Computational Linguistics and Chinese Language Processing
Vol. 8, No. 2, August 2003, pp. 61-76 


# Building A Chinese WordNet Via Class-Based Translation Model

**Jason S. Chang[*], Tracy Lin[+], Geeng-Neng You[**],**

**Thomas C. Chuang[++], Ching-Ting Hsieh[***]**

## Abstract

Semantic lexicons are indispensable to research in lexical semantics and word sense disambiguation (WSD). For the study of WSD for English text, researchers have been using different kinds of lexicographic resources, including machine readable dictionaries (MRDs), machine readable thesauri, and bilingual corpora. In recent years, WordNet has become the most widely used resource for the study of WSD and lexical semantics in general. This paper describes the Class-Based Translation Model and its application in assigning translations to nominal senses in WordNet in order to build a prototype Chinese WordNet. Experiments and evaluations show that the proposed approach can potentially be adopted to speed up the construction of WordNet for Chinese and other languages.

## 1. Introduction

WordNet has received widespread interest since its introduction in 1990 [Miller 1990]. As a large-scale semantic lexical database, WordNet covers a large vocabulary, similar to a typical college dictionary, but its information is organized differently. The synonymous word senses are grouped into so-called synsets. Noun senses are further organized into a deep IS-A hierarchy. The database also contains many semantic relations, including hypernyms, hyponyms, holonyms, meronyms, etc. WordNet has been applied in a wide range of studies on

[*] Department of Computer Science, National Tsing Hua University
101, Sec. 2, Kuang Fu Road, Hsinchu, Taiwan, ROC E-mail: jschang@cs.nthu.edu.tw
[+] Department of Communication Engineering, National Chiao Tung University
1001, University Road, Hsinchu, Taiwan, ROC E-mail: tracylin@mail.nctu.edu.tw
[**] Department of Information Manangement, National Taichung Institute of Technology
San Ming Road, Taichung, Taiwan, ROC E-mail: gny@mail.ntit.edu.tw
[++] Dept of Computer Science, Van Nung Institute of Technology
1 Van-Nung Road, Chung-Li, Taiwan, ROC E-mail: tomchuang@cc.vit.edu.tw
[***] Panasonic Taiwan Laboratories Co., Ltd. (PTL) E-Mail: chingting@ptl.com.tw

such topics as word sense disambiguation [Towell and Voothees, 1998; Mihalcea and Moldovan, 1999], information retrieval [Pasca and Harabagiu, 2001], and computer-assisted language learning [Wible and Liu, 2001].

Thus, there is a universally shared interest in the construction of WordNet in different languages. However, constructing a WordNet for a new language is a formidable task. To exploit the resources of WordNet for other languages, researchers have begun to study ways of speeding up the construction of WordNet for many European languages [Vossen, Diez-Orzas, and Peters, 1997]. One of many ways to build a WordNet for a language other than English is to associate WordNet senses with appropriate translations. Many researchers have proposed using existing monolingual and bilingual Machine Readable Dictionaries (MRD) with an emphasis on nouns [Daude, Padro & Rigau, 1999]. Very little study has been done on using corpora or on covering other parts of speech, including adjectives, verbs, and adverbs. In this paper, we describe a new method for automating the process of constructing Chinese WordNet. The method was developed specifically for nouns and is capable of assigning Chinese translations to some 20,000 nominal synsets in WordNet.

The rest of this paper is divided into four sections. The next section provides the background on using a bilingual dictionary to build a Chinese WordNet and semantic concordance. Section 3 describes a class-based translation model for assigning translations to WordNet senses. Section 4 describes the experimental setup and results. A conclusion is provided in Section 5 along with directions of future work.

## 2. From Bilingual MRD and Corpus to Bilingual Semantic Database

In this section, we describe the proposed method for automating the construction process of a Chinese WordNet. We have experimented to find the simplest way of attaching an appropriate translation to each WordNet sense under a Class-Based Translation Model. The translation candidates are taken from a bilingual word list or Machine Readable Dictionaries (MRDs). We will use an example to show the idea, and a formal description will follow in Section 3.

***Table 1.*** *Words in the same conceptual class that often share common Chinese characters in their translations.*

| Code (set title) | Hyponyms | Chinese translation |
|---|---|---|
| fish (aquatic vertebrate) | carp | 鯉魚 |
| fish (aquatic vertebrate) | catfish | 鯰魚 |
| fish (aquatic vertebrate) | eel | 鰻魚 |
| complex (building) | factory | 工廠 |
| complex (building) | cannery | 罐頭工廠 |
| complex (building) | mill | 製造廠 |
| speech (communication) | discussion | 討論;議論 |

| | | |
|---|---|---|
| speech (communication) | argument | 論據;論點;爭論 |
| speech (communication) | debate | 辯論 |

Let us consider the example of assigning appropriate translations for the nominal senses of "plant" in WordNet 1.7.1. The noun "plant" in WordNet has four senses:

1. plant, works, industrial plant (buildings for carrying on industrial labor);
2. plant, flora, plant life (a living organism lacking the power of locomotion);
3. plant (something planted secretly for discovery by another person);
4. plant (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience).

