TOR, TORMD: Distributional Profiles of Concepts for Unsupervised Word Sense Disambiguation

Saif Mohammad
Dept. of Computer Science
University of Toronto
Toronto, ON M5S 3G4
Canada
smm@cs.toronto.edu

Graeme Hirst
Dept. of Computer Science
University of Toronto
Toronto, ON M5S 3G4
Canada
gh@cs.toronto.edu

Philip Resnik
Dept. of Linguistics and UMIACS
University of Maryland
College Park, MD 20742
USA
resnik@umiacs.umd.edu

Abstract

Words in the context of a target word have long been used as features by supervised word-sense classifiers. Mohammad and Hirst (2006a) proposed a way to determine the strength of association between a sense or concept and co-occurring words—the distributional profile of a concept (DPC)—without the use of manually annotated data. We implemented an unsupervised naïve Bayes word sense classifier using these DPCs that was best or within one percentage point of the best unsupervised systems in the Multilingual Chinese–English Lexical Sample Task (task #5) and the English Lexical Sample Task (task #17). We also created a simple PMI-based classifier to attempt the English Lexical Substitution Task (task #10); however, its performance was poor.

1 Introduction

Determining the intended sense of a word is potentially useful in many natural language tasks including machine translation and information retrieval. The best approaches for word sense disambiguation are supervised and they use words that co-occur with the target as features. These systems rely on sense-annotated data to identify words that are indicative of the use of the target in each of its senses.

However, only limited amounts of sense-annotated data exist and it is expensive to create. In our previous work (Mohammad and Hirst, 2006a), we proposed an unsupervised approach to determine the strength of association between a sense or concept and its co-occurring words—**the distributional profile of a concept (DPC)**—relying simply on raw text and a published thesaurus. The categories in a published thesaurus were used as coarse senses or concepts (Yarowsky, 1992). We now show how distributional profiles of concepts can be used to create an *unsupervised* naïve Bayes word-sense classifier. We also implemented a simple classifier that relies on the pointwise mutual information (PMI) between the senses of the target and co-occurring words. These DPC-based classifiers participated in three SemEval 2007 tasks: the English Lexical Sample Task (task #17), the English Lexical Substitution Task (task #10), and the Multilingual Chinese–English Lexical Sample Task (task #5).

The English Lexical Sample Task (Pradhan et al., 2007) is a traditional word sense disambiguation task wherein the intended (WordNet) sense of a target word is to be determined from its context. We manually mapped the WordNet senses to the categories in a thesaurus and the DPC-based naïve Bayes classifier was used to identify the intended sense (category) of the target words.

The object of the Lexical Substitution Task (McCarthy and Navigli, 2007) is to replace a target word in a sentence with a suitable substitute that preserves the meaning of the utterance. The list of possible substitutes for a given target word is usually contingent on its intended sense. Therefore, word sense disambiguation is expected to be useful in lexical substitution. We used the PMI-based classier to determine the intended sense.

The objective of the Multilingual Chinese–English Lexical Sample Task (Jin et al., 2007) is to select from a given list a suitable English translation of a Chinese target word in context. Mohammad et al. (2007) proposed a way to create **cross-lingual distributional profiles of a concepts (CL-DPCs)**—the strengths of association between the concepts of one language and words of another. For this task, we mapped the list of English translations to appropriate thesaurus categories and used an implementation of a CL-DPC–based unsupervised naïve Bayes classifier to identify the intended senses (and thereby the English translations) of target Chinese words.

2 Distributional profiles of concepts

In order to determine the strength of association between a sense of the target word and its co-occurring words, we need to determine their individual and joint occurrence counts in a corpus. Mohammad and Hirst (2006a) and Mohammad et al. (2007) proposed ways to determine these counts in a monolingual and cross-lingual framework without the use of sense-annotated data. We summarize the ideas in this section; the original papers give more details.

