Word Sense Disambiguation Using Sense Examples Automatically Acquired from a Second Language

Xinglong Wang
School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh
EH8 9LW, UK
xwang@inf.ed.ac.uk

John Carroll
Department of Informatics
University of Sussex
Falmer, Brighton
BN1 9QH, UK
johnca@sussex.ac.uk

Abstract

We present a novel almost-unsupervised approach to the task of Word Sense Disambiguation (WSD). We build sense examples automatically, using large quantities of Chinese text, and English-Chinese and Chinese-English bilingual dictionaries, taking advantage of the observation that mappings between words and meanings are often different in typologically distant languages. We train a classifier on the sense examples and test it on a gold standard English WSD dataset. The evaluation gives results that exceed previous state-of-the-art results for comparable systems. We also demonstrate that a little manual effort can improve the quality of sense examples, as measured by WSD accuracy. The performance of the classifier on WSD also improves as the number of training sense examples increases.

1 Introduction

The results of the recent Senseval-3 competition (Mihalcea et al., 2004) have shown that supervised WSD methods can yield up to 72.9% accuracy[1] on words for which manually sense-tagged data are available. However, supervised methods suffer from the so-called knowledge acquisition bottleneck: they need large quantities of high quality annotated data to produce reliable results. Unfortunately, very few sense-tagged corpora are available and manual sense-tagging is extremely costly and labour intensive. One way to tackle this problem is trying to automate the sense-tagging process. For example, Agirre et al. (2001) proposed a method for building topic signatures automatically, where a topic signature is a set of words, each associated with some weight, that tend to co-occur with a certain concept. Their system queries an Internet search engine with monosemous synonyms of words that have multiple senses in WordNet (Miller et al., 1990), and then extracts topic signatures by processing text snippets returned by the search engine. They trained a classifier on the topic signatures and evaluated it on a WSD task, but the results were disappointing.

In recent years, WSD approaches that exploit differences between languages have shown great promise. Several trends are taking place simultaneously under this multilingual paradigm. A classic one is to acquire sense examples using bilingual parallel texts (Gale et al., 1992; Resnik and Yarowsky, 1997; Diab and Resnik, 2002; Ng et al., 2003): 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, Ng et al. (2003) acquired sense examples using English-Chinese parallel corpora, which were manually or automatically aligned at sentence level and then word-aligned using software. A manual selection of target translations was then performed, grouping together senses that share the same translation in Chinese. Finally, the occurrences of the word on the English side of the parallel

[1] This figure refers to the highest accuracy achieved in the Senseval-3 English Lexical Sample task with fine-grained scoring.

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 547–554, Vancouver, October 2005.

texts were considered to have been disambiguated and "sense tagged" by the appropriate Chinese translations. A classifier was trained on the extracted sense examples and then evaluated on the nouns in Senseval-2 English Lexical Sample dataset. The results appear good numerically, but since the sense groups are not in the gold standard, comparison with other Senseval-2 results is difficult. As discussed by Ng et al., there are several problems with relying on bilingual parallel corpora for data collection. First, parallel corpora, especially accurately aligned parallel corpora are rare, although attempts have been made to mine them from the Web (Resnik, 1999). Second, it is often not possible to distinguish all senses of a word in the source language, by merely relying on parallel corpora, especially when the corpora are relatively small. This is a common problem for bilingual approaches: useful data for some words cannot be collected because different senses of polysemous words in one language often translate to the same word in the other. Using parallel corpora can aggravate this problem, because even if a word sense in the source language has a unique translation in the target language, the translation may not occur in the parallel corpora at all, due to the limited size of this resource.

To alleviate these problems, researchers seek other bilingual resources such as bilingual dictionaries, together with monolingual resources that can be obtained easily. Dagan and Itai (1994) proposed an approach to WSD using monolingual corpora, a bilingual lexicon and a parser for the source language. One of the problems of this method is that for many languages, accurate parsers do not exist. With a small amount of classified data and a large amount of unclassified data in both the source and the target languages, Li and Li (2004) proposed bilingual bootstrapping. This repeatedly constructs classifiers in the two languages in parallel and boosts the performance of the classifiers by classifying data in each of the languages and by exchanging information regarding the classified data between two languages. With a certain amount of manual work, they reported promising results, but evaluated on relatively small datasets.

