3 systems are listed in Table 4 as S1 (Mihalcea & Moldovan 2001), S2 (Hoste, Kool, & Daelemans 2001), and S3 (Crestan, El-Beze, & Loupy 2001), arranged in order of performance. The accuracy of the *WNsI* strategy is also shown in Table 4.

Analysis

One problem encountered in gathering training examples from parallel texts is a lack of matching occurrences for some Chinese translations. For instance, the description for sense 7 of the noun *report* is “the general estimation that the public has for a person” and is assigned the Chinese translation “名声”. However, from the word alignment output of GIZA++, no occurrences of the noun *report* are aligned to “名声”. Thus, no training examples are gathered for sense 7 of *report*, affecting the recall of this approach.

One possible solution would be to gather training examples for sense 7 of the noun *report* from *other English nouns* having the same corresponding Chinese translations. For instance, we first note that no training examples are gathered for sense 7 of *report*, and its assigned Chinese translation is “名声”. Searching through our inventory of Chinese translations for other English nouns, we find that the translation “名声” is also assigned to sense 3 of the noun *name*, which has the sense description “a person’s reputation”. Therefore, from the word alignment output of GIZA++, occurrences of the noun *name* which are aligned to “名声” are gathered as training examples for sense 7 of *report*.

Using this approach, we gathered training examples for senses of nouns originally having no training examples. As a large number of examples could be gathered for those senses originally without any examples, we decide to impose a limit. For each noun, we first gather a maximum of 500 examples per sense. Thereafter, a maximum of 500 examples

for each noun are randomly selected as training data. Using this approach, we obtained an accuracy of 70.7% when evaluating on the *MFSet* set of nouns, as shown in the row P2 in Table 4. When evaluation is expanded to all the nouns in SENSEVAL-2 English all-words task by using the *WNsI* strategy for nouns not covered by *MFSet*, we obtained an accuracy of 76.3%.

Although P2 obtained an accuracy improvement over P1, the substitution of training examples of one noun for another might introduce noise into the training data. Hence, there is a need to introduce a weighting factor for the substituting examples. We introduce the weighting scheme used in the following section.

The Jcn Measure

Past research has shown that the semantic distance measure of Jiang & Conrath (1997) provides a reliable estimate of the distance between two concepts (or synsets) in the WordNet hierarchy. They argued that the strength of the link between a child concept c and its parent concept p is proportional to the conditional probability of encountering an instance c given an instance of p , $P(c|p)$.

Following the notation in information theory, the information content (IC) of a concept c can be represented as:

$$IC(c) = -\log P(c)$$

where $P(c)$ can be estimated from the probability of occurrence of c in a large text corpus.

Therefore, the link strength $LS(c, p)$ of an edge that links a child node c to its parent node p can be defined as:

$$LS(c, p) = -\log P(c|p) = IC(c) - IC(p)$$

where the link strength (LS) is simply the difference of the information content values between a child concept and its parent concept.

Hence, the distance between two synsets $s1$ and $s2$ can be measured as follows:

$$Dist(s1, s2) = IC(s1) + IC(s2) - 2 \times IC(lcs(s1, s2))$$

where $lcs(s1, s2)$ denotes the most specific synset (concept) that subsumes both $s1$ and $s2$.

The WordNet Similarity package (Pedersen, Patwardhan, & Michelizzi 2004) supports a range of WordNet similarity measures. In our work, we made use of the WordNet Similarity package to provide a similarity score between WordNet synsets based on the *jcn* measure. In our experiments, the counts required for the calculation of IC by the *jcn* measure is obtained from the British National Corpus (BNC). The package transformed the distance measure $Dist(s1, s2)$ into a *jcn* similarity score by taking the reciprocal:

$$jcn(s1, s2) = 1/Dist(s1, s2)$$

Incorporating the Jcn Measure

In performing WSD with a naive Bayes classifier, the sense s assigned to an example with features f_1, \dots, f_n is chosen so as to maximize:

$$p(s) \prod_{j=1}^n p(f_j|s) = \frac{Count(s)}{N} \prod_{j=1}^n \frac{Count(f_j, s)}{Count(s)}$$

System	S1	P1	P2	P2jcn	SC	SC+OM	WNs1
S1	*	~	~	~	»	>	»
P1		*	~	«	~	~	»
P2			*	<	>	~	»
P2jcn				*	»	>	»
SC					*	~	»
SC+OM						*	»
WNs1							*

Table 5: Paired t-test between the various results under the column *MFS* of Table 4: “~”, (“>” and “<”), and (“»” and “«”) correspond to the p-value > 0.05, (0.01, 0.05], and ≤ 0.01 respectively.

