Word Sense Disambiguation Using Automatically Translated Sense Examples

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

David Martinez
Department of Computer Science
University of Sheffield
Sheffield, S1 4DP, UK
davidm@dcs.shef.ac.uk

Abstract

We present an unsupervised approach to Word Sense Disambiguation (WSD). We automatically acquire English sense examples using an English-Chinese bilingual dictionary, Chinese monolingual corpora and Chinese-English machine translation software. We then train machine learning classifiers on these sense examples and test them on two gold standard English WSD datasets, one for binary and the other for fine-grained sense identification. On binary disambiguation, performance of our unsupervised system has approached that of the state-of-the-art supervised ones. On multi-way disambiguation, it has achieved a very good result that is competitive to other state-of-the-art unsupervised systems. Given the fact that our approach does not rely on manually annotated resources, such as sense-tagged data or parallel corpora, the results are very promising.

1 Introduction

Results from recent Senseval workshops have shown that supervised Word Sense Disambiguation (WSD) systems tend to outperform their unsupervised counterparts. However, supervised systems rely on large amounts of accurately sense-annotated data to yield good results and such resources are very costly to produce. It is difficult for supervised WSD systems to perform well and reliably on words that do not have enough sense-tagged training data. This is the so-called knowledge acquisition bottleneck.

To overcome this bottleneck, unsupervised WSD approaches have been proposed. Among them, systems under the multilingual paradigm have shown great promise (Gale et al., 1992; Dagan and Itai, 1994; Diab and Resnik, 2002; Ng et al., 2003; Li and Li, 2004; Chan and Ng, 2005; Wang and Carroll, 2005). The underlying hypothesis is that mappings between word forms and meanings can be different from language to language. Much work have been done on extracting sense examples from parallel corpora for WSD. For example, Ng et al. (2003) proposed to train a classifier on sense examples acquired from word-aligned English-Chinese parallel corpora. They grouped senses that share the same Chinese translation, and then the occurrences of the word on the English side of the parallel corpora were considered to have been disambiguated and "sense tagged" by the appropriate Chinese translations. Their system was evaluated on the nouns in Senseval-2 English lexical sample dataset, with promising results. Their follow-up work (Chan and Ng, 2005) has successfully scaled up the approach and achieved very good performance on the Senseval-2 English all-word task.

Despite the promising results, there are problems with relying on parallel corpora. For example, there is a lack of matching occurrences for some Chinese translations to English senses. Thus gathering training examples for them might be difficult, as reported in (Chan and Ng, 2005). Also, parallel corpora themselves are rare resources and not available for many language pairs.

Some researchers seek approaches using monolingual resources in a second language and then try to map the two languages using bilingual dictionaries. For example, Dagan and Itai (1994) carried out WSD experiments using monolingual corpora, a bilingual lexicon and a parser for the source language. One problem of this method is that

for many languages, accurate parsers do not exist. Wang and Carroll (2005) proposed to use monolingual corpora and bilingual dictionaries to automatically acquire sense examples. Their system was unsupervised and achieved very promising results on the Senseval-2 lexical sample dataset. Their system also has better portability, i.e., it runs on any language pair as long as a bilingual dictionary is available. However, sense examples acquired using the dictionary-based word-by-word translation can only provide "bag-of-words" features. Many other features useful for machine learning (ML) algorithms, such as the ordering of words, part-of-speech (POS), bigrams, etc., have been lost. It could be more interesting to translate Chinese text snippets using machine translation (MT) software, which would provide richer contextual information that might be useful for WSD learners. Although MT systems themselves are expensive to build, once they are available, they can be used repeatedly to automatically generate as much data as we want. This is an advantage over relying on other expensive resources such as manually sense-tagged data and parallel copora, which are limited in size and producing additional data normally involves further costly investments.

We carried out experiments on acquiring sense examples using both MT software and a bilingual dictionary. When we had the two sets of sense examples ready, we trained a ML classifier on them and then tested them on coarse-grained and fine-grained gold standard WSD datasets, respectively. We found that on both test datasets the classifier using MT translated sense examples outperformed the one using those translated by a dictionary, given the same amount of training examples used on each word sense. This confirms our assumption that a richer feature set, although from a noisy data source, such as machine translated text, might help ML algorithms. In addition, both systems performed very well comparing to other state-of-the-art WSD systems. As we expected, our system is particularly good on coarse-grained disambiguation. Being an unsupervised approach, it achieved a performance competitive to state-of-the-art supervised systems.