In previous work, we proposed to use Chinese monolingual corpora and Chinese-English bilingual dictionaries to acquire sense examples (Wang, 2004)[2]. We evaluated the sense examples using a vector space WSD model on a small dataset containing words with binary senses, with promising results. This approach does not rely on scarce resources such as aligned parallel corpora or accurate parsers.

This paper describes further progress based on our proposal: we automatically build larger-scale sense examples and then train a Naïve Bayes classifier on them. We have evaluated our system on the English Lexical Sample Dataset from Senseval-2 and the results show conclusively that such sense examples can be used successfully in a full-scale fine-grained WSD task. We tried to analyse whether more sense examples acquired this way would improve WSD accuracy and also whether a little human effort on sense mapping could further improve WSD performance.

The reminder of the paper is organised as follows. Section 2 outlines the acquisition algorithm for sense examples. Section 3 describes details of building this resource and demonstrates our application of sense examples to WSD. We also present results and analysis in this section. Finally, we conclude in Section 4 and talk about future work.

2 Acquisition of Sense Examples

Following our previous proposal (Wang, 2004), we automatically acquire English sense examples using large quantities of Chinese text and English-Chinese and Chinese-English dictionaries. The Chinese language was chosen because it is a distant language from English and the more distant two languages are, the more likely that senses are lexicalised differently (Resnik and Yarowsky, 1999). The underlying assumption of this approach is that in general each sense of an ambiguous English word corresponds to a distinct translation in Chinese. As shown in Figure 1, firstly, the system translates senses of an English word into Chinese words, using an English-Chinese dictionary, and then retrieves text snippets from a large amount of Chinese text, with the Chinese translations as queries. Then, the Chinese text snippets are segmented and then translated back to English word by word, using a Chinese-English dic-

[2]Sense examples were referred to as "topic signatures" in that paper.

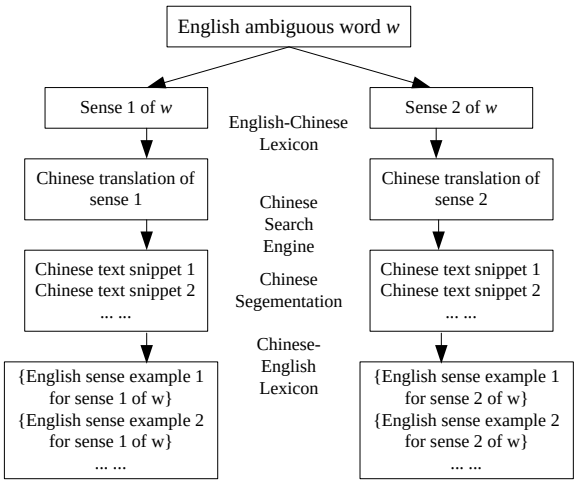

Figure 1. Process of automatic acquisition of sense examples. For simplicity, assume w has two senses.

tionary. In this way, for each sense, a set of sense examples is produced. As an example, suppose one wants to retrieve sense examples for the *financial* sense of *interest*. One first looks up the Chinese translations of this sense in an English-Chinese dictionary, and finds that 利息 is the right Chinese translation corresponding to this particular sense. Then, the next stage is to automatically build a collection of Chinese text snippets by either searching in a large Chinese corpus or on the Web, using 利息 as query. Since Chinese is a language written without spaces between words, one needs to use a segmentor to mark word boundaries before translating the snippets word by word back to English. The result is a collection of sense examples for the *financial* sense of *interest*, each containing a bag of words that tend to co-occur with that particular sense. For example, {*interest rate, bank, annual, economy, ...*} might be one of the sense examples extracted for the *financial* sense of *interest*. Note that words in a sense example are unordered.

