Computational Linguistics and Chinese Language Processing
Vol. 11, No. 3, September 2006, pp. 223-244

Sense Extraction and Disambiguation for Chinese Words from Bilingual Terminology Bank

Ming-Hong Bai[*,+], Keh-Jiann Chen[*] and Jason S. Chang[+]

Abstract

Using lexical semantic knowledge to solve natural language processing problems has been getting popular in recent years. Because semantic processing relies heavily on lexical semantic knowledge, the construction of lexical semantic databases has become urgent. WordNet is the most famous English semantic knowledge database at present; many researches of word sense disambiguation adopt it as a standard. Because of the success of WordNet, there is a trend to construct WordNet in different languages. In this paper, we propose a methodology for constructing Chinese WordNet by extracting information from a bilingual terminology bank. We developed an algorithm of word-to-word alignment to extract the English-Chinese translation-equivalent word pairs first. Then, the algorithm disambiguates word senses and maps Chinese word senses to WordNet synsets to achieve the goal. In the word-to-word alignment experiment, this alignment algorithm achieves the f-score of 98.4%. In the word sense disambiguation experiment, the extracted senses cover 36.89% of WordNet synsets and the accuracy of the three proposed disambiguation rules achieve the accuracies of 80%, 83% and 87%, respectively.

Keywords: Word Alignment, Word Sense Disambiguation, WordNet, EM Algorithm, Sense Tagging.

1. Introduction

Using lexical semantic knowledge to solve natural language processing problems has been getting popular in recent years. Especially for word sense disambiguation, the semantic lexicon plays a very important role. However, all semantic approaches depend on knowledge of some well established semantic lexical databases which provide semantic information of

[*] Institute of Information Science, Academia Sinica
E-mail: mhbai@sinica.edu.tw; kchen@iis.sinica.edu.tw
[+] Department of Computer Science, National Tsing Hua University
E-mail: jschang@cs.nthu.edu.tw

[Received February 16, 2006; Revised November 28, 2006; Accepted November 30, 2006]

words, such as the different senses of a word, the synonymous or hyperonymy relation between words, etc.

WordNet is a famous semantic lexical database which owns rich lexical information. [Miller 1990]. It not only covers a large set of vocabularies but also establishes a complete taxonomic structure for word senses. Synonymous word senses are grouped into synsets. These synsets are further associated by semantic relations, including hypernyms, hyponyms, holonyms, meronyms, etc. The WordNet has been applied to a wide range of applications, such as word sense disambiguation, information retrieval, computer-assisted language learning, etc. It has apparently become the de facto standard for English word senses now.

Because of the success of WordNet, there is a universally shared interest in construction of WordNet-like and WordNet-embedded lexical databases in different languages. One of the most famous projects is EuroWordNet (EWN). Its goal is to construct a WordNet-like system containing several European languages. Since constructing a WordNet for a new language is a difficult and labor intensive task, using the resources of WordNet to speed up the construction has begun a new trend. Many researchers, such as [Atserias *et al.* 1997], [Daude *et al.* 1999] and [Chang *et al.* 2003], have tried to associate WordNet synsets to other languages automatically with appropriate translations from bilingual dictionaries. The limitation of using bilingual dictionaries as mapping tables for translation equivalences between two languages is the narrow scopes of the dictionaries, since dictionaries usually contain prototypical translations only. For example, the first sense of word "plant" in WordNet is "plant, works, industrial plant"; it was translated as "GongChang"(工廠) in a Chinese-English bilingual dictionary. However, in actual text, it may be also translated as "Chang"(廠), "GongChang"(工場), "ChangFang"(廠房), "suo"(所, such as ‘power plant’/發電所), etc. Various translations, obviously, add complexity and difficulty to map word senses into WordNet synsets.

Instead of using bilingual dictionaries, we adopt a bilingual terminology bank as the semantic lexical database. The latter includes various compound words, in which a word in a different compounding structure may have different translations, thus there are more translation candidates which can be chosen. A bilingual terminology bank has not only helped to avoid the problem of the limited scope of prototypical translations made by common bilingual dictionaries, but has also helped to disambiguate word senses by various translations and collocations [Diab *et al.* 2002], [Bhattacharya 2004]. Nevertheless, using bilingual terminology banks has to face two main challenges: Firstly, we have to deal with the problem of word-to-word alignment for multi-words terms. Secondly, we have to solve the problem of sense ambiguity of the English translation. The approaches for solving these two problems are the major focuses of the paper.

The rest of paper is divided into four sections. Section 2 introduces the resources of this paper. Section 3 describes the methodology. Experimental setup and results will be addressed in Section 4. A conclusion is provided in Section 5 along with directions for future research.

2. Resources

In this study, we use two dictionaries as the resources to extract semantic information:

a) The Bilingual Terminology Bank from NICT [NICT 2004]

b) A English-Chinese dictionary [Proctor 1988]

The Bilingual Terminology Bank from NICT contains 63 classes of terminologies, with a total of 1,046,058 Chinese terms with their English translations. Among them, 629,352 terms are compounds, which is about 60 percent of the total. The English-Chinese dictionary contains 208,163 words which are used as a supplement. We also adopt WordNet 2.0 as the medium for sense linking. Figure 1 shows some sample entries of the Bilingual Terminology Bank from NICT.

English	**Chinese**	**Class**
succulent stem	肉質莖	Botany
common base current gain	共基電流增益	Electrical Engineering
sliding brush	滑動電刷	Naval Architecture
point of increase	增値點	Mathematics
group carry	成組進位	Computer Science
swine fever	豬瘟	Animal Science
light measurements	光量測	Metrology
reductional grouping	染色體減數分群	Botany
oil film strength	油膜強度	Metrology
normalized quadrature spectrum	標準化四分譜	Meteorology

Figure 1. sample entries of the Bilingual Terminology Bank from NICT.