The following translations are listed for the noun "plant" in the Longman Dictionary of Contemporary English (English-Chinese Edition) [Longman Group 1992]:

1. 植物, 2. 設備, 3. 機器, 4. 工廠, 5. 內線人, and 6. 栽的贓 .

For words such as "plant" with multiple senses and translations, the question arises: Which translation goes with which synset? We make the following observations that are crucial to the solution of the problem:

1. Each nominal synset has a chain of hypernyms which give ever more general concepts of the word sense. For instance, *plant*-1 is a *building complex*, which in turn is a *structure* and so on and so forth, while *plant*-2 can be generalized as *a life form*.
2. The hyponyms of a certain top concept in WordNet form a set of semantically related word senses.
3. Semantically related senses tend to have surface realization in Chinese with shared characters.

For instance, *building complex* spawns the hyponyms *factory*, *mill*, *assembly plant*, *cannery*, *foundry*, *maquiladora*, etc., all of which realize in Chinese using the characters "廠" or "工廠." Therefore, we can say that there is a high probability that senses which are direct or indirect hyponyms of *building complex* share the Chinese characters "工" and "廠" in their Chinese translations. Therefore, it is clear that one can determine that *plant*-1, a hyponym of *building complex*, should have "工廠" instead of "植物" as its translation. See Table 1 for more examples. That intuition can be expanded into a systematic way of assigning the most appropriate translation to a given word sense. Figure 1 shows how the method works for four senses of *plant*.

In the following, we will consider the task of assigning the most appropriate translation to *plant*-1, the first sense of the noun "plant." First, the system looks up "plant" in the Translation Table (T Table) for candidate translations of *plant*-1:

(*plant*, 植物), (*plant*, 機器), (*plant*, 設備), (*plant*, 工廠), (*plant*, 內線人), (*plant*, 栽的贓).

Next, the semantic class *g* to which *plant*-1 belongs is determined by consulting the Semantic Class Table (SC Table). In this study we use some 1,145 top hypernyms *h* to represent the class of word senses that are direct or transitive hyponyms of *h*. The path designator of *h* in WordNet is used to represent the class. The hypernyms are chosen to correspond roughly to the division of sets of words in the Longman Lexicon of Contemporary English (LLOCE) [McArthur 1992]. Table 2 provides examples of classes related to *plant* and their class codes.

***Table 2.*** *Words in four classes related to the noun* ***plant****.*

| English | WN sense | Class Code | Words in the Class |
|---|---|---|---|
| Plant | 1 | N001004003030 | factory, mill, assembly plant, … |
| Plant | 2 | N001001005 | flora, plant life, … |
| Plant | 3 | N001001015008 | thought, idea, … |
| Plant | 4 | N001001003001001 | producer, supernatural, … |
| Plant | 4 | N001003001002001 | announcer, conceiver, … |

For instance, *plant*-1 belongs to the class *g* represented by the WordNet synset (*structure*, *construction*):

*g* = N001004003030.

Subsequently, the system evaluates the probabilities of each translation conditioned on the semantic class *g*:

P(“植物” | N001004003030),

P(“機器” | N001004003030),

P(“設備” | N001004003030),

P(“工廠” | N001004003030),

P(“內線人” | N001004003030),

P(“栽的贓” | N001004003030).

These probabilities are not evaluated directly. The system takes apart the characters in a translation and looks up P( *u* | *g* ), the probabilities for each translation character *u* conditioned on *g*:

P(“植” | **N001004003030**) = 0.000025,

P(“物” | **N001004003030**) = 0.000025,

P(“機” | **N001004003030**) = 0.00278,

P(“器” | **N001004003030**) = 0.00278,

P(“設” | **N001004003030**) = 0.00306,

P(“備” | **N001004003030**) = 0.00075,

P(“工” | **N001004003030**) = 0.00711,

P(“廠” | **N001004003030**) = 0.01689,

P(“內” | **N001004003030**) = 0.00152,

P(“線” | **N001004003030**) = 0.00152,

P(“人” | **N001004003030**) = 0.00152,

P(“栽” | **N001004003030**) = 0.00152,

P(“的” | **N001004003030**) = 0.00152,

P(“贓” | **N001004003030**) = 0.00152.