Since this method acquires training data for WSD systems from raw monolingual Chinese text, it avoids the problem of the shortage of English sense-tagged corpora, and also of the shortage of aligned bilingual corpora. Also, if existing corpora are not big enough, one can always harvest more text from the Web. However, like all methods based on the cross-language translation assumption mentioned above, there are potential problems. For example, it is possible that a Chinese translation of an English sense is also ambiguous, and thus the contents of text snippets retrieved may be regarding a concept other than the one we want. In general, when the assumption does not hold, one could use the *glosses* defined in a dictionary as queries to retrieve text snippets, as comprehensive bilingual dictionaries tend to include translations to all senses of a word, where multiword translations are used when one-to-one translation is not possible. Alternatively, a human annotator could map the senses and translations by hand. As we will describe later in this paper, we chose the latter way in our experiments.

3 Experiments and Results

We firstly describe in detail how we prepared the sense examples and then describe a large scale WSD evaluation on the English Senseval-2 Lexical Sample dataset (Kilgarriff, 2001). The results show that our system trained with the sense examples achieved significantly better accuracy than comparable systems. We also show that when a little manual effort was invested in mapping the English word senses to Chinese monosemous translations, WSD performance improves accordingly. Based on further experiments on a standard binary WSD dataset, we also show that the technique scales up satisfactorily so that more sense examples help achieve better WSD accuracy.

3.1 Building Sense Examples

Following the approach described in Section 2, we built sense examples for the 44 words in the Senseval-2 dataset[3]. These 44 words have 223 senses in total to disambiguate. The first step was translating English senses to Chinese. We used the *Yahoo! Student English-Chinese On-line Dictionary*[4], as well as a more comprehensive electronic dictionary. This is because the *Yahoo!* dictionary is designed for English learners, and its sense granularity is rather coarse-grained. It is good enough for words with fewer or coarse-grained senses. How-

[3]These 44 words cover all nouns and adjectives in the Senseval-2 dataset, but exclude verbs. We discuss this point in section 3.2.

[4]See: http://cn.yahoo.com/dictionary.

ever, the Senseval-2 Lexical Sample task[5] uses WordNet 1.7 as gold standard, which has very fine sense distinctions and translation granularity in the *Yahoo!* dictionary does not conform to this standard. *PowerWord 2002*[6] was chosen as a supplementary dictionary because it integrates several comprehensive English-Chinese dictionaries in a single application. For each sense of an English word entry, both *Yahoo!* and *PowerWord 2002* dictionaries list not only Chinese translations but also English glosses, which provides a bridge between WordNet synsets and Chinese translations in the dictionaries. In detail, to automatically find a Chinese translation for sense s of an English word w, our system looks up w in both dictionaries and determines whether w has the same or greater number of senses as in WordNet. If it does, in one of the bilingual dictionaries, we locate the English gloss g which has the maximum number of overlapping words with the gloss for s in the WordNet synset. The Chinese translation associated with g is then selected. Although this simple method successfully identified Chinese translations for 23 out of the 44 words (52%), translations for the remaining word senses remain unknown because the sense distinctions are different between our bilingual dictionaries and WordNet. In fact, unless an English-Chinese bilingual WordNet becomes available, this problem is inevitable. For our experiments, we solved the problem by manually looking up dictionaries and identifying translations. For each one of the 44 words, *PowerWord 2002* provides more Chinese translations than the number of its synsets in WordNet 1.7. Thus the annotator simply selects the Chinese translations that he considers a best match to the corresponding English senses. This task took an hour for an annotator who speaks both languages fluently.

It is possible that the Chinese translations are also ambiguous, which can make the topic of a collection of text snippets deviate from what is expected. For example, the *oral* sense of *mouth* can be translated as 口 or 口腔 in Chinese. However, the first translation (口) is a single-character word and is highly ambiguous: by combining with other characters, its meaning varies. For example, 出口 means "an exit" or "to export". On the other hand, the second translation (口腔) is monosemous and should be used. To assess the influence of such "ambiguous translations", we carried out experiments involving more human labour to verify the translations. The same annotator manually eliminated those highly ambiguous Chinese translations and then replaced them with less ambiguous or ideally monosemous Chinese translations. This process changed roughly half of the translations and took about five hours. We compared the basic system with this manually improved one. The results are presented in section 3.2.