2.1 Word–category co-occurrence matrix

We create a **word–category co-occurrence matrix (WCCM)** having English word types w^{en} as one dimension and English thesaurus categories c^{en} as another. We used the *Macquarie Thesaurus* (Bernard, 1986) both as a very coarse-grained sense inventory and a source of words that together represent each category (concept). The WCCM is populated with co-occurrence counts from a large English corpus (we used the *British National Corpus (BNC)*). A particular cell m_{ij}, corresponding to word w_i^{en} and concept c_j^{en}, is populated with the number of times w_i^{en} co-occurs with any word that has c_j^{en} as one of its senses (i.e., w_i^{en} co-occurs with any word listed under concept c_j^{en} in the thesaurus).

$$\begin{array}{c|cccccc}
 & c_1^{en} & c_2^{en} & \dots & c_j^{en} & \dots \\
\hline
w_1^{en} & m_{11} & m_{12} & \dots & m_{1j} & \dots \\
w_2^{en} & m_{21} & m_{22} & \dots & m_{2j} & \dots \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
w_i^{en} & m_{i1} & m_{i2} & \dots & m_{ij} & \dots \\
\vdots & \vdots & \vdots & \dots & \vdots & \ddots
\end{array}$$

A particular cell m_{ij}, corresponding to word w_i^{en} and concept c_j^{en}, is populated with the number of times w_i^{en} co-occurs with any word that has c_j^{en} as one of its senses (i.e., w_i^{en} co-occurs with any word listed under concept c_j^{en} in the thesaurus).

This matrix, created after a first pass of the corpus, is the **base word–category co-occurrence matrix (base WCCM)** and it captures strong associations between a sense and co-occurring words (see discussion of the general principle in Resnik (1998)). From the base WCCM we can determine the number of times a word w and concept c co-occur, the number of times w co-occurs with any concept, and the number of times c co-occurs with any word. A statistic such as PMI can then give the strength of association between w and c. This is similar to how Yarowsky (1992) identifies words that are indicative of a particular sense of the target word.

Words that occur close to a target word tend to be good indicators of its intended sense. Therefore, we make a second pass of the corpus, using the base WCCM to roughly disambiguate the words in it. For each word, the strength of association of each of the words in its context (±5 words) with each of its senses is summed. The sense that has the highest cumulative association is chosen as the intended sense. A new **bootstrapped WCCM** is created such that each cell m_{ij}, corresponding to word w_i^{en} and concept c_j^{en}, is populated with the number of times w_i^{en} co-occurs with any word *used in sense* c_j^{en}.

Mohammad and Hirst (2006a) used the DPCs created from the bootstrapped WCCM to attain near-upper-bound results in the task of determining word sense dominance. Unlike the McCarthy et al. (2004) dominance system, this approach can be applied to much smaller target texts (a few hundred sentences) without the need for a large similarly-sense-distributed text[1]. Mohammad and Hirst (2006b) used the DPC-based monolingual distributional measures of *concept-distance* to rank word pairs by their semantic similarity and to correct real-word spelling errors, attaining markedly better results than monolingual distributional measures of *word-distance*. In the spelling correction task, the

[1] The McCarthy et al. (2004) system needs to first generate a distributional thesaurus from the target text (if it is large enough—a few million words) or from another large text with a distribution of senses similar to the target text.

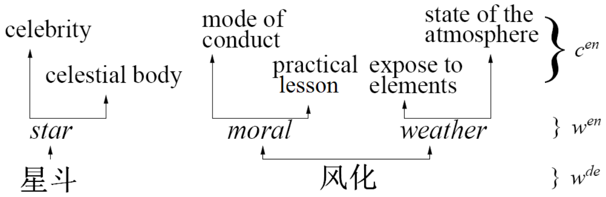

Figure 1: The cross-lingual candidate senses of Chinese words 星斗 and 风化.

distributional concept-distance measures performed better than all WordNet-based measures as well, except for the Jiang and Conrath (1997) measure.

2.2 Cross-lingual word–category co-occurrence matrix

Given a Chinese word w^{ch} in context, we use a Chinese–English bilingual lexicon to determine its different possible English translations. Each English translation w^{en} may have one or more possible coarse senses, as listed in an English thesaurus. These English thesaurus concepts (c^{en}) will be referred to as **cross-lingual candidate senses** of the Chinese word w^{ch}.[2] Figure 1 depicts examples.