Note that to deal with lookup failure, a smoothing probability is given (0.000025, derived using the Good-Turing method). By using a statistical estimate based on simple linear interpolation, we can get

$$\begin{aligned} &P(\text{“工廠”} \mid \textit{plant}\text{-}1) \approx P(\text{“工廠”} \mid \textbf{N001004003030}) \\ &\approx \frac{1}{2} P(\text{“工”} \mid \textbf{N001004003030}) + \frac{1}{2} P(\text{“廠”} \mid \textbf{N001004003030}) \\ &= \frac{1}{2}(0.0178+0.0073) = 0.0124. \end{aligned}$$

Similarly, we have

P(“植物” | **N001004003030**) = 0.0013,

P(“機器” | **N001004003030**) = 0.0023,

P(“設備” | **N001004003030**) = 0.0028,

P(“內線人” | **N001004003030**) = 0.0014,

P(“栽的贓” | **N001004003030**) = 0.0001.

Finally, by choosing the translation with the highest probabilistic value for *g*, we can get an entry for Chinese WordNet (CWN Table):

(plant, 工廠, n, 1, “buildings for carrying on industrial labor”)

After we get the correct translation of *plant*-1 and many other word senses in *g*, we will be able to re-estimate the class-based translation probability for *g* and produce a new CT Table. However, the reader may wonder how we can get the initial CT Table. This dilemma can be resolved by adopting an iterative algorithm that establishes an initial CT Table and makes revision until the values in the CT Table converge. More details will be provided in Section 3.

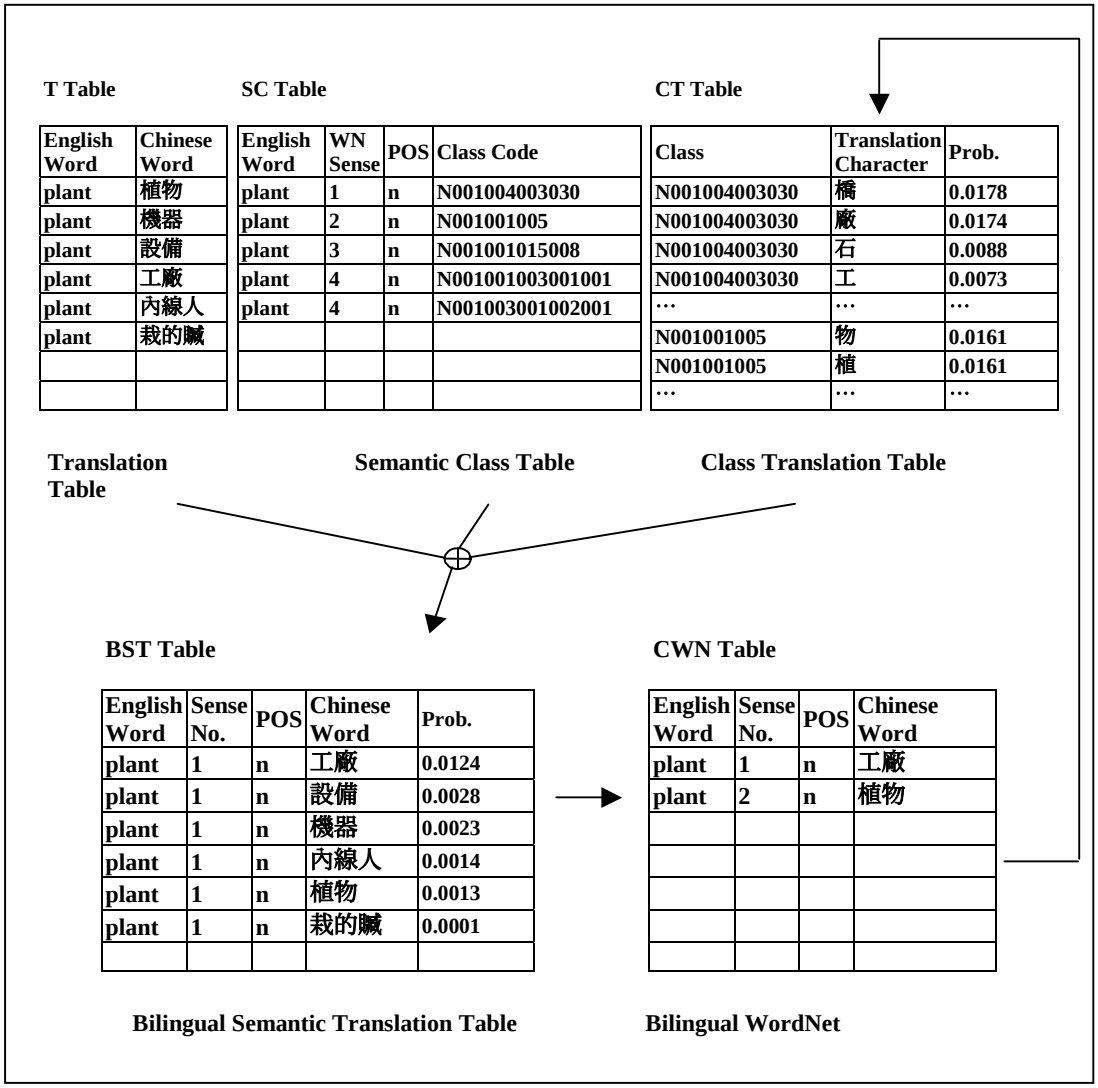


***Fig. 1*** *Using CBTM to build Chinese WordNet. This example shows how the first sense of plant receives an appropriate translation via the Class-Based Translation Model and how the model can be trained iteratively.*