In English, a compound is usually composed of words and blanks; the latter being a natural boundary to separate words. On the contrary, in Chinese there are no blanks in compound words, so we need to segment words before applying word alignment algorithms. In this paper, we adopt the CKIP Chinese Word Segmentation System, which was developed by the CKIP group of Academia Sinica [CKIP 2006].

3. Methodology

The algorithm can be divided into the following two steps:

1. Find the word to word alignment for each entry in the terminology bank,

2. Assign a synset to the Chinese word sense by resolving the sense ambiguities of its aligned English word.

The first step is to find all possible English translations for each Chinese word, which make it possible to link Chinese words to WordNet synsets. Since the English translation may be ambiguous, the purpose of second step is to employ a word sense disambiguation algorithm to select the appropriate synset for the Chinese word. For example, the term pair (*water tank*, *水 槽*) will be aligned as (*water*/*水* *tank*/*槽*) in the first step, so the Chinese word *槽* can be linked to WordNet synsets by its translation *tank*. But *tank* has five senses in WordNet as follows:

*tank*_n_1: an enclosed armored military vehicle,

*tank*_n_2: a large vessel for holding gases or liquids,

*tank*_n_3: as much as a tank will hold,

*tank*_n_4: a freight car that transports liquids or gases in bulk,

*tank*_n_5: a cell for violent prisoners.

The second step is applied to select the best sense translation. In the following subsections, we will describe the detail algorithm of word alignment in section 3.1 and word sense disambiguation in section 3.2.

3.1 Word Alignment

For a Chinese term and its English translation, it is natural to think that the Chinese term is translated from the English term word for word. So, the purpose of word alignment is to connect the words which have a translation relationship between the Chinese term and its English portion. In past years, several statistical-based word alignment methods have been proposed. [Brown *et al*. 1993] proposed a method of word alignment which consists of five translation models, also known as the IBM translation models. Each model focuses on some features of a sentence pair to estimate the translation probability. [Vogel *et al*. 1996] proposed the Hidden-Markov alignment model which makes the alignment probabilities dependent on the alignment position of the previous word rather than on the absolute positions. [Och and Ney 2000] proposed some methods to adjust the IBM models to improve alignment performance.

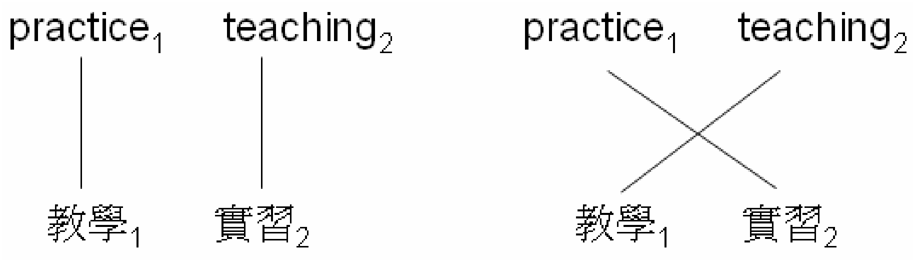
The word alignment task in this paper only focuses on the term pairs of a bilingual terminology bank. Since the length of a term is usually far less than a sentence, some features, such as word position, are no longer important in the task. In this paper, we employ the IBM-1 model, which only focuses on lexical generating probability, to align the words of a bilingual terminology bank.

3.1.1 Modeling Word Alignment

For convenience, we follow the notion of [Brown *et al.* 1993], which defines word alignment as follows:

Suppose we have a English term $\mathbf{e} = e_1,e_2,\ldots,e_n$ where e_i is an English word, and its corresponding Chinese term $\mathbf{c} = c_1,c_2,\ldots,c_m$ where c_j is a Chinese word. An alignment from **e** to **c** can be represented by a series $\mathbf{a}=a_1,a_2,\ldots,a_m$ where each a_j is an integer between 0 and n, such that if c_j is partial (or total) translation of e_i , then $a_j = i$ and if it is not translation of any English word, then a_j=0.

For example, the alignments shown in Figure 2 are two possible alignments from English to Chinese for the term pair (*practice teaching, 教學 實習*), (a) can be represented by **a**=1,2 while (b) can be represented by **a**=2,1.



(a) (b)

Figure 2. two possible alignments from English to Chinese for the term pair (practice teaching, 教學 實習).

In the word alignment stage, given a pair of terms **c** and **e**, we want to find the most likely alignment $\mathbf{a}=a_1,a_2,\ldots,a_\mathrm{m}$, to maximize the alignment probability P(**a**|**c**,**e**) for the pair. The formula can be represented as follows:

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} P(\mathbf{a} \mid \mathbf{c}, \mathbf{e})\,, \qquad (1)$$

where $\hat{\mathbf{a}}$ is the best alignment of the possible alignments. Suppose we already have lexical translation probabilities for each of the lexical pairs, then, the alignment probability $P(\mathbf{a}|\mathbf{c},\mathbf{e})$ can be estimated by means of the lexical translation probabilities as follows:

$$P(\mathbf{a} \mid \mathbf{c}, \mathbf{e}) = \frac{P(\mathbf{a}, \mathbf{c} \mid \mathbf{e})}{P(\mathbf{c} \mid \mathbf{e})} = \prod_{j=1}^{m} P(c_j \mid e_{a_j}) / P(\mathbf{c} \mid \mathbf{e})\,.$$

The probability of c given e, P(c|e), is a constant for a given term pair (c,e), so formula 1 can be estimated as follows:

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \prod_{j=1}^{m} P(c_j \mid e_{a_j}) \; . \tag{2}$$

For example, the probability of the alignment shown in Figure 2 (a) can be estimated by:

$P(c_1|e_1)P(c_2|e_2)$

= *P*(*教學* | *practice*) *P*(*實習* | *teaching*)

= 0.000480 x 1.14x10^{-13} =5.48x10^{-17}.