Using translations as queries, the sense examples were automatically extracted from *the Chinese Gigaword Corpus* (CGC), distributed by the LDC[7], which contains 2.7GB newswire text, of which 900MB are sourced from *Xinhua News Agency of Beijing*, and 1.8GB are drawn from *Central News* from Taiwan. A small percentage of words have different meanings in these two Chinese dialects, and since the Chinese-English dictionary (*LDC Mandarin-English Translation Lexicon Version 3.0*) we use later is compiled with Mandarin usages in mind, we mainly retrieve data from *Xinhua News*. We set a threshold of 100, and only when the amount of snippets retrieved from *Xinhua News* is smaller than 100, do we turn to *Central News* to collect more data. Specifically, for 48 out of the 223 (22%) Chinese queries, the system retrieved less than 100 instances from *Xinhua News* so it extracted more data from *Central News*. In theory, if the training data is still not enough, one could always turn to other text resources, such as the Web.

To decide the optimal length of text snippets to retrieve, we carried out pilot experiments with two length settings: 250 ($\approx$ 110 English words) and 400 (≈ 175 English words) Chinese characters, and found that more context words helped improve WSD performance (results not shown). Therefore, we retrieve text snippets with a length of 400 characters.

We then segmented all text snippets, using an application *ICTCLAS*[8]. After the segmentor marked

[5]The task has two variations: one to disambiguate fine-grained senses and the other to coarse-grained ones. We evaluated our sense examples on the former variation, which is obviously more difficult.

[6]A commercial electronic dictionary application. We used the free on-line version at: http://cb.kingsoft.com.

[7]Available at: http://www.ldc.upenn.edu/Catalog/

[8]See: http://mtgroup.ict.ac.cn/~zhp/ICTCLAS

all word boundaries, the system automatically translated the text snippets word by word using the electronic *LDC Mandarin-English Translation Lexicon 3.0*. As expected, the lexicon does not cover all Chinese words. We simply discarded those Chinese words that do not have an entry in this lexicon. We also discarded those Chinese words with multiword English translations. Since the discarded words can be informative, one direction of our research in the future is to find an up-to-date wide coverage dictionary, and to see how much difference it will make. Finally, we filtered the sense examples with a stop-word list, to ensure only content words were included.

We ended up with 223 sets of sense examples for all senses of the 44 nouns and adjectives in the test dataset. Each sense example contains a set of words that were translated from a Chinese text snippet, whose content should closely relate to the English word sense in question. Words in a sense example are unordered, because in this work we only used bag-of-words information. Except for the very small amount of manual work described above to map WordNet glosses to those in English-Chinese dictionaries, the whole process is automatic.

3.2 WSD Experiments on Senseval-2 Lexical Sample dataset

The Senseval-2 English Lexical Sample Dataset consists of manually sense-tagged training and test instances for nouns, adjectives and verbs. We only tested our system on nouns and adjectives because verbs often have finer sense distinctions, which would mean more manual work would need to be done when mapping WordNet synsets to English-Chinese dictionary glosses. This would involve us in a rather different kind of enterprise since we would have moved from an almost-unsupervised to a more supervised setup.

We did not use the training data supplied with the dataset. Instead, we train a classifier on our automatically built sense examples and test it on the test data provided. In theory, any machine learning classifier can be applied. We chose the Naïve Bayes algorithm with kernel estimation[9] (John and Langley, 1995) which outperformed a few other classifiers in our pilot experiments on other datasets (results not shown). The average length of a sense example is 35 words, which is much shorter than the length of the text snippets, which was set to 400 Chinese characters ($\approx$ 175 English words). This is because function words and words that are not listed in the *LDC Mandarin-English* lexicon were eliminated. We did not apply any weighting to the features because performance went down in our pilot experiments when we applied a TF.IDF weighting scheme (results not shown). We also limited the maximum number of

[9]We used the implementation in the Weka machine learning package, available at: http://www.cs.waikato.ac.nz/~ml/weka.