## 3. The Class-Based Translation Model

In this section, we will formally describe the proposed class-based translation model, how it can be trained, and how it can be applied to the task of assigning appropriate translations to different word senses. Given $E_k$, the *k*th sense of an English word *E* in the WordNet, the probability of its Chinese translation is denoted as $P(C \mid E_k)$. Therefore, the best Chinese

translation $C^*$ is

$$C^*(E_k) \cong \arg\max_{C \in T(E)} P(C \mid E_k), \quad (1)$$

where $T(X)$ is the set of Chinese translations of sense X listed in a bilingual dictionary.

Based on our observation that semantically related senses tend to be realized in Chinese using shared Chinese characters, we tie together the probability functions of translation words in the same semantic class and use the class-based probability as an approximation. Thus, we have

$$P(C \mid E_k) \cong P(C \mid g), \quad (2)$$

where $g = g(E_k)$ is the semantic class containing $E_k$.

The probability of $P(C|g)$ can be estimated using the Expectation and Maximization Algorithm as follows:

(Initialization) $P(C \mid E_k) = \frac{1}{m}$, $m = |T(E)|$ and $C \in T(E)$; (3)

(Maximization) $$P(C \mid g) = \frac{\sum_{E,k,i} P(C_i \mid E_k) I(C = C_i) I(E_k \in g)}{\sum_{E,k,i} P(C_i \mid E_k) I(E_k \in g)}, \quad (4)$$

where $C_i$ = the $i$th translation of $E_k$ in $T(E_k)$,

$I(x) = 1$ if $x$ is true and 0 otherwise;

(Expectation) $P_1(C \mid E_k) = P(C \mid g)$, (5)

where $g = g(E_k)$ is the class that contains $E_k$;

(Normalization) $$P(C \mid E_k) = \frac{P_1(C \mid E_k)}{\sum_{D \in T(E_k)} P_1(D \mid E_k)}. \quad (6)$$

In order to avoid the problem of data sparseness, $P(C|g)$ is estimated indirectly via the unigrams and bigrams in $C$. We also weigh the contribution of each unigram and bigram to avoid the domination of a particular character in the semantic class. Therefore, we rewrite Equations 4 and 5 as follows:

(Maximization) $$P_u(u \mid g) = \frac{\sum_{E,k,i,j} \frac{1}{m} I(E_k \in g) I(u = u_{i,j}) P(u_{i,j} \mid E_k)}{\sum_{E,k,i,j} \frac{1}{m} I(E_k \in g) P(u_{i,j} \mid E_k)}, \quad (4a)$$

where $u_{i,j}$ = the $j$th unigram of the $i$th translation in $T(E_k)$,

$m$ = the number of characters in the $i$th translation in $T(E_k)$,

$$P_b(b \mid g) = \frac{\sum_{E,k,i,j} \frac{1}{m-1} I(E_k \in g) I(b = b_{i,j}) P(b_{i,j} \mid E_k)}{\sum_{E,k,i,j} \frac{1}{m-1} I(E_k \in g) P(b_{i,j} \mid E_k)}, \quad (4b)$$

where $b_{i,j}$ = the $j$th overlapping bigram of the $i$th translation in $T(E_k)$;

(Expectation) $$P_1(C \mid E_k) \cong P(C \mid g) \cong \sum_{i=1}^{m} \frac{P_u(u_i \mid g)}{m} \quad \text{(unigram)}, \quad (5a)$$

$$P_1(C \mid E_k) \cong P(C \mid g) \cong \sum_{i=1}^{m} \frac{P_u(u_i \mid g)}{2m} + \sum_{i=1}^{m-1} \frac{P_b(b_i \mid g)}{2(m-1)} \text{ (+bigram)}, \quad (5b)$$

where $u_i$ is a unigram, $b_i$ is an on overlapping bigram of $C$, and $m$ is the number of characters in $C$ .

For instance, assume that we have the first sense *trunk*-1 of the word *trunk* in WordNet and the translations in LDOCE as follows:

*trunk*-1 (the main stem of a tree; usually covered with bark; the bole is usually the part that is commercially useful for lumber),

Translations of *trunk* — 大皮箱，大衣箱，樹幹，and 象鼻 .