While (b) can be estimated by:

$P(c_1|e_2)\mathrm{p}(c_2|e_1)$

= *P*(*教學* | *teaching*)*P*(*實習* | *practice*)

= 0.6953 x 0.0940 = 0.0654.

In this example, the probability of alignment (b) is larger than (a) in Figure 2. So the alignment (b), (*教學/teaching 實習/practice*), is a better choice than (a), (*教學/practice 實習 /teaching*), for the term pair (*practice teaching*, *教學 實習*). The remaining problem of this stage is how to estimate the translation probability *p*(*c*|*e*) for all possible English-Chinese lexical pairs.

3.1.2 Translation Probability Estimation

The method of our translation probability estimation uses the IBM model 1 [Brown *et al.* 1993], which is based on the EM algorithm [Dempster *et al.* 1977], for maximizing the likelihood of generating the Chinese terms, which is the target language, given the English portion, which is the source language. Suppose we have an English term **e** and its Chinese translation **c** in the terminology bank **T**; *e* is a word in **e,** and *c* is a word in **c**. The probability of word *c* given word *e*, *P*(*c*|*e*), can be estimated by iteratively re-estimating the following EM formulae:

Initialization:

$$P(c \mid e) = \frac{1}{|C|} \; ; \tag{3}$$

E-step:

$$Z(c,e;\mathbf{c},\mathbf{e}) = \sum_{\forall \mathbf{a}} P(\mathbf{a} \mid \mathbf{c},\mathbf{e}) \sum_{j=1}^{m} \delta(c,c_j)\delta(e,e_{a_j}) \; , \tag{4}$$

$$P(\mathbf{a} \mid \mathbf{c},\mathbf{e}) = \frac{P(\mathbf{a},\mathbf{c} \mid \mathbf{e})}{\sum_{\forall \mathbf{a}'} P(\mathbf{a}',\mathbf{c} \mid \mathbf{e})} = \frac{\prod_{j=1}^{m} P(c_j \mid e_{a_j})}{\sum_{\forall \mathbf{a}'} \prod_{j=1}^{m} P(c_j \mid e_{a'_j})} \; ; \tag{5}$$

M-step:

$$P(c\,|\,e)=\frac{\sum_{t=1}^{|T|}Z(c,e;\mathbf{c}^{(t)},\mathbf{e}^{(t)})}{\sum\limits_{\forall v\in C}\sum_{t=1}^{|T|}Z(v,e;\mathbf{c}^{(t)},\mathbf{e}^{(t)})}\,. \tag{6}$$

In the EM training process, we initially assume that the translation probability for any Chinese word *c* given English word *e*, *P*(*c*|*e*), is uniformly distributed as in formula 3, where ***C*** denotes the set of all Chinese words in the terminology bank. In the E-step, we estimate the expected number of times that *e* connects to *c* in the term pair (**c**,**e**). As in formula 4, we sum up the expected counts of the connection from *e* to *c* over all possible alignments which contain the connection. Formula 5 is the detailed definition of the probability of an alignment **a** given (**c**,**e**). Usually, it is hard to evaluate the formulae in E-step. Fortunately, it has been proven [Brown *et al*. 1993] that the expectation formulae, 4 and 5, can be merged and simplified as follows:

$$\begin{aligned}
Z(c,e;\mathbf{c},\mathbf{e})&=\sum_{\mathbf{a}}P(\mathbf{a}\,|\,\mathbf{c},\mathbf{e})\sum_{j=1}^{m}\delta(c,c_j)\delta(e,e_{a_j})\\
&=\frac{\sum\limits_{\mathbf{a}}\prod_{j=1}^{m}P(c_j\,|\,e_{a_j})\sum_{j=1}^{m}\delta(c,c_j)\delta(e,e_{a_j})}{\sum\limits_{\forall\mathbf{a}'}\prod_{j=1}^{m}P(c_j\,|\,e_{a'_j})}\\
&=\frac{P(c\,|\,e)\prod_{j=1,c_j\neq c}^{m}\sum_{i=0,e_i\neq e}^{n}P(c_j\,|\,e_i)}{\prod_{j=1}^{m}\sum_{i=0}^{n}P(c_j\,|\,e_i)}\sum_{j=1}^{m}\delta(c,c_j)\sum_{i=0}^{n}\delta(e,e_i)\\
&=\frac{P(c\,|\,e)}{\sum_{i=0}^{n}P(c\,|\,e_i)}\sum_{j=1}^{m}\delta(c,c_j)\sum_{i=0}^{n}\delta(e,e_i)\,.
\end{aligned} \tag{7}$$

After merging and simplifying, as formula 7, the E-step becomes very simple and effective for computing.

In the M-step, we re-estimate the translation probability, *P*(*c*|*e*). As shown in formula 6, we sum up the expected number of connections from *e* to *c* over the whole bank divide by the expected number of *c*.

The training process will count the expected number, E-step, and re-estimate the translation probability, M-step, iteratively until it has converged.