This paper is organised as follows: Section 2 revisits the process of acquiring sense examples proposed in (Wang and Carroll, 2005) and then describes our adapted approach. Section 3 outlines resources, the ML algorithm and evaluation metrics that we used. Section 4 and Section 5 detail experiments we carried out on gold standard datasets. We also report our results and error analysis. Finally, Section 6 concludes the paper and draws future directions.

2 Acquisition of Sense Examples

Wang and Carroll (2005) proposed an automatic approach to acquire sense examples from large amount of Chinese text and English-Chinese and Chinese-English dictionaries. The acquisition process is summarised as follows:

1. Translate an English ambiguous word w to Chinese, using an English-Chinese lexicon. Given the assumption that mappings between words and senses are different between English and Chinese, each sense s_i of w maps to a distinct Chinese word. At the end of this step, we have produced a set C, which consists of Chinese words $\{c_1, c_2, ..., c_n\}$, where c_i is the translation corresponding to sense s_i of w, and n is the number of senses that w has.
2. Query large Chinese corpora or/and a search engine using each element in C. For each c_i in C, we collect the text snippets retrieved and construct a Chinese corpus.
3. Word-segment these Chinese text snippets.
4. Use an electronic Chinese-English lexicon to translate the Chinese corpora constructed word by word to English.

This process can be completely automatic and unsupervised. However, in order to compare the performance against other WSD systems, one needs to map senses in the bilingual dictionary to those used by gold standard datasets, which are often from WordNet (Fellbaum, 1998). This step is inevitable unless we use senses in the bilingual dictionary as gold standard. Fortunately, the mapping process only takes a very short time[1], comparing to the effort that it would take to manually sense annotate training examples. At the end of the acquisition process, for each sense s_i of an ambiguous word w, we have a large set of English contexts. Note that a context is represented by a bag of words only. We mimicked this process and built a set of sense examples.

To obtain a richer set of features, we adapted the above process and carried out another acquisition experiment. When translating Chinese text snippets to English in the 4th step, we used MT software instead of a bilingual dictionary. The intuition is that although machine translated text contains noise, features like word ordering, POS tags

[1] A similar process took 15 minutes per noun as reported in (Chan and Ng, 2005), and about an hour for 20 nouns as reported in (Wang and Carroll, 2005).

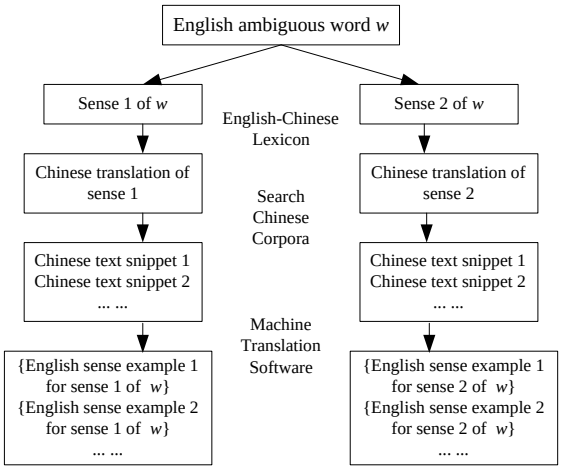

Figure 1:Adapted process of automatic acquisition of sense examples. For simplicity, assume w has two senses.

and bigrams/trigrams may still be of some use for ML classifiers. In this approach, the 3rd step can be omitted, since MT software should be able to take care of segmentation. Figure 1 illustrates our adapted acquisition process.