Initially, the probabilities of each translation for *trunk*-1 are as follows:

$P$( 大皮箱 | *trunk*-1 ) = 1/4, $P$( 大衣箱 | *trunk*-1 ) = 1/4,

$P$( 樹幹 | *trunk*-1 ) = 1/4, $P$( 象鼻 | *trunk*-1 ) = 1/4.

Table 3 shows the words in the semantic class N001004001018013014 (stalk, stem), containing *trunk*-1 and relevant translations. Following Equations 4a and 4b, we took the unigrams and overlapping bigrams from these translations to calculate the probability of unigram and bigram translations for (stalk, stem). Although initially irrelevant translations such as bulb-電燈泡(light bulb) can not be excluded, after one iteration of the maximization step, the noise is suppressed substantially, and the top ranking translations shown in Tables 4 and 5 seem to be the “genus” terms of the class. For instance, the top ranking unigrams for N001004001018013014 include 莖 (stem), 枝 (branch), 條 (branch), 根 (stump) 樹 (tree) 幹 (trunk) etc. Similarly, the top ranking bigrams include 球莖 (bulb), 樹枝 (branch), 柳條 (willow branch), and 樹幹 (trunk). All indicate the general concepts of the class.

With the unigram translation probability $P(u \mid g)$, one can apply Equations 5a and 6 to proceed with the Expectation Step and calculate the probability of each translation candidate for a word sense as shown in Example 1:

**Example 1.**

$P_1$(樹幹|*trunk*-1)=1/2*($P$(樹|N001004001018013014)+$P$(幹| N001004001018013014))

=1/2*(0.0145+0.0103) = **0.0124**,

$P_1$(象鼻|*trunk*-1) =1/2*(*P*(象|N001004001018013014)+*P*(鼻 |N001004001018013014 ))
=1/2* (0.00054+0.00054) = 0.00054,

$P_1$(大皮箱|*trunk*-1) =1/3*(*P*(大|N001004001018013014)+*P*(皮|N001004001018013014 )
+ *P*(箱 |N001004001018013014)) ,

=1/3*(0.0074+0.00036+0.00072) = 0.00283,

$P_1$(大衣箱|*trunk*-1) =1/3*(*P*(大|N001004001018013014)+*P*(衣|N001004001018013014 )
+ *P*(箱 | N001004001018013014))

=1/3*(0.0074 + 0.00043 + 0.00072) = 0.00285

*P* ( 樹幹 | *trunk*-1 ) = 0.0124/(0.0124+0.00054+0.00283+0.00285) = 0.665950591,

*P* ( 象鼻 | *trunk*-1 ) = 0.0124/(0.0124+0.00054+0.00283+0.00285) = 0.0290010741,

*P* ( 大皮箱 | *trunk*-1 ) = 0.0124/(0.0124+0.00054+0.00283+0.00285) = 0.1519871106,

*P* ( 大衣箱 | *trunk*-1 ) = 0.0124/(0.0124+0.00054+0.00283+0.00285) = 0.1530612245.

Using simple linear interpolation of translation unigrams and bigrams (Equation 5b), the probability of each translation candidate for a word sense can be calculated as shown in Example 2:

**Example 2.**

$P_1$( 樹幹 | *trunk*-1 ) = 1/2 * {1/2 * (*P*( 樹 | N001004001018013014 )
+*P*( 幹 | N001004001018013014 ) )
+*P*( 樹幹 | N001004001018013014 ) }

= 1/2 * (0.0124 + 0.0145) = **0.01345**,

$P_1$( 象鼻 | *trunk*-1 ) = 1/2 * {1/2 * (*P*( 象 | N001004001018013014 )
+*P*( 鼻 | N001004001018013014 ) )
+*P*( 象鼻 | N001004001018013014 ) }

= 1/2 * (0.00054 + 0.00107) = 0.000805,

$P_1$( 大皮箱 | *trunk*-1 ) = 1/2 * {1/3 * (*P*( 大 | N001004001018013014 )
+ *P*( 皮 | N001004001018013014 ))
+ *P*( 箱 | N001004001018013014 )}
+ 1/2 * (*P*( 大皮 | N001004001018013014 )
+*P*( 皮箱 | N001004001018013014 ) ) }

= 1/2 * (0.00283 + 0.00054) = 0.001685,

$P_1$( 大衣箱 | *trunk*-1 ) = 1/2 * {1/3 * (*P*( 大 | N001004001018013014 )
+ *P*( 衣 | N001004001018013014 ))
+ *P*( 箱 | N001004001018013014 ) }
+ 1/2 * (*P*( 大衣 | N001004001018013014 )

+*P*( 衣箱 | N001004001018013014 ) ) }

= 1/2 * (0.00285 + 0.00054) = 0.001695

*P* (樹幹|*trunk*-1) = 0.01345/(0.01345+0.000805+0.001685+0.001695)= 0.76268783669,

*P* (象鼻|*trunk*-1) = 0.000805/(0.01345+0.000805+0.001685+0.001695)

= 0.045647859371,

*P* (大皮箱|*trunk*-1) = 0.001685/(0.01345+0.000805+0.001685+0.001695)

= 0.095548624894,

*P* (大衣箱|*trunk*-1) = 0.001695/(0.01345+0.000805+0.001685+0.001695)

= 0.096115679047.

