Proceedings of the 41st Annual Meeting of the Association
for Computational Linguistics, July 2003, pp. 455-462.

Exploiting Parallel Texts for Word Sense Disambiguation:

An Empirical Study

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

Abstract

A central problem of word sense disambiguation (WSD) is the lack of manually sense-tagged data required for supervised learning. In this paper, we evaluate an approach to automatically acquire sense-tagged training data from English-Chinese parallel corpora, which are then used for disambiguating the nouns in the SENSEVAL-2 English lexical sample task. Our investigation reveals that this method of acquiring sense-tagged data is promising. On a subset of the most difficult SENSEVAL-2 nouns, the accuracy difference between the two approaches is only 14.0%, and the difference could narrow further to 6.5% if we disregard the advantage that manually sense-tagged data have in their sense coverage. Our analysis also highlights the importance of the issue of domain dependence in evaluating WSD programs.

1 Introduction

The task of word sense disambiguation (WSD) is to determine the correct meaning, or sense of a word in context. It is a fundamental problem in natural language processing (NLP), and the ability to disambiguate word sense accurately is important for applications like machine translation, information retrieval, etc.

Corpus-based, supervised machine learning methods have been used to tackle the WSD task, just like the other NLP tasks. Among the various approaches to WSD, the supervised learning approach is the most successful to date. In this approach, we first collect a corpus in which each occurrence of an ambiguous word *w* has been manually annotated with the correct sense, according to some existing sense inventory in a dictionary. This annotated corpus then serves as the training material for a learning algorithm. After training, a model is automatically learned and it is used to assign the correct sense to any previously unseen occurrence of *w* in a new context.

While the supervised learning approach has been successful, it has the drawback of requiring manually sense-tagged data. This problem is particular severe for WSD, since sense-tagged data must be collected separately for each word in a language.

One source to look for potential training data for WSD is parallel texts, as proposed by Resnik and Yarowsky (1997). Given a word-aligned parallel corpus, the different translations in a target language serve as the "sense-tags" of an ambiguous word in the source language. For example, some possible Chinese translations of the English noun *channel* are listed in Table 1. To illustrate, if the sense of an occurrence of the noun *channel* is "a path over which electrical signals can pass", then this occurrence can be translated as "频道" in Chinese.

WordNet 1.7 sense id	Lumped sense id	Chinese translations	WordNet 1.7 English sense descriptions
1	1	频道	A path over which electrical signals can pass
2	2	水道 水渠 排水渠	A passage for water
3	3	沟	A long narrow furrow
4	4	海峡	A relatively narrow body of water
5	5	途径	A means of communication or access
6	6	导管	A bodily passage or tube
7	1	频道	A television station and its programs

Table 1: WordNet 1.7 English sense descriptions, the actual lumped senses, and Chinese translations of the noun *channel* used in our implemented approach

Parallel corpora	Size of English texts (in million words (MB))	Size of Chinese texts (in million characters (MB))
Hong Kong News	5.9 (39.4)	10.7 (22.8)
Hong Kong Laws	7.0 (39.8)	14.0 (22.6)
Hong Kong Hansards	11.9 (54.2)	18.0 (32.4)
English translation of Chinese Treebank	0.1 (0.7)	0.2 (0.4)
Xinhua News	3.6 (22.9)	7.0 (17.0)
Sinorama	3.2 (19.8)	5.3 (10.2)
Total	31.7 (176.8)	55.2 (105.4)

Table 2: Size of English-Chinese parallel corpora

This approach of getting sense-tagged corpus also addresses two related issues in WSD. Firstly, what constitutes a valid sense distinction carries much subjectivity. Different dictionaries define a different sense inventory. By tying sense distinction to the different translations in a target language, this introduces a "data-oriented" view to sense distinction and serves to add an element of objectivity to sense definition. Secondly, WSD has been criticized as addressing an isolated problem without being grounded to any real application. By defining sense distinction in terms of different target translations, the outcome of word sense disambiguation of a source language word is the selection of a target word, which directly corresponds to word selection in machine translation.