As described above, we prepared two sets of training examples for each English word sense to disambiguate: one set was translated word-by-word by looking up a bilingual dictionary, as proposed in (Wang and Carroll, 2005), and the other translated using MT software. In detail, we first mapped senses of ambiguous words, as defined in the gold-standard TWA (Mihalcea, 2003) and Senseval-3 lexical sample (Mihalcea et al., 2004) datasets (which we use for evaluation) onto their corresponding Chinese translations. We did this by looking up an English-Chinese dictionary *PowerWord 2002*[2]. This mapping process involved human intervention, but it only took an annotator (fluent speaker in both Chinese and English) 4 hours. Since some Chinese translations are also ambiguous, which may affect WSD performance, the annotator was asked to select the Chinese words that are relatively unambiguous (or ideally monosemous) in Chinese for the target word senses, when it was possible. Sometimes multiple senses of an English word can map to the same Chinese word, according to the English-Chinese dictionary. In such cases, the annotator was advised to try to capture the subtle difference between these English word senses and then to select different Chinese translations for them, using his knowledge on the languages. Then, using the translations as queries, we retrieved as many text snippets as possible from *the Chinese Gigaword Corpus*. For efficiency purposes, we randomly chose maximumly 200 text snippets for each sense, when acquiring data for nouns and adjectives from Senseval-3 lexical sample dataset. The length of the snippets was set to 400 Chinese characters.

From here we prepared two sets of sense examples differently. For the approach of dictionary-based translation, we segmented all text snippets, using the application *ICTCLAS*[3]. After the segmentor marked all word boundaries, the system automatically translated the text snippets word by word using the electronic *LDC Mandarin-English Translation Lexicon 3.0*. All possible translations of each word were included. 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. Finally we got a set of sense examples for each sense. Note that a sense example produced here is simply a bag of words without ordering.

We prepared the other set of sense examples by translating text snippets with the MT software *Systran 5.0 Standard*, where each example contains much richer features that potentially can be exploited by ML algorithms.

3 Experimental Settings

3.1 Training

We applied the Vector Space Model (VSM) algorithm on the two different kinds of sense examples (i.e., dictionary translated ones vs. MT software translated ones), as it has been shown to perform well with the features described below (Agirre and Martinez, 2004a). In VSM, we represent each context as a vector, where each feature has an 1 or 0 value to indicate its occurrence or absence. For each sense in training, a centroid vector is obtained, and these centroids are compared to the vectors that represent test examples, by means of the cosine similarity function. The closest centroid assigns its sense to the test example.

For the sense examples translated by MT software, we analysed the sentences using different

[2]PowerWord is a commercial electronic dictionary application. There is a free online version at: http://cb.kingsoft.com.

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

tools and extracted relevant features. We applied stemming and POS tagging, using the fnTBL toolkit (Ngai and Florian, 2001), as well as shallow parsing[4]. Then we extracted the following types of topical and domain features[5], which were then fed to the VSM machine learner:

- Topical features: we extracted lemmas of the content words in two windows around the target word: the whole context and a ± 4 word window. We also obtained salient bigrams in the context, with the methods and the software described in (Pedersen, 2001). We included another feature type, which match the closest words (for each POS and in both directions) to the target word (e.g. *LEFT NOUN "dog"* or *LEFT VERB "eat"*).

- Domain features: The "WordNet Domains" resource was used to identify the most relevant domains in the context. Following the relevance formula presented in (Magnini and Cavagliá, 2000), we defined two feature types: (1) the most relevant domain, and (2) a list of domains above a threshold[6].

For the dictionary-translated sense examples, we simply used bags of words as features.

3.2 Evaluation

We evaluated our WSD classifier on both coarse-grained and fine-grained datasets. For coarse-grained WSD evaluation, we used TWA dataset (Mihalcea, 2003), which is a binarily sense-tagged corpus drawn from the British National Corpus (BNC), for 6 nouns. For fine-grained evaluation, we used Senseval-3 English lexical sample dataset (Mihalcea et al., 2004), which comprises 7,860 sense-tagged instances for training and 3,944 for testing, on 57 words (nouns, verbs and adjectives). The examples were mainly drawn from BNC. WordNet 1.7.1[7] was used as sense inventory for nouns and adjectives, and *Wordsmyth*[8] for verbs. We only evaluated our WSD systems on nouns and adjectives.

[4]This software was kindly provided by David Yarowsky's group at Johns Hopkins University.

[5]Preliminary experiments using local features (bigrams and trigrams) showed low performance, which was expected because of noise in the automatically acquired data.

[6]This software was kindly provided by Gerard Escudero's group at Universitat Politecnica de Catalunya. The threshold was set in previous work.

[7]http://wordnet.princeton.edu

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

We also used the SemCor corpus (Miller et al., 1993) for tuning our relative-threshold heuristic. It contains a number of texts, mainly from the Brown Corpus, comprising about 200,000 words, where all content words have been manually tagged with senses from WordNet.