We create a cross-lingual word–category co-occurrence matrix (CL-WCCM) with Chinese word types w^{ch} as one dimension and English thesaurus concepts c^{en} as another.

$$\begin{array}{c|cccccc} & c_1^{en} & c_2^{en} & \dots & c_j^{en} & \dots \\ \hline w_1^{ch} & m_{11} & m_{12} & \dots & m_{1j} & \dots \\ w_2^{ch} & m_{21} & m_{22} & \dots & m_{2j} & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ w_i^{ch} & m_{i1} & m_{i2} & \dots & m_{ij} & \dots \\ \vdots & \vdots & \vdots & \dots & \vdots & \ddots \end{array}$$

The matrix is populated with co-occurrence counts from a large Chinese corpus; we used a collection of LDC-distributed corpora[3]—Chinese Treebank English Parallel Corpus, FBIS data, Xinhua Chinese–English Parallel News Text Version 1.0 beta 2, Chinese English News Magazine Parallel Text, Chinese News Translation Text Part 1, and Hong Kong Parallel Text. A particular cell m_{ij}, corresponding to word w_i^{ch} and concept c_j^{en}, is populated with the number of times the Chinese word w_i^{ch} co-occurs with any Chinese word having c_j^{en} as one of its *cross-lingual candidate senses*. For example, the cell for 太空 ('space') and 'celestial body' will have the sum of the number of times 太空 co-occurs with 天体, 日, 太阳, 星, 星斗, and so on (see Figure 2). We used the *Macquarie Thesaurus* (Bernard, 1986) (about 98,000 words). The possible Chinese translations of an English word were taken from the Chinese–English Translation Lexicon version 3.0 (Huang and Graff, 2002) (about 54,000 entries).

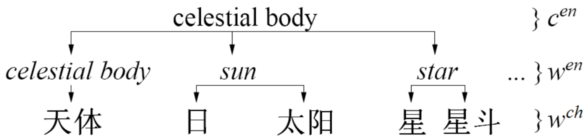

Figure 2: Chinese words having 'celestial body' as one of their cross-lingual candidate senses.

This base word–category co-occurrence matrix (base WCCM), created after a first pass of the corpus, captures strong associations between a category (concept) and co-occurring words. For example, even though we increment counts for both 太空–'celestial body' and 太空–'celebrity' for a particular instance where 太空 co-occurs with 星斗, 太空 will co-occur with a number of words such as 天体, 太阳, and 日 that each have the sense of *celestial body* in common (see Figure 2), whereas all their other senses are likely different and distributed across the set of concepts. Therefore, the co-occurrence count of 太空 and 'celestial body' will be relatively higher than that of 太空 and 'celebrity'.

As in the monolingual case, a second pass of the corpus is made to disambiguate the (Chinese) words in it. For each word, the strength of association of each of the words in its context (±5 words) with each of its cross-lingual candidate senses is summed. The sense that has the highest cumulative association with co-occurring words is chosen as the intended sense. A new bootstrapped WCCM is created by populating each cell m_{ij}, corresponding to word w_i^{ch} and concept c_j^{en}, with the number of times the Chinese word w_i^{ch} co-occurs with any Chi-

[2] Some of the cross-lingual candidate senses of w^{ch} might not really be senses of w^{ch} (e.g., 'celebrity', 'practical lesson', and 'state of the atmosphere' in Figure 1). However, as substantiated by experiments by Mohammad et al. (2007), our algorithm is able to handle the added ambiguity.

[3] http://www.ldc.upenn.edu

nese word *used in cross-lingual sense* c_j^{en}. A statistic such as PMI is then applied to these counts to determine the strengths of association between a target concept and co-occurring words, giving the distributional profile of the concept.

Mohammad et al. (2007) combined German text with an English thesaurus using a German–English bilingual lexicon to create German–English DPCs. These DPCs were used to determine semantic distance between German words, showing that state-of-the-art accuracies for one language can be achieved using a knowledge source (thesaurus) from another.

Given that a published thesaurus has about 1000 categories and the size of the vocabulary N is at least 100,000, the CL-WCCM and the WCCM are much smaller matrices (about $1000 \times N$) than the traditional word–word co-occurrence matrix ($N \times N$). Therefore the WCCMs are relatively inexpensive both in terms of memory and computation.