While this use of parallel corpus for word sense disambiguation seems appealing, several practical issues arise in its implementation:

(i) What is the size of the parallel corpus needed in order for this approach to be able to disambiguate a source language word accurately?

(ii) While we can obtain large parallel corpora in the long run, to have them manually word-aligned would be too time-consuming and would defeat the original purpose of getting a sense-tagged corpus without manual annotation. However, are current word alignment algorithms accurate enough for our purpose?

(iii) Ultimately, using a state-of-the-art supervised WSD program, what is its disambiguation accuracy when it is trained on a "sense-tagged" corpus obtained via parallel text alignment, compared with training on a manually sense-tagged corpus?

Much research remains to be done to investigate all of the above issues. The lack of large-scale parallel corpora no doubt has impeded progress in this direction, although attempts have been made to mine parallel corpora from the Web (Resnik, 1999).

However, large-scale, good-quality parallel corpora have recently become available. For example, six English-Chinese parallel corpora are

now available from Linguistic Data Consortium. These parallel corpora are listed in Table 2, with a combined size of 280 MB. In this paper, we address the above issues and report our findings, exploiting the English-Chinese parallel corpora in Table 2 for word sense disambiguation. We evaluated our approach on all the nouns in the English lexical sample task of SENSEVAL-2 (Edmonds and Cotton, 2001; Kilgarriff 2001), which used the WordNet 1.7 sense inventory (Miller, 1990). While our approach has only been tested on English and Chinese, it is completely general and applicable to other language pairs.

2 Approach

Our approach of exploiting parallel texts for word sense disambiguation consists of four steps: (1) parallel text alignment (2) manual selection of target translations (3) training of WSD classifier (4) WSD of words in new contexts.

2.1 Parallel Text Alignment

In this step, parallel texts are first sentence-aligned and then word-aligned. Various alignment algorithms (Melamed 2001; Och and Ney 2000) have been developed in the past. For the six bilingual corpora that we used, they already come with sentences pre-aligned, either manually when the corpora were prepared or automatically by sentence-alignment programs. After sentence alignment, the English texts are tokenized so that a punctuation symbol is separated from its preceding word. For the Chinese texts, we performed word segmentation, so that Chinese characters are segmented into words. The resulting parallel texts are then input to the GIZA++ software (Och and Ney 2000) for word alignment.

In the output of GIZA++, each English word token is aligned to some Chinese word token. The alignment result contains much noise, especially for words with low frequency counts.

2.2 Manual Selection of Target Translations

In this step, we will decide on the sense classes of an English word *w* that are relevant to translating *w* into Chinese. We will illustrate with the noun *channel*, which is one of the nouns evaluated in the English lexical sample task of SENSEVAL-2. We rely on two sources to decide on the sense classes of *w*:

(i) The sense definitions in WordNet 1.7, which lists seven senses for the noun *channel*. Two senses are lumped together if they are translated in the same way in Chinese. For example, sense 1 and 7 of *channel* are both translated as "频道" in Chinese, so these two senses are lumped together.

(ii) 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 (as listed in Table 1). These occurrences of the noun *channel* in the English side of the parallel texts are considered to have been disambiguated and "sense-tagged" by the appropriate Chinese translations. Each such occurrence of *channel* together with the 3-sentence context in English surrounding *channel* then forms a training example for a supervised WSD program in the next step.

The average time taken to perform manual selection of target translations for one SENSEVAL-2 English noun is less than 15 minutes. This is a relatively short time, especially when compared to the effort that we would otherwise need to spend to perform manual sense-tagging of training examples. This step could also be potentially automated if we have a suitable bilingual translation lexicon.

2.3 Training of WSD Classifier

Much research has been done on the best supervised learning approach for WSD (Florian and Yarowsky, 2002; Lee and Ng, 2002; Mihalcea and Moldovan, 2001; Yarowsky et al., 2001). In this paper, we used the WSD program reported in (Lee and Ng, 2002). In particular, our method made use of the knowledge sources of part-of-speech, surrounding words, and local collocations. We used naïve Bayes as the learning algorithm. Our previous research demonstrated that such an approach leads to a state-of-the-art WSD program with good performance.