For instance, as the example shown in Figure 2, the English term **e=** *practice teaching* and Chinese term **c**=*教學 實習* are given. Assume the total number of Chinese words in the terminology bank is 100,000. Initially, the probabilities of each translation are as follows:

$$P(\text{教學} \mid practice) = \frac{1}{|C|} = 0.00001, \qquad P(\text{教學} \mid teaching) = \frac{1}{|C|} = 0.00001,$$

$$P(\text{實習} \mid practice) = \frac{1}{|C|} = 0.00001, \qquad P(\text{實習} \mid teaching) = \frac{1}{|C|} = 0.00001.$$

In E-step, we count the expected number for all possible connections in the term pair:

$$\text{Z}(\text{教學}, practice; \mathbf{e}, \mathbf{c}) = \frac{P(\text{教學} \mid practice)}{P(\text{教學} \mid practice) + P(\text{教學} \mid teaching)} = 0.5,$$

$$\text{Z}(\text{教學}, teaching; \mathbf{e}, \mathbf{c}) = \frac{P(\text{教學} \mid teaching)}{P(\text{教學} \mid practice) + P(\text{教學} \mid teaching)} = 0.5,$$

$$\text{Z}(\text{實習}, practice; \mathbf{e}, \mathbf{c}) = \frac{P(\text{實習} \mid practice)}{P(\text{實習} \mid practice) + P(\text{實習} \mid teaching)} = 0.5,$$

$$\text{Z}(\text{實習}, teaching; \mathbf{e}, \mathbf{c}) = \frac{P(\text{實習} \mid practice)}{P(\text{實習} \mid practice) + P(\text{實習} \mid teaching)} = 0.5.$$

In M-step, we first count the global expected number of each translation by summing up the expected number of each data entry over the whole term bank:

$$\sum_{t=1}^{|T|} Z(\text{教學}, practice; \mathbf{e}^{(t)}, \mathbf{c}^{(t)}) = 0.7,$$

$$\sum_{t=1}^{|T|} Z(\text{教學}, teaching; \mathbf{e}^{(t)}, \mathbf{c}^{(t)}) = 43.72,$$

$$\sum_{t=1}^{|T|} Z(\text{實習}, practice; \mathbf{e}^{(t)}, \mathbf{c}^{(t)}) = 5.37,$$

$$\sum_{t=1}^{|T|} Z(\text{實習}, teaching; \mathbf{e}^{(t)}, \mathbf{c}^{(t)}) = 0.95.$$

After the global expected number of each translation has been counted, we can re-estimate the translation probabilities by means of the expected numbers:

$$\text{P}(\text{教學} \mid practice) = \frac{\sum_{t=1}^{|T|} Z(\text{教學}, practice; \mathbf{e}^{(t)}, \mathbf{c}^{(t)})}{\sum_{v \in C} \sum_{t=1}^{|T|} Z(v, practice; \mathbf{e}^{(t)}, \mathbf{c}^{(t)})} = \frac{0.7}{110.67} = 0.00632,$$

$$\text{P}(\text{教學} \mid teaching) = \frac{\sum_{t=1}^{|T|} Z(\text{教學}, teaching; \mathbf{e}^{(t)}, \mathbf{c}^{(t)})}{\sum_{v \in C} \sum_{t=1}^{|T|} Z(v, teaching; \mathbf{e}^{(t)}, \mathbf{c}^{(t)})} = \frac{43.72}{121.88} = 0.35871,$$

$$\mathrm{P}(\text{實習} \mid practice) = \frac{\sum_{t=1}^{|T|} Z(\text{實習}, practice; \mathbf{e}^{(t)}, \mathbf{c}^{(t)})}{\sum_{v \in C} \sum_{t=1}^{|T|} Z(v, practice; \mathbf{e}^{(t)}, \mathbf{c}^{(t)})} = \frac{5.37}{110.67} = 0.04852,$$

$$\mathrm{P}(\text{實習} \mid teaching) = \frac{\sum_{t=1}^{|T|} Z(\text{實習}, teaching; \mathbf{e}^{(t)}, \mathbf{c}^{(t)})}{\sum_{v \in C} \sum_{t=1}^{|T|} Z(v, teaching; \mathbf{e}^{(t)}, \mathbf{c}^{(t)})} = \frac{0.95}{121.88} = 0.00779.$$

The training process will count the expected number and re-estimate the translation iteratively until it has converged. There are some translation probabilities estimated in this experiment shown in Figures 3-6.

English	Chinese	P(*c* | *e*)
water	水	0.599932
water	水位	0.048781
water	水分	0.011677
water	用水	0.011427
water	地下水	0.010800
water	水壓	0.009310
water	水量	0.007905
water	水管	0.007640
water	位	0.007471
water	水面	0.006704

Figure 3. translation probabilities for water.

English	Chinese	P(*c* | *e*)
tank	槽	0.292606
tank	櫃	0.176049
tank	艙	0.077515
tank	箱	0.034325
tank	水	0.025067
tank	液	0.018411
tank	水槽	0.016570
tank	池	0.016157
tank	罐	0.015687
tank	水箱	0.012206

Figure 4. translation probabilities for tank.

English	Chinese	P(*c* | *e*)
practice	練習	0.163636
practice	實習	0.093320
practice	演習	0.058102
practice	實務	0.056980
practice	操作	0.051331
practice	優良	0.042036
practice	作業	0.038144
practice	方法	0.036161
practice	實作	0.034805
practice	實際	0.025800

Figure 5. translation probabilities for practice.

English	Chinese	P(*c* | *e*)
teaching	教學	0.698757
teaching	教學法	0.137614
teaching	教材	0.045780
teaching	單元	0.015502
teaching	教具	0.010315
teaching	教導	0.007246
teaching	教會	0.007246
teaching	教授	0.007246
teaching	教訓	0.007246
teaching	教	0.007246

Figure 6. translation probabilities for teaching.