Throughout the paper we will use the concepts of precision and recall to measure the performance of WSD systems, where precision refers to the ratio of correct answers to the total number of answers given by the system, and recall indicates the ratio of correct answers to the total number of instances. Our ML systems attempt every instance and always give a unique answer, and hence precision equals to recall. When comparing with other systems that participated in Senseval-3 in Table 7, both recall and precision are shown. When POS and overall averages are given, they are calculated by micro-averaging the number of examples per word.

4 Experiments on TWA dataset

First we trained a VSM classifier on the sense examples translated with the *Systran* MT software (we use notion "MT-based approach" to refer to this process), and then tested it on the TWA test dataset. We tried two combinations of features: one only used topical features and the other used the whole feature set (i.e., topical and domain features). Table 1 summarises the sizes of the training/test data, the Most Frequent Sense (MFS) baseline and performances when applying the two different feature combinations. We can see that best results were obtained when using all the features. It also shows that both our systems achieved a significant improvement over the MFS baseline. Therefore, in the subsequent WSD experiments following the MT-based approach, we decided to use the entire feature set. To compare the machine-translated sense examples with the ones translated word-by-word, we then trained the same VSM classifier on the examples translated with a bilingual dictionary (we use notion "dictionary-based approach" to refer to this process) and evaluated it on the same test dataset. Table 2 shows results of the dictionary-based approach and the MT-based approach. For comparison, we include results from another system (Mihalcea, 2003), which uses monosemous relatives to automatically acquire sense examples. The right-most column shows results of a 10-fold

Word	Train ex.	Test ex.	MFS	Topical	All
bass	3,201	107	90.7	92.5	93.5
crane	3,656	95	74.7	84.2	83.2
motion	2,821	201	70.1	78.6	84.6
palm	1,220	201	71.1	82.6	85.1
plant	4,183	188	54.4	76.6	76.6
tank	3,898	201	62.7	79.1	77.1
Overall	18,979	993	70.6	81.1	82.5

Table 1:Recall(%) of the VSM classifier trained on the MT-translated sense examples, with different sets of features. The MFS baseline(%) and the number of training and test examples are also shown.

Word	(Mihalcea, 2003)	Dictionary-based	MT-based	Hand-tagged
bass	92.5	91.6	93.5	90.7
crane	71.6	74.5	83.2	81.1
motion	75.6	72.6	84.6	93.0
palm	80.6	81.1	85.1	87.6
plant	69.1	51.6	76.6	87.2
tank	63.7	66.7	77.1	84.1
Overall	76.6	71.3	82.5	87.6

Table 2:Recall(%) on TWA dataset for 3 unsupervised systems and a supervised cross-validation on test data.

cross-validation on the TWA data, which indicates the score that a supervised system would attain, taking additional advantage that the examples for training and test are drawn from the same corpus.

We can see that our MT-based approach has achieved significantly better recall than the other two automatic methods. Besides, the results of our unsupervised system are approaching the performance achieved with hand-tagged data. It is worth mentioning that Mihalcea (2003) applied a similar supervised cross-validation method on this dataset that scored 83.35%, very close to our unsupervised system[9]. Thus, we can conclude that the MT-based system is able to reach the best performance reported on this dataset for an unsupervised system.

5 Experiments on Senseval-3

In this section we describe the experiments carried out on the Senseval-3 lexical sample dataset. First, we introduce a heuristic method to deal with the problem of fine-grainedness of WordNet senses. The remaining two subsections will be devoted to the experiments of the baseline system and the contribution of the heuristic to the final system.

[9]The main difference to our hand-tagged evaluation, apart from the ML algorithm, is that we did not remove the bias from the "one sense per discourse" factor, as she did.

Threshold	Remove Senses	Remove Tokens	Sn.-Tk. ratio
4	7,669 (40.6)	11,154 (15.9)	2.55
5	9,759 (51.6)	15,516 (22.1)	2.34
6	11,341 (60.0)	18,827 (26.8)	2.24
7	12,569 (66.5)	21,775 (31.0)	2.14
8	13,553 (71.7)	24,224 (34.5)	2.08
9	14,376 (76.0)	27,332 (38.9)	1.95
10	14,914 (78.9)	29,418 (41.9)	1.88

Table 3:Sense filtering by relative-threshold on SemCor. For each threshold the number of removed senses/tokens and ambiguity are shown.