2.4 WSD of Words in New Contexts

Given an occurrence of *w* in a new context, we then used the naïve Bayes classifier to determine the most probable sense of *w*.

noun	No. of senses before lumping	No. of senses after lumping	M1	P1	P1-Baseline	M2	M3	P2	P2-Baseline
child	4	1	-	-	-	-	-	-	-
detention	2	1	-	-	-	-	-	-	-
feeling	6	1	-	-	-	-	-	-	-
holiday	2	1	-	-	-	-	-	-	-
lady	3	1	-	-	-	-	-	-	-
material	5	1	-	-	-	-	-	-	-
yew	2	1	-	-	-	-	-	-	-
bar	13	13	0.619	0.529	0.500	-	-	-	-
bum	4	3	0.850	0.850	0.850	-	-	-	-
chair	4	4	0.887	0.895	0.887	-	-	-	-
day	10	6	0.921	0.907	0.906	-	-	-	-
dyke	2	2	0.893	0.893	0.893	-	-	-	-
fatigue	4	3	0.875	0.875	0.875	-	-	-	-
hearth	3	2	0.906	0.844	0.844	-	-	-	-
mouth	8	4	0.877	0.811	0.846	-	-	-	-
nation	4	3	0.806	0.806	0.806	-	-	-	-
nature	5	3	0.733	0.756	0.522	-	-	-	-
post	8	7	0.517	0.431	0.431	-	-	-	-
restraint	6	3	0.932	0.864	0.864	-	-	-	-
sense	5	4	0.698	0.684	0.453	-	-	-	-
stress	5	3	0.921	0.921	0.921	-	-	-	-
art	4	3	0.722	0.494	0.424	0.678	0.562	0.504	0.424
authority	7	5	0.879	0.753	0.538	0.802	0.800	0.709	0.538
channel	7	6	0.735	0.487	0.441	0.715	0.715	0.526	0.441
church	3	3	0.758	0.582	0.573	0.691	0.629	0.609	0.572
circuit	6	5	0.792	0.457	0.434	0.683	0.438	0.446	0.438
facility	5	3	0.875	0.764	0.750	0.874	0.893	0.754	0.750
grip	7	7	0.700	0.540	0.560	0.655	0.574	0.546	0.556
spade	3	3	0.806	0.677	0.677	0.790	0.677	0.677	0.677

Table 3: List of 29 SENSEVAL-2 nouns, their number of senses, and various accuracy figures

3 An Empirical Study

We evaluated our approach to word sense disambiguation on all the 29 nouns in the English lexical sample task of SENSEVAL-2 (Edmonds and Cotton, 2001; Kilgarriff 2001). The list of 29 nouns is given in Table 3. The second column of Table 3 lists the number of senses of each noun as given in the WordNet 1.7 sense inventory (Miller, 1990).

We first lump together two senses s_1 and s_2 of a noun if s_1 and s_2 are translated into the same Chinese word. The number of senses of each noun after sense lumping is given in column 3 of Table 3. For the 7 nouns that are lumped into one sense (i.e., they are all translated into one Chinese word), we do not perform WSD on these words. The average number of senses before and after sense lumping is 5.07 and 3.52 respectively.

After sense lumping, we trained a WSD classifier for each noun *w*, by using the lumped senses in the manually sense-tagged training data for *w* provided by the SENSEVAL-2 organizers. We then tested the WSD classifier on the official SENSEVAL-2 test data (but with lumped senses) for *w*. The test accuracy (based on fine-grained scoring of SENSEVAL-2) of each noun obtained is listed in the column labeled M1 in Table 3.

We then used our approach of parallel text alignment described in the last section to obtain the training examples from the English side of the parallel texts. Due to the memory size limitation of our machine, we were not able to align all six parallel corpora of 280MB in one alignment run of

GIZA++. For two of the corpora, Hong Kong Hansards and Xinhua News, we gathered all English sentences containing the 29 SENSEVAL-2 noun occurrences (and their sentence-aligned Chinese sentence counterparts). This subset, together with the complete corpora of Hong Kong News, Hong Kong Laws, English translation of Chinese Treebank, and Sinorama, is then given to GIZA++ to perform one word alignment run. It took about 40 hours on our 2.4 GHz machine with 2 GB memory to perform this alignment.