	Sys A		Sys B		Baselines & A Senseval-2 Entry			
Word	**Basic**	**MW**	**Basic**	**MW**	**RB**	**MFB**	**Lesk(U)**	**UNED**
art-n(5)	29.9	51.0	39.8	59.6	16.3	41.8	16.3	50.0
authority-n(7)	20.7	22.8	21.5	23.7	10.0	39.1	30.4	34.8
bar-n(13)	41.1	48.3	44.7	52.0	3.3	38.4	2.0	27.8
blind-a(3)	74.5	74.5	74.5	75.0	40.0	78.2	32.7	74.5
bum-n(4)	60.0	60.0	64.4	62.2	15.6	68.9	53.3	11.1
chair-n(4)	81.2	82.6	80.0	82.9	23.2	76.8	56.5	81.2
channel-n(7)	31.5	35.6	32.4	36.5	12.3	13.7	21.9	17.8
child-n(4)	56.3	56.3	56.3	56.3	18.8	54.7	56.2	43.8
church-n(3)	53.1	56.3	59.4	59.4	29.7	56.2	45.3	62.5
circuit-n(6)	48.2	68.2	48.8	69.8	10.6	27.1	5.9	55.3
colourless-a(2)	45.7	45.7	66.7	69.4	42.9	65.7	54.3	31.4
cool-a(6)	26.9	26.9	50.9	50.9	13.5	46.2	9.6	46.2
day-n(9)	32.2	32.9	32.4	33.1	7.6	60.0	0	20.0
detention-n(2)	62.5	84.4	60.6	84.8	43.8	62.5	43.8	78.1
dyke-n(2)	85.7	85.7	82.8	86.2	28.6	53.6	57.1	35.7
facility-n(5)	20.7	22.4	27.1	28.8	13.8	48.3	46.6	25.9
faithful-a(3)	69.6	69.6	66.7	66.7	21.7	78.3	26.1	78.3
fatigue-n(4)	76.7	74.4	77.3	77.3	25.6	76.7	44.2	86.0
feeling-n(6)	11.7	11.7	50.0	50.0	9.8	56.9	2.0	60.8
fine-a(9)	8.6	11.4	34.3	32.9	7.1	42.9	5.7	44.3
fit-a(3)	44.8	44.8	44.8	44.8	31.0	58.6	3.4	48.3
free-a(8)	29.3	37.8	37.3	48.2	15.9	35.4	7.3	35.4
graceful-a(2)	58.6	58.6	70.0	73.3	62.1	79.3	72.4	79.3
green-a(7)	53.2	58.5	53.2	58.5	21.3	75.5	10.6	78.7
grip-n(7)	35.3	37.3	35.3	37.3	19.6	35.3	17.6	21.6
hearth-n(3)	46.9	50.0	48.5	51.5	31.2	71.9	81.2	65.6
holiday-n(2)	64.5	74.2	64.5	75.0	38.7	77.4	29.0	54.8
lady-n(3)	69.2	73.6	71.7	77.8	28.3	64.2	50.9	58.5
local-a(3)	36.8	42.1	38.5	43.6	26.3	55.3	31.6	34.2
material-n(5)	39.1	44.9	47.8	49.3	10.1	20.3	44.9	53.6
mouth-n(8)	38.3	41.7	38.3	41.7	11.7	36.7	31.7	48.3
nation-n(3)	35.1	35.1	39.5	39.5	21.6	78.4	18.9	70.3
natural-a(10)	14.6	32.0	14.6	34.0	6.8	27.2	6.8	44.7
nature-n(5)	23.9	26.1	27.7	31.9	15.2	45.7	41.3	23.9
oblique-a(2)	72.4	72.4	72.4	73.3	44.8	69.0	72.4	27.6
post-n(8)	34.7	45.6	34.7	45.6	10.1	31.6	6.3	41.8
restraint-n(6)	6.7	6.7	17.4	19.6	11.1	28.9	28.9	17.8
sense-n(5)	20.7	43.4	25.9	46.3	24.5	24.5	24.5	30.2
simple-a(7)	45.5	45.5	49.3	50.7	13.6	51.5	12.1	51.5
solemn-a(2)	64.0	76.0	73.1	76.9	32.0	96.0	24.0	96.0
spade-n(3)	66.7	63.6	67.6	70.6	18.2	63.6	60.6	54.5
stress-n(5)	37.5	37.5	45.0	45.0	12.8	48.7	2.6	20.5
vital-a(4)	42.1	42.1	41.0	46.2	21.1	92.1	0	94.7
yew-n(2)	21.4	25.0	82.8	89.7	57.1	78.6	17.9	71.4
Avg.	**40.7**	**46.0**	**46.1**	**52.0**	**18.1**	**50.5**	**24.6**	**46.4**