3 Classification

We implemented two unsupervised classifiers. The words in context were used as features.

3.1 Unsupervised Naïve Bayes Classifier

The naïve Bayes classifier has the following formula to determine the intended sense c_{nb}:

$$c_{nb} = \operatorname*{argmax}_{c_j \in C} P(c_j) \prod_{w_i \in W} P(w_i|c_j) \tag{1}$$

where C is the set of possible senses (as listed in the *Macquarie Thesaurus*) and W is the set of words that co-occur with the target (we used a window of ±5 words).

Traditionally, prior probabilities of the senses ($P(c_j)$) and the conditional probabilities in the likelihood ($\prod_{w_i \in W} P(w_i|c_j)$) are determined by simple counts in sense-annotated data. We approximate these probabilities using counts from the word–category co-occurrence matrix (monolingual or cross-lingual), thereby obviating the need for manually-annotated data.

$$P(c_j) = \frac{\sum_i m_{ij}}{\sum_{i,j} m_{ij}} \tag{2}$$

$$P(w_i|c_j) = \frac{m_{ij}}{\sum_i m_{ij}} \tag{3}$$

For the English Lexical Task, m_{ij} is the number of times the English word w_i co-occurs with the English category c_j—as listed in the word–category co-occurrence matrix (WCCM). For the Multilingual Chinese–English Lexical Task, m_{ij} is the number of times the Chinese word w_i co-occurs with the English category c_j—as listed in the cross-lingual word–category co-occurrence matrix (CL-WCCM).

3.2 PMI-based classifier

We calculate the pointwise mutual information between a sense of the target word and a co-occurring word using the following formula:

$$PMI(w_i, c_j) = \log \frac{P(w_i, c_j)}{P(w_i) \times P(c_j)} \tag{4}$$

$$\text{where} \quad P(w_i, c_j) = \frac{m_{ij}}{\sum_{i,j} m_{ij}} \tag{5}$$

$$\text{and} \quad P(w_i) = \frac{\sum_j m_{ij}}{\sum_{i,j} m_{ij}} \tag{6}$$

m_{ij} is the count in the WCCM or CL-WCCM (as described in the previous subsection). For each sense of the target, the sum of the strength of association (PMI) between it and each of the co-occurring words (in a window of ±5 words) is calculated. The sense with the highest sum is chosen as the intended sense.

$$c_{pmi} = \operatorname*{argmax}_{c_j \in C} \sum_{w_i \in W} PMI(w_i, c_j) \tag{7}$$

Note that this PMI-based classifier does not capitalize on prior probabilities of the different senses.

4 Data

4.1 English Lexical Sample Task

The English Lexical Sample Task training and test data (Pradhan et al., 2007) have 22281 and 4851 instances respectively for 100 target words (50 nouns and 50 verbs). WordNet 2.1 is used as the sense inventory for most of the target words, but certain words have one or more senses from OntoNotes (Hovy et al., 2006). Many of the fine-grained senses are grouped into coarser senses.

Our approach relies on representing a sense with a number of near-synonymous words, for which a thesaurus is a natural source. Even though the approach can be ported to WordNet[4], there was no easy

[4] The synonyms within a synset, along with its one-hop neighbors and all its hyponyms, can represent that sense.

		TRAINING DATA		TEST DATA		
WORDS	BASELINE	PMI-BASED	NAÏVE BAYES	PRIOR	LIKELIHOOD	NAÏVE BAYES
all	27.8	41.4	50.8	37.4	49.4	52.1
nouns only	25.6	43.4	53.6	18.1	49.6	49.7
verbs only	29.2	38.4	44.5	58.9	49.1	54.7

Table 1: English Lexical Sample Task: Results obtained using the PMI-based classifier on the training data and the naïve Bayes classifier on both training and test data

way of representing OntoNotes senses with near-synonymous words. Therefore, we asked four native speakers of English to map the WordNet and OntoNotes senses of the 100 target words to the *Macquarie Thesaurus* and use it as our sense inventory. We also wanted to examine the effect of using a very coarse sense inventory such as the categories in a published thesaurus (811 in all).