After word alignment, each 3-sentence context in English containing an occurrence of the noun *w* that is aligned to a selected Chinese translation then forms a training example. For each SENSEVAL-2 noun *w*, we then collected training examples from the English side of the parallel texts using the *same* number of training examples for each sense of *w* that are present in the manually sense-tagged SENSEVAL-2 official training corpus (lumped-sense version). If there are insufficient training examples for some sense of *w* from the parallel texts, then we just used as many parallel text training examples as we could find for that sense. We chose the same number of training examples for each sense as the official training data so that we can do a fair comparison between the accuracy of the parallel text alignment approach versus the manual sense-tagging approach.

After training a WSD classifier for *w* with such parallel text examples, we then evaluated the WSD classifier on the same official SENSEVAL-2 test set (with lumped senses). The test accuracy of each noun obtained by training on such parallel text training examples (averaged over 10 trials) is listed in the column labeled P1 in Table 3.

The baseline accuracy for each noun is also listed in the column labeled "P1-Baseline" in Table 3. The baseline accuracy corresponds to always picking the most frequently occurring sense in the training data.

Ideally, we would hope M1 and P1 to be close in value, since this would imply that WSD based on training examples collected from the parallel text alignment approach performs as well as manually sense-tagged training examples. Comparing the M1 and P1 figures, we observed that there is a set of nouns for which they are relatively close. These nouns are: *bar, bum, chair, day, dyke, fatigue, hearth, mouth, nation, nature, post, restraint, sense, stress*. This set of nouns is relatively easy to disambiguate, since using the most-frequently-occurring-sense baseline would have done well for most of these nouns.

The parallel text alignment approach works well for *nature* and *sense*, among these nouns. For *nature*, the parallel text alignment approach gives better accuracy, and for *sense* the accuracy difference is only 0.014 (while there is a relatively large difference of 0.231 between P1 and P1-Baseline of *sense*). This demonstrates that the parallel text alignment approach to acquiring training examples can yield good results.

For the remaining nouns (*art, authority, channel, church, circuit, facility, grip, spade*), the accuracy difference between M1 and P1 is at least 0.10. Henceforth, we shall refer to this set of 8 nouns as "difficult" nouns. We will give an analysis of the reason for the accuracy difference between M1 and P1 in the next section.

4 Analysis

4.1 Sense-Tag Accuracy of Parallel Text Training Examples

To see why there is still a difference between the accuracy of the two approaches, we first examined the quality of the training examples obtained through parallel text alignment. If the automatically acquired training examples are noisy, then this could account for the lower P1 score.

The word alignment output of GIZA++ contains much noise in general (especially for the low frequency words). However, note that in our approach, we only select the English word occurrences that align to our manually selected Chinese translations. Hence, while the complete set of word alignment output contains much noise, the subset of word occurrences chosen may still have high quality sense tags.

Our manual inspection reveals that the annotation errors introduced by parallel text alignment can be attributed to the following sources:

(i) Wrong sentence alignment: Due to erroneous sentence segmentation or sentence alignment, the correct Chinese word that an English word *w* should align to is not present in its Chinese sentence counterpart. In this case, word alignment will align the wrong Chinese word to *w*.

(ii) Presence of multiple Chinese translation candidates: Sometimes, multiple and distinct Chi-

nese translations appear in the aligned Chinese sentence. For example, for an English occurrence *channel*, both "频道" (sense 1 translation) and "途径" (sense 5 translation) happen to appear in the aligned Chinese sentence. In this case, word alignment may erroneously align the wrong Chinese translation to *channel*.

(iii) Truly ambiguous word: Sometimes, a word is truly ambiguous in a particular context, and different translators may translate it differently. For example, in the phrase "the church meeting", *church* could be the physical building sense (教堂), or the institution sense (教会). In manual sense tagging done in SENSEVAL-2, it is possible to assign two sense tags to *church* in this case, but in the parallel text setting, a particular translator will translate it in one of the two ways (教堂 or 教会), and hence the sense tag found by parallel text alignment is only one of the two sense tags.