System	S1	P1	P2	P2jcn	SC	SC+OM	WNs1
S1	*	>	~	~	~	~	»
P1		*	~	<	~	~	»
P2			*	~	~	~	»
P2jcn				*	~	~	»
SC					*	~	»
SC+OM						*	»
WNs1							*

Table 6: Paired t-test between the various results under the column *All nouns* of Table 4.

where N is the total number of training examples, $Count(s)$ denotes the number of training examples having sense label s , and the number of such examples with feature value f_j is denoted by $Count(f_j, s)$. Ordinarily, each training example will contribute either a count of 1 or 0 to $Count(f_j, s)$ and $Count(s)$. Now, by incorporating the jcn measure as a weighting scheme, a training example gathered from another English noun based on a common Chinese translation contributes a fractional count $jcn(s1, s2)$. In our earlier example, we substituted sense 3 examples of the noun *name* as examples for sense 7 of the noun *report*. For this particular example, $s1$ and $s2$ refer to sense 7 of *report* and sense 3 of *name*, respectively.

After incorporating the jcn measure into the P2 system, we obtained an accuracy of 72.7% when evaluating on the *MFS* set of nouns, listed in Table 4 as P2jcn. By using the *WNs1* strategy for nouns not covered by *MFS*, we expanded our evaluation to all the nouns in SENSEVAL-2 English all-words task and obtained an accuracy of 77.2%.

Significance Test

Paired t tests were conducted to see if one system is significantly better than another. The t statistic of the difference between each test instance pair is computed, giving rise to a p value. The results of significance tests for the various experiments on the *MFS* set of nouns are given in Table 5. Corresponding results of significance tests when evaluating on all the nouns in SENSEVAL-2 English all-words task are

given in Table 6.

Results in Table 5 show that all systems significantly outperform the baseline of always choosing sense 1 of WordNet. Performance of the best system (S1) in SENSEVAL-2 English all-words task is significantly better than the WSD system (SC) trained on manually annotated examples gathered from SEMCOR. The performance of S1 is also significantly better than the WSD system (SC+OM) trained on an aggregate set of examples gathered from SEMCOR and OMWE. In contrast, the performance of the various WSD systems (P1, P2, P2jcn) trained on examples gathered from parallel texts is comparable to the performance of S1. Also, note that P2 and P2jcn perform significantly better than the SC system. In particular, the P2jcn system, which incorporates the jcn measure and is trained on examples automatically gathered from parallel texts, outperforms the SC+OM system, which is trained on an aggregate set of manually annotated examples. From Table 5, one can also see that the P2jcn system performs significantly better than both P1 and P2, confirming the effectiveness of the approach where examples gathered from other nouns sharing the same Chinese translation are weighted by the jcn measure.

When evaluation is done on all the nouns in SENSEVAL-2 English all-words task, we see from Table 6 that all systems still significantly outperform the baseline of always choosing sense 1 of WordNet. Performance of the SC and SC+OM systems are now comparable with S1. We note that S1 now outperforms the P1 system. However, this is due to the fact that we had gathered training examples from parallel texts only for the *MFS* set of nouns. In moving the parallel text systems (P1, P2, P2jcn) to evaluation on all the nouns in SENSEVAL-2 English all-words task, we had resorted to the strategy of always choosing sense 1 of WordNet for nouns not covered by *MFS*. Nevertheless, note that the performance of P2 and P2jcn are comparable with S1. The previous reason, plus the fact that the systems SC and SC+OM were trained on all available relevant examples from SEMCOR and OMWE, enable the performance of SC and SC+OM to catch up with P2 and P2jcn. Finally, note that the performance of P2jcn is still significantly better than P1, confirming the effectiveness of the jcn approach.

Other Difficulties of the Parallel Text Approach

Our previous work (Ng, Wang, & Chan 2003) alluded to another problem with using parallel text as a source of training examples for WSD. In that work, senses were lumped together if they were translated in the same way in Chinese. This reflects the difficulty of assigning unique Chinese translations to each sense of a noun. In this paper, whenever multiple senses were translated in the same way in Chinese, we assigned a valid translation only to the least numbered sense. Although this will inevitably affect the recall of our approach, this was done in order to evaluate our experiments using fine-grained senses. A possible remedy would be to introduce translations in a third language, so that more senses of an English word could be identified with the parallel text approach.