***Table 3.*** *Words and their translations in the semantic class N001004001018013014*

| English *E* | WN sense *k* | G($E_k$) | Chinese Translation |
|---|---|---|---|
| Beanstalk | 1 | N001004001018013014 | 豆莖 |
| Bole | 2 | N001004001018013014 | 樹幹 |
| Branch | 2 | N001004001018013014 | 分枝 |
| Branch | 2 | N001004001018013014 | 部門 |
| Branch | 2 | N001004001018013014 | 樹枝 |
| Brier | 2 | N001004001018013014 | 荊棘 |
| Bulb | 1 | N001004001018013014 | 球莖狀物 |
| Bulb | 1 | N001004001018013014 | 電燈泡 |
| Cane | 2 | N001004001018013014 | 籐條 |
| Cutting | 2 | N001004001018013014 | 剪報 |
| Cutting | 2 | N001004001018013014 | 插枝 |
| Stick | 2 | N001004001018013014 | 小樹枝 |
| Stick | 2 | N001004001018013014 | 手杖 |
| Stem | 2 | N001004001018013014 | 家系 |
| Stem | 2 | N001004001018013014 | 幹 |

***Table 4.*** *Probabilities of each unigram for the semantic class containing* trunk*-1, etc.*

| Unigram (u) | Semantic Class Code (g) | *P*( *u* \| *g* ) |
|---|---|---|
| 莖 | N001004001018013014 | 0.0706 |
| 枝 | N001004001018013014 | 0.0274 |
| 豆 | N001004001018013014 | 0.0216 |
| 條 | N001004001018013014 | 0.0162 |
| 樹 | N001004001018013014 | 0.0145 |
| 根 | N001004001018013014 | 0.0134 |

| | | |
|---|---|---|
| 幹 | N001004001018013014 | 0.0103 |
| 籐 | N001004001018013014 | 0.0080 |
| … | …… | … |

***Table 5.*** *Probabilities of each bigram for the semantic class containing* trunk*-1, etc.*

| **Bigram (b)** | **Semantic Class Code (g)** | ***P*( *b* \| *g* )** |
|---|---|---|
| 球莖 | N001004001018013014 | 0.0287 |
| 柳條 | N001004001018013014 | 0.0269 |
| 樹幹 | N001004001018013014 | 0.0145 |
| 樹枝 | N001004001018013014 | 0.0144 |
| 嫩枝 | N001004001018013014 | 0.0134 |
| … | …… | … |

Both examples show that the class-based translation model produces reasonable probabilistic values. The examples also show that for *trunk*-1, the linear interpolation method gives a higher probabilistic value for the correct translation "樹幹" than the unigram-based approach does (0.76268783669 vs. 0.665950591). In this case, linear interpolation is a better parameter estimation scheme. Our experiments showed, in general, that combining both unigrams and bigrams does lead to better overall performance.

## 4. Experiments

We carried out two experiments to see how well CBTM can be applied to assign appropriate translations to nominal senses in WordNet. In the first experiment, the translation probability was estimated using Chinese character unigrams, while in the second experiment, both unigrams and bigrams were used. The linguistic resources used in the experiments included:

1. **WordNet 1.6**: WordNet contains approximately 116,317 nominal word senses organized into approximately 57,559 word meanings (synsets).
2. **Longman English-Chinese Dictionary of Contemporary English (LDOCE E-C)**: LDOCE is a learner's dictionary with 55,000 entries. Each word sense contains information, such as a definition, the part-of-speech, examples, and so on. In our method, we take advantage of its wide coverage of frequently used senses and corresponding Chinese translations. In the experiments, we tried to restrict the translations to lexicalized words rather than descriptive phrases. We set a limit on the length of a translation: nine Chinese characters or less. Many of the nominal entries in WordNet are not covered by learner dictionaries; therefore, the experiments focused on those senses for which Chinese translations are available in LDOCE.
3. **Longman Lexicon of Contemporary English (LLOCE)**: LLOCE is a bilingual

taxonomy, which brings together words with related meanings and lists them in topical/semantic classes with definitions, examples, and illustrations.

The three tables shown in Figure 1 were generated in the course of the experiments:

1. The Translation Table has 44,726 entries and was easily constructed by extracting Chinese translations from LDOCE E-C [Proctor 1988].
2. We obtained the Sense Class Table by finding the common hypernyms of sets of words in LLOCE. 1,145 classes were used in the experiments.
3. The Class Translation Table was constructed using the EM algorithm based on the T Table and SC Table. The CT Table contains 155,512 entries.