By manually examining a subset of about 1,000 examples, we estimate that the sense-tag error rate of training examples (tagged with lumped senses) obtained by our parallel text alignment approach is less than 1%, which compares favorably with the quality of manually sense tagged corpus prepared in SENSEVAL-2 (Kilgarriff, 2001).

4.2 Domain Dependence and Insufficient Sense Coverage

While it is encouraging to find out that the parallel text sense tags are of high quality, we are still left with the task of explaining the difference between M1 and P1 for the set of difficult nouns. Our further investigation reveals that the accuracy difference between M1 and P1 is due to the following two reasons: domain dependence and insufficient sense coverage.

Domain Dependence The accuracy figure of M1 for each noun is obtained by training a WSD classifier on the manually sense-tagged training data (with lumped senses) provided by SENSEVAL-2 organizers, and testing on the corresponding official test data (also with lumped senses), both of which come from similar domains. In contrast, the P1 score of each noun is obtained by training the WSD classifier on a mixture of six parallel corpora, and tested on the official SENSEVAL-2 test set, and hence the training and test data come from dissimilar domains in this case.

Moreover, from the "docsrc" field (which records the document id that each training or test example originates) of the official SENSEVAL-2 training and test examples, we realized that there are many cases when some of the examples from a document are used as training examples, while the rest of the examples from the *same* document are used as test examples. In general, such a practice results in higher test accuracy, since the test examples would look a lot closer to the training examples in this case.

To address this issue, we took the official SENSEVAL-2 training and test examples of each noun *w* and combined them together. We then randomly split the data into a new training and a new test set such that no training and test examples come from the same document. The number of training examples in each sense in such a new training set is the same as that in the official training data set of *w*.

A WSD classifier was then trained on this new training set, and tested on this new test set. We conducted 10 random trials, each time splitting into a different training and test set but ensuring that the number of training examples in each sense (and thus the sense distribution) follows the official training set of *w*. We report the average accuracy of the 10 trials. The accuracy figures for the set of difficult nouns thus obtained are listed in the column labeled M2 in Table 3.

We observed that M2 is *always* lower in value compared to M1 for all difficult nouns. This suggests that the effect of training and test examples coming from the same document has inflated the accuracy figures of SENSEVAL-2 nouns.

Next, we randomly selected 10 sets of training examples from the parallel corpora, such that the number of training examples in each sense followed the official training set of *w*. (When there were insufficient training examples for a sense, we just used as many as we could find from the parallel corpora.) In each trial, after training a WSD classifier on the selected parallel text examples, we tested the classifier on the same test set (from SENSEVAL-2 provided data) used in that trial that generated the M2 score. The accuracy figures thus obtained for all the difficult nouns are listed in the column labeled P2 in Table 3.

Insufficient Sense Coverage We observed that there are situations when we have insufficient training examples in the parallel corpora for some

of the senses of some nouns. For instance, no occurrences of sense 5 of the noun *circuit* (racing circuit, a racetrack for automobile races) could be found in the parallel corpora. To ensure a fairer comparison, for each of the 10-trial manually sense-tagged training data that gave rise to the accuracy figure M2 of a noun *w*, we extracted a new subset of 10-trial (manually sense-tagged) training data by ensuring adherence to the number of training examples found for each sense of *w* in the corresponding parallel text training set that gave rise to the accuracy figure P2 for *w*. The accuracy figures thus obtained for the difficult nouns are listed in the column labeled M3 in Table 3. M3 thus gave the accuracy of training on manually sense-tagged data but restricted to the number of training examples found in each sense from parallel corpora.

4.3 Discussion

The difference between the accuracy figures of M2 and P2 averaged over the set of all difficult nouns is **0.140**. This is smaller than the difference of **0.189** between the accuracy figures of M1 and P1 averaged over the set of all difficult nouns. This confirms our hypothesis that eliminating the possibility that training and test examples come from the same document would result in a fairer comparison.