5.1 Unsupervised methods on fine-grained senses

When applying unsupervised WSD algorithms to fine-grained word senses, senses that rarely occur in texts often cause problems, as these cases are difficult to detect without relying on hand-tagged data. This is why many WSD systems use sense-tagged corpora such as SemCor to discard or penalise low-frequency senses.

For our work, we did not want to rely on hand-tagged corpora, and we devised a method to detect low-frequency senses and to remove them before using our translation-based approach. The method is based on the hypothesis that word senses that have few close relatives (synonyms, hypernyms, and hyponyms) tend to have low frequency in corpora. We collected all the close relatives to the target senses, according to WordNet, and then removed all the senses that did not have a number of relatives above a given threshold. We used this method on nouns, as the WordNet hierarchy is more developed for them.

First, we observed the effect of sense removal in the SemCor corpus. For all the polysemous nouns, we applied different thresholds (4-10 relatives) and measured the percentage of senses and SemCor tokens that were removed. Our goal was to remove as many senses as we could, while keeping as many tokens as possible. Table 3 shows the results of the process on all $5,438$ polysemous nouns in SemCor for a total of 18,912 senses and 70,238 tokens. The average number of senses per token initially is 3.47.

For the lowest threshold (4) we can see that we are able to remove a large number of senses from consideration (40%), keeping 85% of the tokens in SemCor. Higher thresholds can remove more senses, but it forces us to discard more valid tokens. In Table 3, the best ratios are given by lower thresholds, suggesting that conservative ap-

proaches would be better. However, we have to take into account that unsupervised state-of-the-art WSD methods on fine-grained senses perform below 50% recall on this dataset[10], and therefore an approach that is more aggressive may be worth trying.

We applied this heuristic method in our experiments and decided to measure the effect of the threshold parameter by relying on SemCor and the Senseval-3 training data. Thus, we tested the MT-based system for different threshold values, removing the senses for consideration when the relative number was below the threshold. The results of the experiments using this technique will be described in Section 5.3.

5.2 Baseline system

We performed experiments on Senseval-3 test data with both MT-based and dictionary-based approaches. We show the results for nouns and adjectives in Table 4, together with the MFS baseline (obtained from the Senseval-3 lexical sample training data). We can see that the results are similar for nouns, while for adjectives the MT-based system achieves significantly better recall. Overall, the performance was much lower than our previous 2-way disambiguation. The system also ranks below the MFS baseline.

One of the main reasons for the low performance was that senses with few examples in the test data are over-represented in training. This is because we trained the classifiers on equal number of maximumly 200 sense examples for every sense, no matter how rarely a sense actually occurs in real text. As we explained in the previous section, this problem could be alleviated for nouns by using the relative-based heuristics. We only implemented the MT-based approach for the rest of the experiments, as it performed better than the dictionary-based one.

5.3 Relative threshold

In this section we explored the contribution of the relative-based threshold to the system. We tested the system only on nouns. In order to tune the threshold parameter, we first applied the method on SemCor and the Senseval-3 training data. We used hand-tagged corpora from two different sources to see whether the method was generic enough to be applied on unseen test data. Note also that we used this experiment to define a general threshold for the heuristic, instead of optimising it for different words. Once the threshold is fixed, it will be used for all target words.

[10]Best score in Senseval-3 for nouns without SemCor or hand-tagged data: 47.5% recall (figure obtained from http://www.senseval.org).

Word	Test Ex.	MFS	Dictionary-based	MT-based
Nouns	1807	54.23	40.07	**40.73**
Adjs	159	49.69	15.74	**23.29**
Overall	1966	53.86	38.10	**39.32**

Table 4:Averaged recall(%) for the dictionary-based and MT-based methods in Senseval-3 lexical-sample data. The MFS baseline(%) and the number of testing examples are also shown.

Threshold	Avg. test ambiguity	Senseval-3	SemCor
0	5.80	40.68	30.11
4	3.60	40.15	32.99
5	3.32	39.43	32.82
6	2.76	40.53	34.18
7	2.52	43.89	35.94
8	2.36	46.90	39.15
9	2.08	45.37	38.98
10	1.88	**48.62**	46.16
11	1.80	48.59	**47.68**
12	1.68	48.34	43.63
13	1.40	47.23	45.31
14	1.28	44.32	42.05

Table 5:Average ambiguity and recall(%) for the relative-based threshold on Senseval-3 training data and SemCor (for nouns only). Best results shown in bold.