The annotators were presented with a target word, its WordNet/OntoNotes senses, and the Macquarie senses. WordNet senses were represented by synonyms, gloss, and example usages. The OntoNotes senses were described through syntactic patterns and example usages (provided by the task organizers). The Macquarie senses (categories) were described by the category head (a representative word for the category) and five other words in the category. Specifically, words in the same semicolon group[5] as the target were chosen. Annotators 1 and 2 labeled each WordNet/OntoNotes sense of the first 50 target words with one or more appropriate Macquarie categories. Annotators 3 and 4 labeled the senses of the other 50 words. We combined all four annotations into a WordNet–Macquarie mapping file by taking, for each target word, the union of categories chosen by the two annotators.

4.2 English Lexical Substitution Task

The English Lexical Substitution Task has 1710 test instances for 171 target words (nouns, verbs, adjectives, and adverbs) (McCarthy and Navigli, 2007). Some instances were randomly extracted from an Internet corpus, whereas others were selected manually from it. The target word might or might not be part of a multiword expression. The task is not tied to any particular sense inventory.

4.3 Multilingual Chinese–English Lexical Sample Task

The Multilingual Chinese–English Lexical Sample Task training and test data (Jin et al., 2007) have 2686 and 935 instances respectively for 40 target words (19 nouns and 21 verbs). The instances are taken from a corpus of *People's Daily News*. The organizers used the *Chinese Semantic Dictionary (CSD)*, developed by the Institute of Computational Linguistics, Peking University, both as a sense inventory and bilingual lexicon (to extract a suitable English translation of the target word once the intended Chinese sense is determined).

In order to determine the English translations of Chinese words in context, our system relies on Chinese text and an English thesaurus. As the thesaurus is used as our sense inventory, the first author and a native speaker of Chinese mapped the English translations of the target to appropriate Macquarie categories. We used three examples (from the training data) per English translation for this purpose.

5 Evaluation

5.1 English Lexical Sample Task

Both the naïve Bayes classifier and the PMI-based one were applied to the training data. For each instance, the Macquarie category c that best captures the intended sense of the target was determined. The instance was labeled with all the WordNet senses that are mapped to c in the WordNet–Macquarie mapping file (described earlier in Section 4.1).

5.1.1 Results

Table 1 shows the performances of the two classifiers. The system attempted to label all instances and so we report accuracy values instead of precision and recall. The naïve Bayes classifier performed markedly better in training than the PMI-

[5] Words within a semicolon group of a thesaurus tend to be more closely related than words across groups.

based one and so was applied to the test data. The table also lists baseline results obtained when a system randomly guesses one of the possible senses for each target word. Note that since this is a completely unsupervised system, it is not privy to the dominant sense of the target words. We do not rely on the ranking of senses in WordNet as that would be an implicit use of the sense-tagged SemCor corpus. Therefore, the most-frequent-sense baseline does not apply. Table 1 also shows results obtained using just the prior probability and likelihood components of the naïve Bayes formula. Note that the combined accuracy is higher than individual components for nouns but not for verbs.

5.1.2 Discussion

The naïve Bayes classifier's accuracy is only about one percentage point lower than that of the best unsupervised system taking part in the task (Pradhan et al., 2007). One reason that it does better than the PMI-based one is that it takes into account prior probabilities of the categories. However, using just the likelihood also outperforms the PMI classifier. This may be because of known problems of using PMI with low frequencies (Manning and Schütze, 1999). In case of verbs, lower combined accuracies compared to when using just prior probabilities suggests that the bag-of-words type features are not very useful. It is expected that more syntactically oriented features will give better results. Using window sizes ($\pm1, \pm2$, and ±10) on the training data resulted in lower accuracies than that obtained using a window of ±5 words. A smaller window size is probably missing useful co-occurring words, whereas a larger window size is adding words that are not indicative of the target's intended sense.

The use of a sense inventory (*Macquarie Thesaurus*) different from that used to label the data (WordNet) clearly will have a negative impact on the results. The mapping from WordNet/OntoNotes to Macquarie is likely to have some errors. Further, for 19 WordNet/OntoNotes senses, none of the annotators found a thesaurus category close enough in meaning. This meant that our system had no way of correctly disambiguating instances with these senses. Also impacting accuracy is the significantly fine-grained nature of WordNet compared to the thesaurus. For example, following are the three coarse senses for the noun *president* in WordNet: (1) executive officer of a firm or college, (2) the chief executive of a republic, and (3) President of the United States. The last two senses will fall into just one category for most, if not all, thesauri.