3.1.3 Imposing Alignment Constraints

As was mentioned in Section 3.1.1, the goal of word alignment is to find the best alignment candidate to maximize the translation probability of a term pair. However, in real situations there are some problems that have to be solved:

1. Cross connections: assume there is a series of words, c_j,c_{j+1},c_{j+2} in a Chinese term, if c_j and c_{j+2} connect to the same English word while $\mathrm{c}_{\mathrm{j+1}}$ connects to any other word, we call this

alignment contains a cross connection. There is an example of cross connection shown in Figure 7. The Chinese word 校 is more likely to connect to *examination* shown in Figure 8.

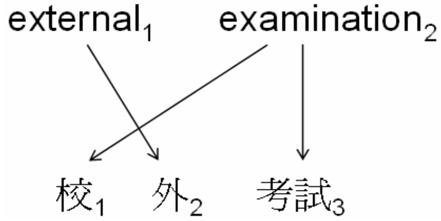


Figure 7. example of cross connection, 校 and 考試 connected to examination while 外 connected to external.

	校	外	考試
external	1.4x10^{-7}	**0.575537**	5.3x10^{-9}
examination	**5.2x10^{-6}**	5.2x10-6	**0.172751**

Figure 8. example of cross connection: the translation probabilities of the example, it shows that 校 is more likely to connect to examination.

2. Function words: in word alignment stage, function words are usually ignored except when they are part of compound words. For example, Figure 9, *of* is a part of a compound which can not be skipped, while in Figure 10, *of* can be skipped.

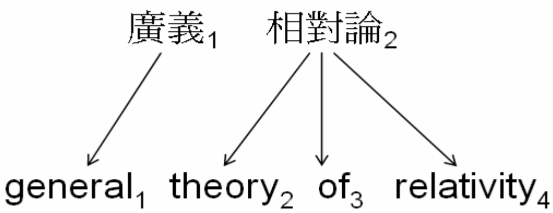


Figure 9. of is part of compound.

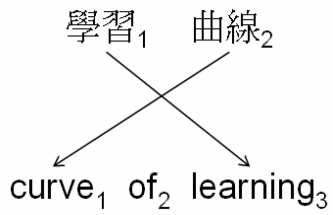


Figure 10. of is not part of compound.

In order to solve this problem, two constraints are imposed on the alignment algorithm. Formula 1 is altered by using a cost function instead of probability, defined as follows:

$$\mathbf{a} = \arg\min_{\mathbf{a}} cost(\mathbf{a}), \tag{8}$$

where cost function is given by:

$$\cos\mathrm{t}(\mathbf{a}) = \begin{cases} \infty, & \textit{if } \textit{cross\_connection}(\mathbf{a}) = \textit{true} \\ \infty, & \begin{array}{l} \textit{if } \ a_i \text{ connects } c_i \text{ to any word} \\ \quad \text{and } c_i \text{ is a function word} \\ \quad \text{and } c_i \text{ is not part of compound} \end{array} \\ \sum_{j=1}^{k} -\log(p(c_j \mid e_{a_j})) & \textit{else} \end{cases} . \tag{9}$$

The *cross connection* function is used to detect the cross connection in an alignment candidate. If a cross connection is found, the alignment candidate will be assigned a large cost value. The function was given by:

$$cross\_connection(\mathbf{a}) = \begin{cases} true, & \textit{if } \ a_i \neq a_{i+1} \text{ and } a_i = a_{i+2} \\ false, & else \end{cases} . \tag{10}$$

3.1.4 Connection Directions

There are two connection directions in word alignment: from Chinese to English, (where Chinese is the source language while English is the target language), and from English to Chinese. The alignment method of the IBM models has a restriction; a word of target language can only be connected to exactly one word of the source language. This restriction causes two words in the source language not to be able to connect to a word in the target language.

For example, in Figure 11, for alignment from Chinese to English, *cedar* should be connected to both 雪 and 松, but the model does not allow the connection in this direction. Figure 12 is another example of the same problem from English to Chinese.

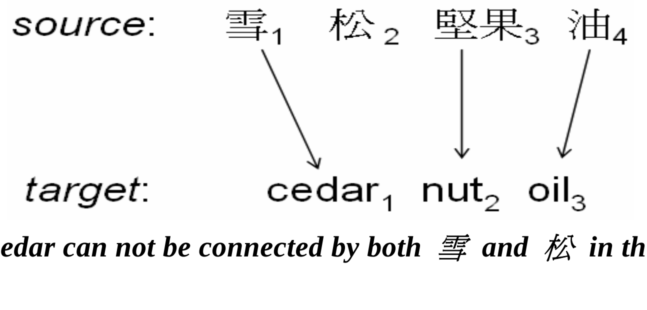


Figure 11. cedar can not be connected by both 雪 and 松 in this direction.

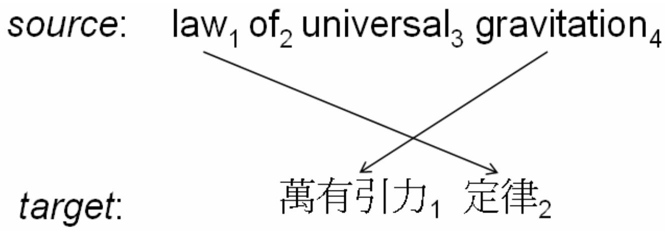


***Figure 12.**萬有引力* **can not be connected by both universal and gravitation in this direction.**

In order to solve this problem, the alignments of these two directions are merged using the following steps: 1. Align from Chinese to English. Each word of an English compound will be connected by the same Chinese word in this step which will be treated as an alignment unit in the next step. 2. Align from English to Chinese. Each word of a Chinese compound will be connected to the same English unit, a word or merged compound, in this step.

For example, *universal gravitation* was merged in step 1 while *雪* and *松* were not merged in the same step, as shown in Figure 13. In step2, *雪* and *松* were merged and *universal gravitation* will be treated as a unit in the same step, as shown in Figure 14.