In addition, the difference between the accuracy figures of M3 and P2 averaged over the set of all difficult nouns is **0.065**. That is, eliminating the advantage that manually sense-tagged data have in their sense coverage would reduce the performance gap between the two approaches from 0.140 to 0.065. Notice that this reduction is particularly significant for the noun *circuit*. For this noun, the parallel corpora do not have enough training examples for sense 4 and sense 5 of *circuit*, and these two senses constitute approximately 23% in each of the 10-trial test set.

We believe that the remaining difference of 0.065 between the two approaches could be attributed to the fact that the training and test examples of the manually sense-tagged corpus, while not coming from the same document, are however still drawn from the same general domain. To illustrate, we consider the noun *channel* where the difference between M3 and P2 is the largest. For *channel*, it turns out that a substantial number of the training and test examples contain the collocation "Channel tunnel" or "Channel Tunnel". On average, about 9.8 training examples and 6.2 test examples contain this collocation. This alone would have accounted for 0.088 of the accuracy difference between the two approaches.

That domain dependence is an important issue affecting the performance of WSD programs has been pointed out by (Escudero et al., 2000). Our work confirms the importance of domain dependence in WSD.

As to the problem of insufficient sense coverage, with the steady increase and availability of parallel corpora, we believe that getting sufficient sense coverage from larger parallel corpora should not be a problem in the near future for most of the commonly occurring words in a language.

5 Related Work

Brown et al. (1991) is the first to have explored statistical methods in word sense disambiguation in the context of machine translation. However, they only looked at assigning at most two senses to a word, and their method only asked a single question about a single word of context. Li and Li (2002) investigated a bilingual bootstrapping technique, which differs from the method we implemented here. Their method also does not require a parallel corpus.

The research of (Chugur et al., 2002) dealt with sense distinctions across multiple languages. Ide et al. (2002) investigated word sense distinctions using parallel corpora. Resnik and Yarowsky (2000) considered word sense disambiguation using multiple languages. Our present work can be similarly extended beyond bilingual corpora to multilingual corpora.

The research most similar to ours is the work of Diab and Resnik (2002). However, they used machine translated parallel corpus instead of human translated parallel corpus. In addition, they used an unsupervised method of noun group disambiguation, and evaluated on the English all-words task.

6 Conclusion

In this paper, we reported an empirical study to evaluate an approach of automatically acquiring sense-tagged training data from English-Chinese parallel corpora, which were then used for disambiguating the nouns in the SENSEVAL-2 English lexical sample task. Our investigation reveals that

this method of acquiring sense-tagged data is promising and provides an alternative to manual sense tagging.

Acknowledgements

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

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 264-270.

Irina Chugur, Julio Gonzalo, and Felisa Verdejo. 2002. Polysemy and sense proximity in the Senseval-2 test suite. In *Proceedings of the ACL SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 32-39.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 255-262.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, pages 1-5.

Gerard Escudero, Lluis Marquez, and German Rigau. 2000. An empirical study of the domain dependence of supervised word sense disambiguation systems. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 172-180.

Radu Florian and David Yarowsky. 2002. Modeling consensus: Classifier combination for word sense disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 25-32.

Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54-60.

Adam Kilgarriff. 2001. English lexical sample task description. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, pages 17-20.

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 2002 Conference on Empirical Methods in Natural Language Processing*, pages 41-48.

Cong Li and Hang Li. 2002. Word translation disambiguation using bilingual bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 343-351.

I. Dan Melamed. 2001. *Empirical Methods for Exploiting Parallel Texts.* MIT Press, Cambridge.

Rada F. Mihalcea and Dan I. Moldovan. 2001. Pattern learning and active feature selection for word sense disambiguation. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, pages 127-130.

George A. Miller. (Ed.) 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235-312.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440-447.

Philip Resnik. 1999. Mining the Web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527-534.

Philip Resnik and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pages 79-86.

Philip Resnik and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113-133.

David Yarowsky, Silviu Cucerzan, Radu Florian, Charles Schafer, and Richard Wicentowski. 2001. The Johns Hopkins SENSEVAL2 system descriptions. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, pages 163-166.