Table 1. WSD accuracy on words in the English Senseval-2 Lexical Sample dataset. The left most column shows words, their POS tags and how many senses they have. "Sys A" and "Sys B" are our systems, and "MW" denotes a multi-word detection module was used in conjunction with the "Basic" system. For comparison, it also shows two baselines: "RB" is the random baseline and "MFB" is the most-frequent-sense baseline. "UNED" is one of the best unsupervised participants in the Senseval-2 competition and "Lesk(U)" is the highest unsupervised-baseline set in the workshop. All accuracies are expressed as percentages.

training sense examples to 6000, for efficiency purposes. We attempted to tag every test data instance, so our coverage (on nouns and adjectives) is 100%.

To assess the influence of ambiguous Chinese translations, we prepared two sets of training data. As described in section 3.1: sense examples in the first set were prepared without taking ambiguity in Chinese text into consideration, while those in the second set were prepared with a little more human effort involved trying to reduce ambiguity by using less ambiguous translations. We call the system trained on the first set "Sys A" and the one trained on the second "Sys B".

In this lexical sample task, multiwords are expected to be picked out by participating WSD systems. For example, the answer *art collection* should be supplied when this multiword occurs in a test instance. It would be judged wrong if one tagged the *art* in *art collection* as the *artworks* sense, even though one could argue that this was also a correct answer. To deal with multiwords, we implemented a very simple detection module, which tries to match multiword entries in WordNet to the ambiguous word and its left and right neighbours. For example, if the module finds *art collection* is an entry in WordNet, it tags all occurrences of this multiword in the test data, regardless of the prediction by the classifier.

The results are shown in Table 1. Our "Sys B" system, with and without the multiword detection module, outperformed "Sys A", which shows that sense examples acquired with less ambiguous Chinese translations contain less noise and therefore boost WSD performance. For comparison, the table also shows various baseline performance figures and a system that participated in Senseval-2[10]. Considering that the manual work involved in our approach is negligible compared with manual sense-tagging, we classify our systems as unsupervised and we should aim to beat the random baseline. This all four of our systems do easily. We also easily beat another unsupervised baseline – the Lesk (1986) baseline, which disambiguates words using WordNet definitions. The MFB baseline is actually a 'supervised' baseline, since an unsupervised system does not have such prior knowledge beforehand. McCarthy et al. (2004) argue that this is a very tough baseline for an unsupervised WSD system to beat. Our "Sys B" with multiword detection exceeds it. "Sys B" also exceeds the performance of UNED (Fernández-Amorós et al., 2001), which was the second-best ranked[11] unsupervised systems in the Senseval-2 competition.