The results of the MT-based system applying threshold values from 4 to 14 are given in Table 5. We can see clearly that the algorithm benefits from the heuristic, specially when ambiguity is reduced to around 2 senses in average. Also observe that the contribution of the threshold is quite similar for SemCor and Senseval-3 training data. From this table, we chose 11 as threshold value for the test data, as it obtained the best performance on SemCor.

Thus, we performed a single run of the algorithm on the test data applying the chosen threshold. The performance for all nouns is given in Table 6. We can see that the recall has increased significantly, and is now closer to the MFS baseline, which is a very hard baseline for unsupervised systems (McCarthy et al., 2004). Still, the performance is significantly lower than the score achieved by supervised systems, which can reach above 72% recall (Mihalcea et al., 2004). Some of the reasons for the gap are the following:

- The acquisition process: problems can arise

Word	Test Ex.	MFS	Our System
argument	111	51.40	45.90
arm	133	82.00	85.70
atmosphere	81	66.70	35.80
audience	100	67.00	67.00
bank	132	67.40	67.40
degree	128	60.90	60.90
difference	114	40.40	40.40
difficulty	23	17.40	39.10
disc	100	38.00	27.00
image	74	36.50	17.60
interest	93	41.90	11.80
judgment	32	28.10	40.60
organization	56	73.20	19.60
paper	117	25.60	37.60
party	116	62.10	52.60
performance	87	26.40	26.40
plan	84	82.10	82.10
shelter	98	44.90	39.80
sort	96	65.60	65.60
source	32	65.60	65.60
Overall	1807	54.23	48.58

Table 6:Final results(%) for all nouns in Senseval-3 test data. Together with the number of test examples and MFS baseline(%).

from ambiguous Chinese words, and the acquired examples can contain noise generated by the MT software.

- Distribution of fine-grained senses: As we have seen, it is difficult to detect rare senses for unsupervised methods, while supervised systems can simply rely on frequency of senses.

- Lack of local context: Our system does not benefit from local bigrams and trigrams, which for supervised systems are one of the best sources of knowledge.

5.4 Comparison with Senseval-3 unsupervised systems

Finally, we compared the performance of our system with other unsupervised systems in the Senseval-3 lexical-sample competition. We evaluated these systems for nouns, using the outputs provided by the organisation[11], and focusing on the systems that are considered unsupervised. However, we noticed that most of these systems used the information of SemCor frequency, or even Senseval-3 examples in their models. Thus, we classified the systems depending on whether they used SemCor frequencies (Sc), Senseval-3 examples (S-3), or did not (Unsup.). This is an important distinction, as simply knowing the most frequent sense in hand-tagged data is a big advantage for unsupervised systems (applying the MFS heuristic for nouns in Senseval-3 would achieve 54.2% precision, and 53.0% recall when using SemCor). At this point, we would like to remark that, unlike other systems using Semcor, we have applied it to the minimum extent. Its only contribution has been to indirectly set the threshold for our general heuristic based on WordNet relatives. We are exploring better ways to integrate the relative information in the model.

[11]http://www.senseval.org

System	Type	Prec.	Recall
wsdiit	S-3	67.96	67.96
Cymfony	S-3	57.94	57.94
Prob0	S-3	55.01	54.13
clr04	Sc	48.86	48.75
upv-unige-CIAOSENSO	Sc	53.95	48.70
MT-based	**Unsup.**	**48.58**	**48.58**
duluth-senserelate	Unsup.	47.48	47.48
DFA-Unsup-LS	Sc	46.71	46.71
KUNLP.eng.ls	Sc	45.10	45.10
DLSI-UA-ls-eng-nosu.	Unsup.	20.01	16.05

Table 7:Comparison of unsupervised S3 systems for nouns (sorted by recall(%)). Our system given in bold.

The results of the Senseval-3 systems are given in Table 7. There are only 2 systems that do not require any hand-tagged data, and our method is able to improve both when using the relative-threshold. The best systems in Senseval-3 benefited from the training examples from the training data, particularly the top-scoring system, which is clearly supervised. The 2nd ranked system requires 10% of the training examples in Senseval-3 to map the clusters that it discovers automatically, and the 3rd simply applies the MFS heuristic.