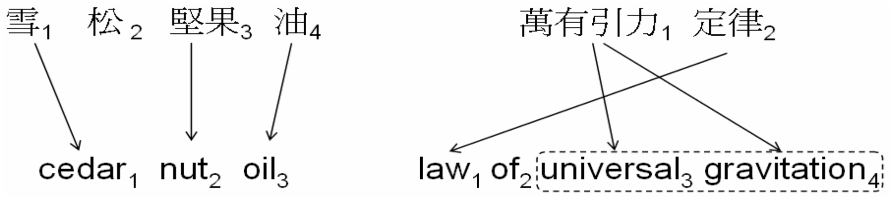


Figure 13.* 雪 *and* 松 *were not merged in step 1 while universal gravitation was merged in the same step.

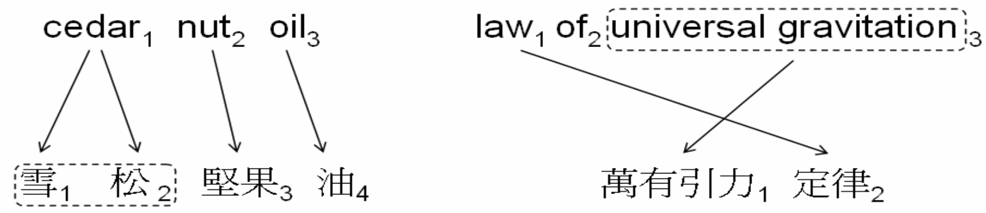


Figure 14. step 2,* 雪 *and* 松 *were merged in step 2 and universal gravitation was treated as a unit in the same step.

After these two steps, all of the compounds in each language will be merged. Figure 15 shows some examples of word alignment in these experiments.

English Term	Chinese Term	Alignment
evaporation tank	蒸發 槽	evaporation/蒸發 tank/槽
wind-wave tank	風浪 水槽	wind-wave/風浪 tank/水槽
wave tank	波浪 水槽	wave/波浪 tank/水槽
volumetric tank	量 水箱	volumetric/量 tank/水箱
curve of learning	學習 曲線	curve/曲線 of/ learning/學習
exchange of students	學生 交換	exchange/交換 of/ students/學生
practice teaching	教學 實習	practice/實習 teaching/教學
wall cloud	雲 牆	wall/牆 cloud/雲
gas mixture	混合 氣體	gas/氣體 mixture/混合
air choke valve	阻 氣 閥	air/氣 choke/阻 valve/閥

Figure 15. some examples of word alignment.

3.2 Sense Tagging

When we tag Chinese words with WordNet senses, if the translation of a word has only one sense, a monosemous word, it can be tagged with that sense directly. If the translation has more than one sense, we should use a disambiguation method to get the appropriate sense. In past years, a lot of word sense disambiguation (WSD) methods have been proposed, including supervised, bootstrapping, and unsupervised. Supervised and bootstrapping methods usually resolve an ambiguity in the collocations of the target word, which implies that the target word should be in a complete sentence. These are not appropriate for this project’s data. When some statistical based unsupervised methods are not accurate enough, they will add too much noise to the results. For the purpose of building a high quality dictionary, we tend to use a high precision WSD method which should also be appropriate for a bilingual term bank. We employ some heuristic rules, which are motivated by [Atserias *et al*. 1997], described as follows:

Heuristic 1.

If e_i is a morpheme of **e** then pick the sense of e_i, say s_j, which contains hyponym **e**.

This heuristic rule works for head morphemes of compounds. For example, as shown in figure 16, the term pair (*water tank*, *水 槽*) is aligned as (*water*/*水 tank/槽*). There are five senses for *tank*. The above heuristic rule will select *tank-2* as the sense of *tank/槽* because there is only one sense of *water tank* and the sense is a hyponym of *tank-2*. In this case, the sense of *water tank* can be tagged as *water tank-1* and *tank* can be tagged as *tank-2*.

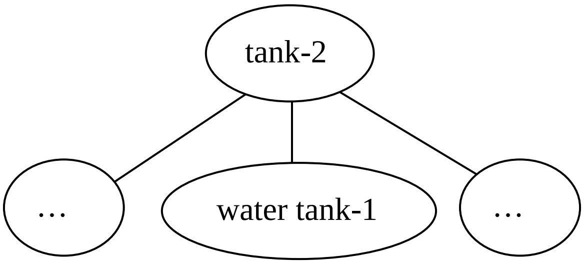


Figure 16. water tank-1 is a hyponym of tank-2.

Heuristic 2.

Suppose the set {$e_1, e_2, \ldots, e_k$} contains all possible translations of Chinese word *c*,

Case 1: If {$e_1, e_2, \ldots, e_k$} share a common sense s_t, then pick s_t as their sense.

Case 2: If one element of the set {$e_1, e_2, \ldots, e_k$}, say e_i, has a sense s_t which is the hypernym of synsets corresponding to the rest of the words. We say that they nearly share the same sense and pick s_t as the sense e_i, pick the corresponding hyponyms as the sense of the rest of words.

An example of case 1 is the translations of *腳踏車*, {*bicycle*, *bike*, *wheel*}, which are a subset of a synset. This means that the synset is the common sense of these words and we can pick it as the words' sense. An example of case 2, as shown in figure 17, is the translations of *信號旗*, {*signal*, *signal flag*, *code flag*}, although these words do not exactly share the same sense, one sense of *signal* is the hypernym of *signal flag* and *code flag*. This means that they nearly share the same sense; we pick the hypernym, *signal-1*, as the sense of *signal* and the corresponding hyponyms as the sense of *signal flag* and *code flag*.