	BEST		OOT	
	Acc	Mode Acc	Acc	Mode Acc
all	2.98	4.72	11.19	14.63
Further Analysis				
NMWT	3.22	5.04	11.77	15.03
NMWS	3.32	4.90	12.22	15.26
RAND	3.10	5.20	9.98	13.00
MAN	2.84	4.17	12.61	16.49

Table 2: English Lexical Substitution Task: Results obtained using the PMI-based classifier

5.2 English Lexical Substitution Task

We used the PMI-based classifier[6] for the English Lexical Substitution Task. Once it identifies a suitable thesaurus category as the intended sense for a target, ten candidate substitutes are chosen from that category. Specifically, the category head word and up to nine words in the same semicolon group as the target are selected (words within a semicolon group are closer in meaning). Of the ten candidates, the single-word expression that is most frequent in the BNC is chosen as the best substitute; the motivation is that the annotators, who created the gold standard, were instructed to give preference to single words over multiword expressions as substitutes.

5.2.1 Results

The system was evaluated not only on the best substitute (BEST) but also on how good the top ten candidate substitutes are (OOT). Table 2 presents the results.[7] The system attempted all instances. The table also lists performances of the system on instances where the target is not part of a multiword expression (NMWT), on instances where the substitute is not a multiword expression (NMWS), on instances randomly extracted from the corpus (RAND), and on instances manually selected (MAN).

[6] Due to time constraints, we were able to upload results only with the PMI-based classifier by the task deadline.

[7] The formulae for accuracy and mode accuracy are as described by Pradhan et al. (2007).

	Training Data						Test Data					
	Baseline		PMI-based		Naïve Bayes		Prior		Likelihood		Naïve Bayes	
Words	micro	macro	micro	macro	micro	macro	micro	macro	micro	macro	micro	macro
all	33.1	38.3	33.9	40.0	38.5	44.7	35.4	41.7	38.8	44.6	37.5	43.1
nouns only	41.9	43.5	43.6	45.0	49.4	50.5	45.3	47.1	48.1	50.8	50.0	51.6
verbs only	28.0	34.1	28.0	35.6	31.9	39.6	29.1	36.8	32.9	39.0	29.6	35.5

Table 3: Multilingual Chinese–English Lexical Sample Task: Results obtained using the PMI-based classifier on the training data and the naïve Bayes classifier on both training and test data

5.2.2 Discussion

Competitive performance of our DPC-based system on the English Lexical Sample Task and the Chinese–English Lexical Sample Task (see next subsection) suggests that DPCs are useful for sense disambiguation. Poor results on the substitution task can be ascribed to several factors. First, we used the PMI-based classifier that we found later to be markedly less accurate than the naïve Bayes classifier in the other two tasks. Second, the words in the thesaurus categories may not always be near-synonyms; they might just be strongly related. Such words will be poor substitutes for the target. Also, we chose as the best substitute simply the most frequent of the ten candidates. This simple technique is probably not accurate enough. On the other hand, because we chose the candidates without any regard to frequency in a corpus, the system chose certain infrequent words such as *wellnigh* and *ecchymosed*, which were not good candidate substitutes.

5.3 Multilingual Chinese–English Lexical Sample Task

In the Multilingual Chinese–English Lexical Sample Task, both the naïve Bayes classifier and the PMI-based classifier were applied to the training data. For each instance, the Macquarie category, say c, that best captures the intended sense of the target word is determined. Then the instance is labeled with all the English translations that are mapped to c in the English translations–Macquarie mapping file (described earlier in Section 4.3).

5.3.1 Results

Table 3 shows accuracies of the two classifiers. Macro average is the ratio of number of instances correctly disambiguated to the total, whereas micro average is the average of the accuracies achieved on each target word. As in the English Lexical Sample Task, both classifiers, especially the naïve Bayes classifier, perform well above the random baseline. Since the naïve Bayes classifier also performed markedly better than the PMI-based one in training, it was applied to the test data. Table 3 also shows results obtained using just the likelihood and prior probability components of the naïve Bayes classifier on the test data.