Related Work

Diab & Resnik (2002) used a machine translated parallel corpus and through an unsupervised method of noun group disambiguation, achieved an accuracy of mid to high 50s when evaluated on the nouns of SENSEVAL-2 English all-words task. This accuracy is much lower than the best system S1 in the all-words task. In recent work, Diab (2004) bootstrapped a supervised WSD system using annotated data produced by the unsupervised approach described in (Diab & Resnik 2002), and evaluated on SENSEVAL-2 English lexical sample task. Building on the work of Diab & Resnik (2002), Bhattacharya, Getoor, & Bengio (2004) built probabilistic models using parallel corpus with an unsupervised approach. Performance on a selected subset of nouns in SENSEVAL-2 English all-words task is promising, but still lags behind the top 3 systems of SENSEVAL-2 English all-words task.

Conclusion

In order to build a wide-coverage WSD system, tackling the data acquisition bottleneck for WSD is crucial. In this paper, we showed that the approach of automatically gathering training examples for WSD from parallel texts is scalable to a large set of nouns. When evaluated on the nouns in SENSEVAL-2 English all-words task using fine-grained scoring, classifiers trained on examples from parallel texts achieve high accuracy, significantly outperforming the strategy of always choosing sense 1 of WordNet. We also explored the various practical problems faced by the parallel text approach, and a solution was provided to improve recall by utilizing examples of nouns having the same assigned Chinese translations. When the *jcn* weighting scheme is incorporated and evaluation is done on the set of nouns for which examples are gathered from parallel texts, classifiers trained on parallel text examples outperform classifiers trained on manually annotated examples drawn from SEMCOR and OMWE. When evaluated on all the nouns of SENSEVAL-2 English all-words task, classifiers trained on parallel text examples achieved performance comparable to the best system of SENSEVAL-2 English all-words task.

Acknowledgements

The first author is supported by a Singapore Millennium Foundation Scholarship (ref no. SMF-2004-1076). This research is also partially supported by a research grant R252-000-125-112 from National University of Singapore Academic Research Fund.

References

Bhattacharya, I.; Getoor, L.; and Bengio, Y. 2004. Unsupervised sense disambiguation using bilingual probabilistic models. In *Proceedings of ACL04*, 287–294.

Chklovski, T., and Mihalcea, R. 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*, 116–122.

Crestan, E.; El-Beze, M.; and Loupy, C. D. 2001. Improving wsd with multi-level view of context monitored by similarity measure. In *Proceedings of SENSEVAL-2*, 67–70.

Diab, M., and Resnik, P. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of ACL02*, 255–262.

Diab, M. 2004. Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of ACL04*, 303–310.

Hoste, V.; Kool, A.; and Daelemans, W. 2001. Classifier optimization and combination in the english all words task. In *Proceedings of SENSEVAL-2*, 83–86.

Jiang, J., and Conrath, D. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of ROCLING X*, 19–33.

Kilgarriff, A. 2001. English lexical sample task description. In *Proceedings of SENSEVAL-2*, 17–20.

Lee, Y. K., and Ng, H. T. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of EMNLP02*, 41–48.

Mihalcea, R., and Moldovan, D. 2001. Pattern learning and active feature selection for word sense disambiguation. In *Proceedings of SENSEVAL-2*, 127–130.

Miller, G. A.; Chodorow, M.; Landes, S.; Leacock, C.; and Thomas, R. G. 1994. Using a semantic concordance for sense identification. In *Proceedings of ARPA Human Language Technology Workshop*, 240–243.

Miller, G. A. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography* 3(4):235–312.

Ng, H. T., and Lee, H. B. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of ACL96*, 40–47.

Ng, H. T.; Wang, B.; and Chan, Y. S. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of ACL03*, 455–462.

Och, F. J., and Ney, H. 2000. Improved statistical alignment models. In *Proceedings of ACL00*, 440–447.

Palmer, M.; Fellbaum, C.; Cotton, S.; Delfs, L.; and Dang, H. T. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of SENSEVAL-2*, 21–24.

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of AAAI04, Intelligent Systems Demonstration*.

Resnik, P., and Yarowsky, D. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of ACL97 SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, 79–86.