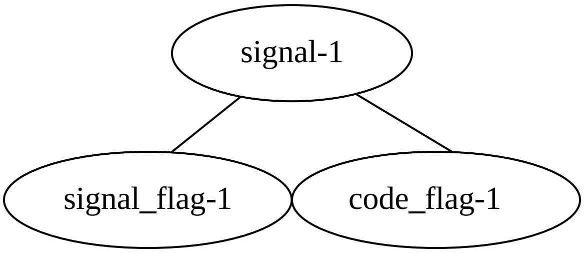


Figure 17. the translations of 信號旗, {signal, signal flag, code flag}, are nearly share the same sense.

Heuristic 3.

If some of the translations of *c* are tagged in the previous steps and the results show that the translations of *c* is always tagged with the same sense, we think *c* to have mono sense, so pick that sense as the sense of untagged translations.

In the previous steps, many Chinese-English pairs have been tagged with WordNet senses. In these tagged instances, we found that some Chinese words were always tagged with the same synset, although they may have many different English translations, and these English words may be ambiguous themselves. The untagged translations of the Chinese word can be tagged with the same synset.

For example, as shown in Figure 18, *防波堤* has many different translations and some of them are ambiguous in WordNet, (*groin* has 3 senses in WordNet). In fact, those seemingly different senses tagged by previous steps actually are indexed by the same synset in WordNet, so we guess that *防波堤* has mono sense and will be tagged the same synset for all instances.

Chinese word	**English word**	**Sense**
防波堤	breakwater	breakwater-1
防波堤	groin	groin-2
防波堤	groyne	groyne-1
防波堤	mole	mole-5
防波堤	bulwark	bulwark-3
防波堤	seawall	seawall-1
防波堤	jetty	jetty-1

Figure 18. the possible translations of 防波堤 and its sense tagged by the previous steps.

4. Experiments

In the experiment of word alignment, we extract 840,187 English-Chinese translation pairs which contain 445,830 Chinese word types and 318,048 English word types. On average, each Chinese word has 1.88 English translations while each English word has 2.64 Chinese translations.

In word sense disambiguation, 124,752 Chinese words were linked to 42,589 WordNet synsets, which contain 165,775 (Chinese word, synset) translation pairs. On average, each Chinese word was discovered to have 1.33 senses in terms of WordNet synsets. In the following subsection, we will evaluate the performance of the word alignments and WSD results.

4.1 Results of Word Alignment

In order to evaluate the performance of word alignment, we randomly select 500 term pairs from a terminology bank and align them manually as the gold standard, As single-morpheme terms do not need to be aligned, compound words were considered only. We follow the evaluation method defined by [Och and Ney 2000], which defined precision, recall and alignment error rate (AER) as follows:

$$\text{recall} = \frac{|A \cap S|}{|S|},$$

$$\text{precision} = \frac{|A \cap P|}{|A|},$$

$$\text{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|},$$

where S denotes the annotated set of sure alignments, P denotes the annotated set of possible alignments, and A denotes the set of alignments produced by the alignment method.

The results are shown in Table 1. The recall and precision figures show that the word alignment results are quite accurate. As we expected, the word alignment in phrases is much easier and accurate than in complete sentences. Note that the f-scores of word alignment tasks in complete sentences, even the current state-of-the-art alignments for naturally related languages such as English and French, are still less than 95 [Blunsom *et al*. 2006].

Table 1. the performance of our word alignment method.

recall	precision	f-score	AER
98.2	98.6	98.4	1.6

Table 2. typical errors of word alignment.

Error Type	Error Samples
Word Segmentation	half-wave/半 length/波長 criterion/準則 spiral/螺旋 coal/煤機 cleaner/洗 american/西 ginseng/洋參 second/再 wind/生氣 microlen/微透鏡藕 coupler/合器 atomic/原子能 energy/階
transliteration	san/聖胡 julian/連安
asymmetric translation	navigation/航行參考 star/星
abbreviation	double/ III/托克馬克熱核反應器

The main alignment errors are caused by the following reasons as shown in Table 2. The first error type was caused by the errors of word segmentation. For example, *西洋參* should be segmented as *西洋 參* instead of *西 洋參* and *再生氣* should be segmented as *再生氣* instead of *再 生氣*. The second error type was the mapping of transliterations which is a different type of word alignment. The third type was caused by the asymmetric translation of the data. For example, in the term pair (navigation star, 航行 參考 星), the Chinese word 參考 has no appropriate mapping in the English portion. The fourth type was caused by abbreviation which is also a difficult problem in regards to word alignment.

4.2 Result of Word Sense Disambiguation

Since the goal of these experiments is to build a Chinese WordNet automatically, we concerned more with the quality of WSD than the quantity. To evaluate the accuracy of these heuristic rules, we randomly selected 200 sense tagged words for each heuristic rule and checked the sense of each word manually. The accuracy rate of WSD results are defined as follows:

$$\text{accuracy rate} = \frac{\text{\# of selected words with correct sense}}{\text{\# of selected words}}.$$

The accuracy of each heuristic rule is shown in Table 3. It shows that the accuracy of heuristic rules is all over 80 %. Note that, in the lexical sample tasks of Senseval 3 [Mihalcea *et al*. 2004], the precision of the best supervised WSD methods is less than 73%, the unsupervised methods are even worse. Furthermore, these methods depend highly on the contexts of target words, which is not suitable in these experiments. These are the reasons why we use the heuristic rules instead of conventional WSD methods.

Table 3. Disambiguation accuracy of each heuristic rule.