Table 6 shows the results of using CBTM and Equation 1 to find the best translations for a word sense. We are concerned with the coverage of word senses in average text. In that sense, the translation of *plant*-3 is incorrect, but this error is not very significant, since this word sense is used infrequently. We chose the WordNet semantic concordance, SEMCOR, as our testing corpus. There are 13,494 distinct nominal word senses in SEMCOR. After the translation probability calculation step, our results covered 10,314 word senses in SEMCOR; thus, the coverage rate was 76.43%.

***Table 6.*** *The results and appropriate translations for each sense of the English word.*

| English | WN sense | Chinese Translation | Appropriate Chinese Translation |
|---|---|---|---|
| Plant | 1 | 工廠 | 工廠 |
| Plant | 2 | 植物 | 植物 |
| Plant | 3 | 內線人 | 栽的贓 |
| Plant | 4 | 內線人 | 內線人 |
| Spur | 1 | 鼓勵 | 鼓勵 |
| Spur | 2 | 激勵 | 刺, 針 |
| Spur | 4 | 馬刺 | 馬刺 |
| Spur | 5 | 支線 | 支線 |
| Bank | 1 | 銀行 | 銀行 |
| Bank | 2 | 邊坡 | 沙洲 |
| Bank | 3 | 庫 | 庫, 儲存所 |
| Scale | 1 | 記數法或基準 | 記數法或基準 |
| Scale | 2 | 比例 | 規模 |
| Scale | 3 | 比例 | 比例 |
| Scale | 5 | 脫下的乾燥皮屑 | 脫下的乾燥皮屑 |
| Scale | 6 | 音階 | 音階 |

To see how well the model assigns translations to WordNet senses appearing in average text, we randomly selected 500 noun instances from SEMCOR as our test data. There were 410 distinct words. Only 75 words had a unique sense in WordNet. There were 77 words with

two senses in WordNet, while 70 words had three senses in WordNet, and so on. The average degree of sense ambiguity was 4.2.

***Table 7****. The degree of ambiguity and number of words in the test data with different degree of ambiguity.*

| **Degree of ambiguity # of senses in WordNet** | **# of word types in the test data** | **Examples** |
|---|---|---|
| 1 | 75 | aptitude, controversy, regret |
| 2 | 77 | camera, fluid, saloon |
| 3 | 70 | drain, manner, triviality |
| 4 | 51 | confusion, fountain, lesson |
| 5 | 35 | isolation, pressure, spur |
| 6 | 25 | blood, creation, seat |
| 7 | 28 | column, growth, mind |
| 8 | 9 | contact, hall. program |
| 9 | 7 | body, company, track |
| 10 | 8 | bank, change, front |
| >10 | 25 | control, corner, deaft |

Among our 500 test data, 280 entries were the first sense, while 112 entries were the second sense. Over half of the words had the meaning of the first sense. Therefore, the first sense was most frequently used. Therefore, it was found to be more important to get the first and the second senses right. We manually gave each word sense an appropriate Chinese translation whenever one was available from LDOCE. From these translations, we found the following:

1. There were 491 word senses for which corresponding translations were available from LDOCE.
2. There were 5 word senses for which no relevant translations could be found in LDOCE due to the limited coverage of this learner's dictionary. Those word senses and relevant translations included assignment-2 (轉讓), marriage-3 (婚禮), snowball-1(繡球莢), prime-1(質數), and program-7 (政綱).
3. There were 4 words, that have no translations due to the particular cross-referencing scheme of LDOCE. Under this scheme, some nouns in LDOCE are not directly given a definition and translation, but rather a pointer to a more frequently used spelling. For instance, "groom" is given a pointer to "BRIDEGROOM" rather than the relevant definition and translation ("新郎").

In the first experiment, we started out by ranking the relevant translations for each noun sense using the class-based translation model. If two translations had the same probabilistic value, we gave them the same rank. For instance, Table 8 shows that the top 1 translation for *plant*-1 was "工廠."

***Table 8.*** *The rank of each translation corresponding to each word sense. (plant-2, 栽的贓) and (plant-2, 設備) have the same probability and rank.*

| English | Semantic class | WN sense | Chinese Translation | Probability | Rank |
|---|---|---|---|---|---|
| Plant | N001004003030 (structure) | 1 | 工廠 | 0.012372 | 1 |
| Plant | N001004003030 (structure) | 1 | 設備 | 0.002823 | 2 |
| Plant | N001004003030 (structure) | 1 | 機器 | 0.002270 | 3 |
| Plant | N001004003030 (structure) | 1 | 內線人 | 0.001375 | 4 |
| Plant | N001004003030 (structure) | 1 | 植物 | 0.001278 | 5 |
| Plant | N001004003030 (structure) | 1 | 栽的贓 | 0.000130 | 6 |
| Plant | N001001005 (flora) | 2 | 植物 | 0.016084 | 1 |
| Plant | N001001005 (flora) | 2 | 機器 | 0.002623 | 2 |
| Plant | N001001005 (flora) | 2 | 工廠 | 0.000874 | 3 |
| Plant | N001001005 (flora) | 2 | 設備 | 0.000525 | 4 |
| Plant | N001001005 (flora) | 2 | 栽的贓 | 0.000525 | 4 |
| Plant | N001001005 (flora) | 2 | 內線人 | 0.000360 | 5 |