There are a number of factors that can influence WSD performance. The distribution of training data for senses is one. In our experiments, we used all sense examples that we built for a sense (with an upper bound of 6000). However, the distribution of senses in English text often does not match the distribution of their corresponding Chinese translations in Chinese text. For example, suppose an English word w has two senses: s_1 and s_2, where s_1 rarely occurs in English text, whereas sense s_2 is used frequently. Also suppose s_1's Chinese translation is much more frequently used than s_2's translation in Chinese text. Thus, the distribution of the two senses in English is different from that of the translations in Chinese. As a result, the numbers of sense examples we would acquire for the two senses would be distributed as if they were in Chinese text. A classifier trained on this data would then tend to predict unseen test instances in favour of the wrong distribution. The word *nation*, for example, has three senses, of which the *country* sense is used more frequently in English. However, in Chinese, the *country* sense and the *people* sense are almost equally distributed, which might be the reason for its WSD accuracy being lower with our systems than most of the other words. A possible way to alleviate this problem is to select training sense examples according to an estimated distribution in natural English text, which can be done by analysing available sense-tagged corpora with help of smoothing techniques, or with the unsupervised approach of (McCarthy et al., 2004).

Cultural differences can cause difficulty in retrieving sufficient training data. For example, translations of senses of *church* and *hearth* appear only infrequently in Chinese text. Thus, it is hard to build sense examples for these words. Another problem,

[10] Accuracies for each word and averages were calculated by us, based on the information on Senseval-2 Website. See: http://www.sle.sharp.co.uk/senseval2/.

[11] One system performed better but their answers were not on the official Senseval-2 website so that we could not do the comparison. Also, that system did not attempt to disambiguate as many words as UNED and us.

as mentioned above, is that translations of English senses can be ambiguous in Chinese. For example, Chinese translations of the words *vital*, *natural*, *local* etc. are also ambiguous to some extent, and this might be a reason for their low performance. One way to solve this, as we described, is to manually check the translations. Another automatic way is that, before retrieving text snippets, we could segment or even parse the Chinese corpora, which should reduce the level of ambiguity and lead to better sense examples.

3.3 Further WSD Experiments

One of the strengths of our approach is that training data come cheaply and relatively easily. However, the sense examples are acquired automatically and they inevitably contain a certain amount of noise, which may cause problems for the classifier. To assess the relationship between accuracy and the size of training data, we carried out a series of experiments, feeding the classifier with different numbers of sense examples as training data.

For these experiments, we used another standard WSD dataset, the TWA dataset. This is a manually sense-tagged corpus (Mihalcea, 2003), which contains 2-way sense-tagged text instances, drawn from the British National Corpus, for 6 nouns. We first built sense examples for all the 12 senses using the approach described above, then trained the same Naïve Bayes algorithm (NB) on different numbers of sense examples.

In detail, for all of the 6 words, we did the following: given a word w_i, we randomly selected n sense examples for each of its senses s_i, from the total amount of sense examples built for s_i. Then the NB algorithm was trained on the $2 * n$ examples and tested on w_i's test instances in TWA. We recorded the accuracy and repeated this process 200 times and calculated the mean and variance of the 200 accuracies. Then we assigned another value to n and iterated the above process until n took all the predefined values. In our experiments, n was taken from $\{50, 100, 150, 200, 400, 600, 800, 1000, 1200\}$ for words *motion*, *plant* and *tank* and from $\{50, 100, 150, 200, 250, 300, 350\}$ for *bass*, *crane* and *palm*, because there were less sense example data available for the latter three words. Finally, we used the t-test ($p = 0.05$) on pairwise sets of means and variances to see if improvements were statistically significant.