The remaining systems introduce bias of the SemCor distribution in their models, which clearly helped their performance for each word. Our system is able to obtain a similar performance to the best of those systems without relying on hand-tagged data. We also evaluated the systems on the coarse-grained sense groups provided by the Senseval-3 organisers. The results in Table 8 show that our system is comparatively better on this coarse-grained disambiguation task.

6 Conclusions and Future Work

We automatically acquired English sense examples for WSD using large Chinese corpora and MT software. We compared our sense examples with those reported in previous work (Wang and Car-

System	Type	Prec.	Recall
wsdiit	S-3	75.3	75.3
Cymfony	S-3	66.6	66.6
Prob0	S-3	61.9	61.9
MT-based	**Unsup.**	**57.9**	**57.9**
clr04	Sc.	57.6	57.6
duluth-senserelate	Unsup.	56.1	56.1
KUNLP-eng-ls	Sc.	55.6	55.6
upv-unige-CIAOSENSO-	Sc.	61.3	55.3
DFA-Unsup-LS	Sc.	54.5	54.5
DLSI-UA-ls-eng-nosu.	Unsup.	27.6	27.6

Table 8:Coarse-grained evaluation of unsupervised S3 systems for nouns (sorted by recall(%)). Our system given in bold.

roll, 2005), by training a ML classifier on them and then testing the classifiers on both coarse-grained and fine-grained English gold standard datasets. On both datasets, our MT-based sense examples outperformed dictionary-based ones. In addition, evaluations show our unsupervised WSD system is competitive to the state-of-the-art supervised systems on binary disambiguation, and unsupervised systems on fine-grained disambiguation.

In the future, we would like to combine our approach with other systems based on automatic acquisition of sense examples that can provide local context (Agirre and Martinez, 2004b). The goal would be to construct a collection of examples automatically obtained from different sources and to apply ML algorithms on them. Each example would have a different weight depending on the acquisition method used.

Regarding the influence of sense distribution in the training data, we will explore the potential of using a weighting scheme on the "relative threshold" algorithm. Also, we would like to analyse if automatically obtained information on sense distribution (McCarthy et al., 2004) can improve WSD performance. We may also try other MT systems and possibly see if our WSD can in turn help MT, which can be viewed as a bootstrapping learning process. Another interesting direction is automatically selecting the most informative sense examples as training data for ML classifiers.

References

E. Agirre and D. Martinez. 2004a. The Basque Country University system: English and Basque tasks. In *Proceedings of the 3rd ACL workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL)*, Barcelona, Spain.

E. Agirre and D. Martinez. 2004b. Unsupervised wsd based on automatically retrieved examples: The importance of bias. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain.

Y. S. Chan and H. T. Ng. 2005. Scaling up word sense disambiguation via parallel texts. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*, Pittsburgh, Pennsylvania, USA.

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

M. Diab and P. 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.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

W. A. Gale, K. W. Church, and D. 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.

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

B. Magnini and G. Cavagliá. 2000. Integrating subject field codes into WordNet. In *Proceedings of the Second International LREC Conference*, Athens, Greece.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding Predominant Word Senses in Untagged Text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain.

R. Mihalcea, T. Chklovski, and Adam Killgariff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of the 3rd ACL workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL)*, Barcelona, Spain.

R. 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*.

G. A. Miller, C. Leacock, R. Tengi, and R. Bunker. 1993. A Semantic Concordance. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 303–308, Princeton, NJ, March. distributed as *Human Language Technology* by San Mateo, CA: Morgan Kaufmann Publishers.

H. T. Ng, B. Wang, and Y. S. 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*.

G. Ngai and R. Florian. 2001. Transformation-based learning in the fast lane. *Proceedings of the Second Conference of the North American Chapter of the Association for Computational Linguistics, pages 40-47, Pittsburgh, PA, USA.*

T. Pedersen. 2001. A decision tree of bigrams is an accurate predictor of word sense. *Proceedings of the Second Meeting of the NAACL, Pittsburgh, PA.*

X. Wang and J. Carroll. 2005. Word sense disambiguation using sense examples automatically acquired from a second language. In *Proceedings of HLT/EMNLP*, Vancouver, Canada.