***Table 9.*** *The recall rate in the first experiment*

| The number of top-ranking translations | Correct Entries (Total entries =500) | Recall rate (unigram) | Recall rate (unigram+bigram) |
|---|---|---|---|
| Top 1 | 344 | 68.8% | 70.2% |
| Top 2 | 408 | 81.6% | 83.2% |
| Top 3 | 441 | 88.2% | 89.0% |
| Top 4 | 449 | 89.8% | 91.4% |
| Top 5 | 462 | 92.4% | 93.2% |

We used the same method to evaluate the recall rate in the second experiment, where both unigrams and bigrams were used. The experimental results show a slight improvement over the results obtained using only unigrams.

In these experiments, we estimated the translation probability based on unigrams and bigrams. The evaluation results confirm our observation that we can exploit shared characters in translations of semantically related senses to obtain relevant translations. We evaluated the experimental results based on whether the Top 1 to Top 5 translations covered all appropriate translations. If we selected the Top 1 translation in the first experiment as the most appropriate translation, there were 344 correct entries, and the recall rate was 68.8%. The Top 2 translations covered 408 correct entries, and the recall rate was 81.6%. Table 9 shows the recall rate with regard to the number of top-ranking translations used for the purpose of evaluation.

## 5. Conclusion

In this paper, a statistical class-based translation model for the semi-automatic construction of a Chinese WordNet has been proposed. Our approach is based on selecting the appropriate Chinese translation for each word sense in WordNet. We observe that a set of semantically related words tend to share some Chinese characters in their Chinese translations. We propose to rely on the knowledge base of a Class Based Translation Model derived from statistical analysis of the relationship between semantic classes in WordNet and translations in the bilingual version of the Longman Dictionary of Contemporary English (LDOCE). We carried out two experiments that show that CBTM is effective in speeding up the construction of a Chinese WordNet.

The first experiment was based on the translation probability of unigrams, and the second was based on both unigrams and bigrams. Experimental results show that the method produces a Chinese WordNet covering 76.43% of the nominal senses in SEMCOR, which implies that a high percentage of the word senses can be effectively handled. Among our 500 testing cases, the recall rate was around 70%, 80% and 90%, respectively, when the Top 1, Top 2, and Top 3 translations were evaluated. The recall rate when using both unigrams and bigrams was slightly higher than that when using only unigrams. Our results can be used to assist the manual editing of word sense translations.

A number of interesting future directions present themselves. First, obviously, there is potential for combining two or more methods to get even better results in connecting WordNet senses with translations. Second, although nouns are most important for information retrieval, other parts of speech are important for other applications. We plan to extend the method to verbs, adjectives and adverbs. Third, the translations in a machine readable dictionary are at times not very well lexicalized. The translations in a bilingual corpus cauld be used to improve the degree of lexicalization.

## Acknowledgement

This study was partially supported by grants from the National Science Council (NSC 90-2411-H-007-033-MC) and the MOE (project EX 91-E-FA06-4-4).

## References

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

Daudé, J., L. Padró and G. Rigau, "Mapping WordNets using Structural Information," *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000.

McArthur, T., "Longman Lexicon of Contemporary English," Longman Group (Far East) Ltd., Hong Kong, 1992.

Mihalcea, R. and D. Moldovan., "A method for Word Sense Disambiguation of unrestricted text," *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999, pp. 152-158.

Miller, G., "Five papers on WordNet," *International Journal of Lexicography*, 3(4), 1990.

Pasca, M. and S. Harabagiu, "The Informative Role of WordNet in Open-Domain Question Answering," in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, June 2001, Carnegie Mellon University, Pittsburgh PA, pp. 138-143.

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

Towell, G. and E. Voothees, "Disambiguating Highly Ambiguous Words," *Computational Linguistics*, 24(1) 1998, pp. 125-146.

Vossen, P., P. Diez-Orzas and W. Peters, "The Multilingual Design of the EuroWordNet Database," *Processing of the IJCAI-97 workshop Multilingual Ontologies for NLP Applications*, 1997.

Wible, D. and A. Liu, "A syntax-lexical semantics interface analysis of collocation errors," *PacSLRF* 2001.