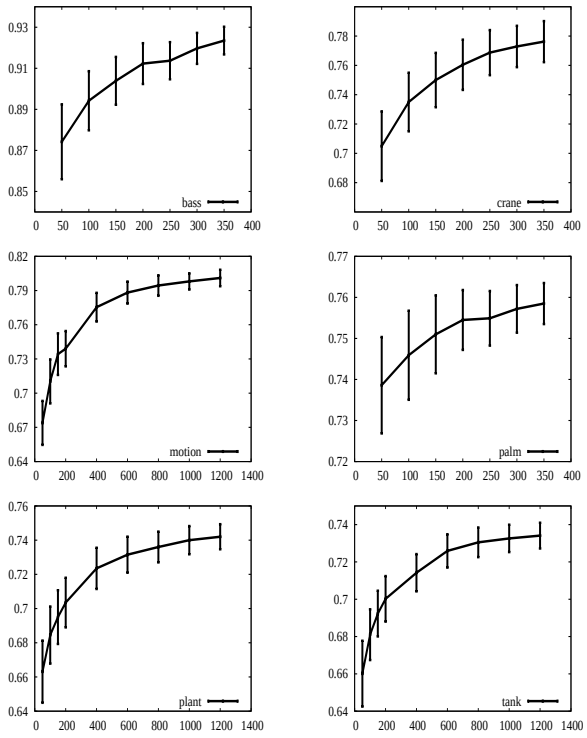

Figure 2. Accuracy scores with increasing number of training sense examples. Each bar is a standard deviation.

The results are shown in Figure 2[12]. 34 out of 42 t-scores are greater than the t-test critical values, so we are fairly confident that the more training sense examples used, the more accurate the NB classifier becomes on this disambiguation task.

4 Conclusions and Future Work

We have presented WSD systems that use sense examples as training data. Sense examples are acquired automatically from large quantities of Chinese text, with the help of Chinese-English and English-Chinese dictionaries. We have tested our WSD systems on the English Senseval-2 Lexical Sample dataset, and our best system outperformed comparable state-of-the-art unsupervised systems. Also, we found that increasing the number of the sense examples significantly improved WSD performance. Since sense examples can be obtained very cheaply from any large Chinese text collection, in-

[12]These experiments showed that our systems outperformed the most-frequent-sense baseline and Mihalcea's unsupervised system (2003).

cluding the Web, our approach is a way to tackle the knowledge acquisition bottleneck.

There are a number of future directions that we could investigate. Firstly, instead of using a bilingual dictionary to translate Chinese text snippets back to English, we could use machine translation software. Secondly, we could try this approach on other language pairs, Japanese-English, for example. This is also a possible solution to the problem that ambiguity may be preserved between Chinese and English. In other words, when a Chinese translation of an English sense is still ambiguous, we could try to collect sense examples using translation in a third language, Japanese, for instance. Thirdly, it would be interesting to try to tackle the problem of Chinese WSD using sense examples built using English, the reverse process to the one described in this paper.

Acknowledgements

This research was funded by EU IST-2001-34460 project MEANING: Developing Multilingual Web-Scale Language Technologies.

References

Eneko Agirre, Olatz Ansa, David Martinez, and Eduard Hovy. 2001. Enriching WordNet concepts with topic signatures. In *Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. Pittsburgh, USA.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the* 40^{th} *Anniversary Meeting of the Association for Computational Linguistics (ACL-02)*. Philadelphia, USA.

David Fernández-Amorós, Julio Gonzalo, and Felisa Verdejo. 2001. The UNED systems at Senseval-2. In *Poceedings of Second International Wordshop on Evaluating Word Sense Disambiguation Systems (Senseval-2)*. Toulouse, France.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101–112.

George H. John and Pat Langley. 1995. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345.

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

Michael E. Lesk. 1986. Automated sense disambiguation using machine-readable dictionaries: how to tell a pinecone from an ice cream cone. In *Proceedings of the SIGDOC Conference*.

Hang Li and Cong Li. 2004. Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, 20(4):563–596.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. Barcelona, Spain.

Rada Mihalcea, Timothy Chklovski, and Adam Killgariff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*.

Rada Mihalcea. 2003. The role of non-ambiguous words in natural language disambiguation. In *Proceedings of the Conference on Recent Advances in Natural Language Processing, RANLP 2003*. Borovetz, Bulgaria.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: an empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

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. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Philip Resnik. 1999. Mining the Web for bilingual text. In *Proceedings of the* 37^{th} *Annual Meeting of the Association for Computational Linguistics*.

Xinglong Wang. 2004. Automatic acquisition of English topic signatures based on a second language. In *Proceedings of the Student Research Workshop at ACL 2004*. Barcelona, Spain.