	# words	#words with correct sense	accuracy rate
Heuristic 1	200	160	80.0 %
Heuristic 2	200	167	83.5 %
Heuristic 3	200	174	87.0 %

We also concerned with how many WordNet senses can be linked with Chinese words. There are two coverage rates, defined as follows:

$$\text{coverage rate of word-sense pairs} = \frac{\text{\# of word sense pairs are linked}}{\text{\# of word sense pairs in WordNet}},$$

$$\text{coverage rate of synsets} = \frac{\text{\# of synsets are linked}}{\text{\# of synsets in WordNet}}.$$

In the WSD steps, 484,771 tokens are tagged with WordNet synsets, in which 54,654 distinct word-sense pairs are contained. In other words, there are 54,654 distinct word-sense pairs which are linked with any Chinese word. The coverage of word-sense pairs and synsets are shown in Table 4. The synset coverage of heuristic rule 3 is not listed in the table, because it just tags the Chinese words which have been disambiguated in the previous steps and does not link any Chinese word with new synset. The table shows that the coverage of word-sense pairs in WordNet 2.0 is 26.9% and the coverage of synsets is 36.89 %.

Table 4. the coverage of each heuristic rule in WordNet 2.0.

	#tokens	#word-sense pairs	word-sense pair coverage	#synsets	synset coverage
monosemous word	370,991	48,623	23.94 %	39,953	34.61 %
Heuristic 1	29,422	4,211	2.07 %	3,452	2.99 %
Heuristic 2	29,311	2,050	1.00 %	1,685	1.46 %
Heuristic 3	81,734	1,931	0.95 %	-	-
Total	484,771	54,654	26.90 %	42,589	36.89 %

It seems the coverage of the experiments is too low. One possible reason is that most of the synsets in WordNet are infrequent. To prove this phenomenon, we use the frequencies of each sense provided by WordNet, which are the occurrence frequencies for each synset in the SemCor Corpus. As per analysis, there are 115,423 synsets in WordNet 2.0, but only 28,688 (24.8%) synsets appear in the SemCor. It shows that most of the senses are low frequency senses in WordNet.

Another issue is that, the coverage is contributed mostly by monosemous words. About 17% of words are ambiguous in WordNet. It seems that there is still room to improve.

5. Conclusions and Future Researches

In this paper, we propose a methodology to extract Chinese-English translation pairs from a large-scale bilingual terminology bank, and link the translation pairs to WordNet synsets. We faced two problems in this study: 1. Word-to-word alignment for each entry in the terminology bank, which helps to extract corresponding English translations for each Chinese word. 2. Word sense disambiguation, which helps to select the appropriate sense when the English translation of a Chinese word is ambiguous.

The evaluation of the experiments shows that the f-score of word alignment archives 98.4%. In the word sense disambiguation stage, the word-sense pairs extracted from the terminology bank cover 26.9% of WordNet word-sense pairs. Also, the distinct senses cover 36.89% of WordNet synsets. The accuracy of the three heuristic rules achieves 80%, 83 %, and 87 %.

A bilingual terminology bank provides some advantages over a bilingual parallel corpus for extracting information. For example, we can extract more Chinese-English translation pairs through the various appearances of a word which is contained in different compounds. The other advantage is that most of compound words in terminology bank are composed of only 2-3 words, which results in the word alignment accuracy of a terminology bank being much higher than a bilingual corpus.

In the future we will try to use some other word sense disambiguation methods to increase the coverage of words and senses in WordNet and to extract more information from terminology bank.

References

Atserias, J., S. Climent, X. Farreres, G. Rigau and H. Rodríguez, “Combining Multiple Methods for the Automatic Construction of Multilingual WordNets,” In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 1997, Tzigov Chark, Bulgaria, pp. 143-149.

Bhattacharya, I., L. Getoor and Y. Bengio, “Unsupervised Sense Disambiguation Using Bilingual Probabilistic Models, ” In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004, Barcelona, Spain, pp. 287-294.

Blunsom, P. and T. Cohn, “Discriminative Word Alignment with Conditional Random Fields,” In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006, Sydney, Australia, pp. 65-72.

Brown, P.F., S.A.D. Pietra, V.J.D. Pietra, and R.L. Mercer, “The Mathematics of Machine Translation: Parameter Estimation,” *Computational Linguistics*, 19(2), 1993, pp. 263–311.

Chang, J.S., T. Lin, G.-N. You, T.C. Chuang and C.-T. Hsieh, “Building A Chinese WordNet Via Class-Based Translation Model,” *International Journal of Computational Linguistics and Chinese Language Processing*, 8(2), 2003, pp. 61-76.

CKIP, Chinese Word Segmentation System, http://ckipsvr.iis.sinica.edu.tw/, 2006

Daudé, J., L. Padró and G. Rigau, “Mapping Multilingual Hierarchies using Relaxation Labelling,” In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, College Park, Maryland.

Dempster, A.P., N.M. Laird and D.B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, 39(1), 1977, pp. 1-38.

Diab, M. and P. Resnik, “An Unsupervised Method for Word Sense Tagging using Parallel Corpora,” In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, NJ, USA, pp. 255-262.

Mihalcea, R. and T. Chklovski, “The SENSEVAL–3 English Lexical Sample Task,” In *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain, pp. 25-28.

Miller, G., “WordNet: An online lexical database,” *International Journal of Lexicography*, 3(4), 1990, pp. 235-312.

NICT, 學術名詞資訊網, http://terms.nict.gov.tw/, 2006.

Och, F.J. and Hermann N., "Improved Statistical Alignment Models," In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, Hong Kong, pp. 440-447.

Proctor, P., "Longman English-Chinese Dictionary of Contemporary English," *Longman Group (Far East) Ltd*., Hong Kong, 1988.

Vogel, S., H. Ney, C. Tillmann, "HMM-based word alignment in statistical translation," In *Proceedings of the 16th conference on Computational linguistics*, 1996, Morristown, NJ, pp. 836-841.