5.3.2 Discussion

Our naïve Bayes classifier scored highest of all unsupervised systems taking part in the task (Jin et al., 2007). As in the English Lexical Sample Task, using just the likelihood again outperforms the PMI classifier on the training data. The use of a sense inventory different from that used to label the data again will have a negative impact on the results as the mapping may have a few errors. The annotator believed none of the given Macquarie categories could be mapped to two Chinese Semantic Dictionary senses. This meant that our system had no way of correctly disambiguating instances with these senses.

There were also a number of cases where more than one CSD sense of a word was mapped to the same Macquarie category. This occurred for two reasons: First, the categories of the *Macquarie Thesaurus* act as very coarse senses. Second, for certain target words, the two CSD senses may be different in terms of their syntactic behavior, yet semantically very close (for example, the ‘be shocked’ and ‘shocked’ senses of 震惊). This many-to-one mapping meant that for a number of instances more than one English translation was chosen. Since the task required us to provide exactly one answer (and there was no partial credit in case of multiple answers), a category was chosen at random.

6 Conclusion

We implemented a system that uses distributional profiles of concepts (DPCs) for unsupervised word sense disambiguation. We used words in the context as features. Specifically, we used the DPCs to create a naïve Bayes word-sense classifier and a simple PMI-based classifier. Our system attempted three SemEval-2007 tasks. On the training data of the English Lexical Sample Task (task #17) and the Multilingual Chinese–English Lexical Sample Task (task #5), the naïve Bayes classifier achieved markedly better results than the PMI-based classifier and so was applied to the respective test data. On both test and training data of both tasks, the system achieved accuracies well above the random baseline. Further, our system placed best or close to one percentage point from the best among the unsupervised systems. In the English Lexical Substitution Task (task #10), for which there was no training data, we used the PMI-based classifier. The system performed poorly, which is probably a result of using the weaker classifier and a simple brute force method for identifying the substitute among the words in a thesaurus category. Markedly higher-than-baseline performance of the naïve Bayes classifier on task #17 and task #5 suggests that the DPCs are useful for word sense disambiguation.

Acknowledgments

We gratefully acknowledge Xiaodan Zhu, Michael Demko, Christopher Parisien, Frank Rudicz, and Timothy Fowler for mapping training-data labels to categories in the *Macquarie Thesaurus*. We thank Michael Demko, Siddharth Patwardhan, Xinglong Wang, Vivian Tsang, and Afra Alishahi for helpful discussions. This research is financially supported by the Natural Sciences and Engineering Research Council of Canada, the University of Toronto, ONR MURI Contract FCPO.810548265 and Department of Defense contract RD-02-5700.

References

J.R.L. Bernard, editor. 1986. *The Macquarie Thesaurus*. Macquarie Library, Sydney, Australia.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 57–60, New York, NY.

Shudong Huang and David Graff. 2002. Chinese–english translation lexicon version 3.0. *Linguistic Data Consortium*.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research on Computational Linguistics*, Taiwan.

Peng Jin, Yunfang Wu, and Shiwen Yu. 2007. SemEval-2007 task 05: Multilingual Chinese-English lexical sample task. In *Proceedings of the Fourth International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Prague, Czech Republic.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SemEval-2007)*, Prague, Czech Republic.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 280–267, Barcelona, Spain.

Saif Mohammad and Graeme Hirst. 2006a. Determining word sense dominance using a thesaurus. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy.

Saif Mohammad and Graeme Hirst. 2006b. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, Sydney, Australia.

Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL-2007)*, Prague, Czech Republic.

Sameer Pradhan, Martha Palmer, and Edward Loper. 2007. SemEval-2007 task 17: English lexical sample, English SRL and English all-words tasks. In *Proceedings of the Fourth International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SemEval-2007)*, Prague, Czech Republic.

Philip Resnik. 1998. Wordnet and class-based probabilities. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 239–263. The MIT Press, Cambridge, Massachusetts.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 454–460, Nantes, France.