Building a Chinese-English WordNet for Translingual Applications

HSIN-HSI CHEN, CHI-CHING LIN, AND WEN-CHENG LIN
National Taiwan University

A WordNet-like linguistic resource is useful, but difficult to construct. This article proposes a method to integrate five linguistic resources, including English/Chinese sense-tagged corpora, English/Chinese thesauruses, and a bilingual dictionary. Chinese words are mapped into WordNet. A Chinese WordNet and a Chinese-English WordNet are derived by following the structures of WordNet. Experiments with Chinese-English information retrieval are developed to evaluate the applicability of the Chinese-English WordNet. The best model achieves 0.1010 average precision, 69.23% of monolingual information retrieval. It also gains a 10.02% increase relative to a model that resolves translation ambiguity and target polysemy problems together.

Categories and Subject Descriptors: H3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing - *Linguistic processing*; *Thesauruses*; H3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *Query formulation*; I2.7 [**Artificial Intelligence**]: Natural Language Processing - *Machine translation*; *Text analysis*
General Terms: Experimentation, Languages
Additional Key Words and Phrases: Bilingual WordNet, cross language information retrieval, sense tagging, thesaurus construction, translingual application, word sense disambiguation

1. INTRODUCTION

Translingual knowledge management is very important in the network era, where knowledge is disseminated easily and quickly to users of different languages via the Internet. Hence, linguistic resources are indispensable for translingual applications. WordNet, an electronic English lexical database [Fellbaum 1998], has been widely used to solve a variety of problems [Harabagiu 1998; Rila 1998; Ruiz et al. 1999], such as information retrieval, lexical acquisition, natural language generation, word sense disambiguation, and so on. Several works were proposed to extend the design idea of WordNet to other languages, e.g., Italian [Artale et al. 1997]; German [Hamp and Feldweg 1997], and so on. EuroWordNet [Vossen 1998], which aims to build a multilingual database with word nets for several European languages, is a successful example. Construction of word nets is a long-term task. Take WordNet as an example. Within four years, the scope of WordNet grew steadily from 13,688 glosses and 44,983 synsets (July 1991), 19,382 glosses and 49,771 synsets (January 1992), 36,880 glosses and 61,023 synsets (January 1993), 58,705 glosses and 79,542 synsets (January 1994), to 75,389 glosses and 91,050 synsets (January 1995). Thus how to build a word net for a specific language on the basis of WordNet is a

Authors' address: Department of Computer Science and Information Engineering, National Taiwan University, #1 Roosevelt Rd. Sec. 4, Taipei, Taiwan 106, ROC

research issue. Farreres et al. [1998] proposed a set of methodologies and techniques to construct multilingual word nets for Spanish and Catalan, which belong to the occidental family of languages.

A similar lexical database for Chinese has been unavailable up to now. The only public resources are tong2yi4ci2ci2lin2 ("同義詞詞林") abbreviated as Cilin [Mei et al. 1982] and HowNet [Dong and Dong 2000]. Cilin is composed of 12 large categories, 94 middle categories, 1,428 small categories, and 3,925 word clusters. Although it is a four-layer semantic structure, it does not provide relationships for hypernym, hyponym, similar, derived, antonym, and so on. HowNet is not a thesaurus [Dong and Dong 2000], but the authors attempted to construct a knowledge base from the interconcept relations and interattribute relations. Some work [Chen et al. 2000; Dorr et al. 2000] has been done to integrate these two linguistic resources with WordNet. Dorr et al. [2000] linked verb concepts between WordNet and HowNet; Chen et al. [2000] explored the mapping of Cilin to WordNet, including nouns, verbs, adjectives, and adverbs.

Two of the major problems in translingual knowledge management are translation ambiguity, which originates from the source language side, and target polysemy, which occurs in the target language [Chen et al. 1999]. Take cross-language information retrieval as an example. We have to disambiguate the meanings of source words and select the target words in the target language. The most suitable translation may have more than one sense. These two problems are often mixed together and have to be resolved both in building and in using a bilingual word net. For word sense disambiguation, the common approach is to postulate that each sense has a characteristic context that is different from the context of all the other senses [Yarowsky 1992]. In addition, all the words expressing the same sense share the same characteristic context. We can use a context vector to represent a sense. The context information can be obtained from a sense-tagged corpus, and thus a sense-tagged corpus is indispensable. In English, only some sense-tagged corpora such as HECTOR [Atkins 1993]; DSO [Ng and Lee 1996]; SEMCOR [Fellbaum 1998]; and SENSEVAL [Kilgarriff 1998] are available. In SENSEVAL, Kilgarriff and Rosenzweig [2000] report a performance of 75% for a fine-grained word sense disambiguation task for English. There does not exist any large-scale sense-tagged corpus for Chinese that is comparable to English.

Although the ability to tag a large corpus by hand is useful, it is also time consuming. Hence, computer-aided tools to sense-tag a Chinese corpus were investigated. Lua [1997] proposed an inductive unsupervised semantic tagger for Chinese words. He adopted the middle categories of Cilin and five additional semantic tags as a tagging set. A total of 100 semantic tags were used to tag a corpus of 348,393 words, and, finally, 2,000 semantic tags were checked manually. Accuracy was about 91%. Because the granularity of the tagging set is an important issue that affects tagging accuracy, we cannot conclude that the performance of a Chinese sense tagger is better than that of an English sense tagger.

This article focuses on two topics, i.e., Chinese-English WordNet and the translingual applications. It is organized as follows: Section 2 sketches an overview, showing how to integrate five linguistic resources to build a Chinese-English WordNet. The basic components, including a sense tagger, a translingual context mapper, and a Chinese-English WordNet constructor, are presented in Sections 3, 4, and 5, respectively. Section 6 demonstrates how to use the bilingual WordNet in a translingual application, i.e., Chinese-English information retrieval. Section 7 concludes.

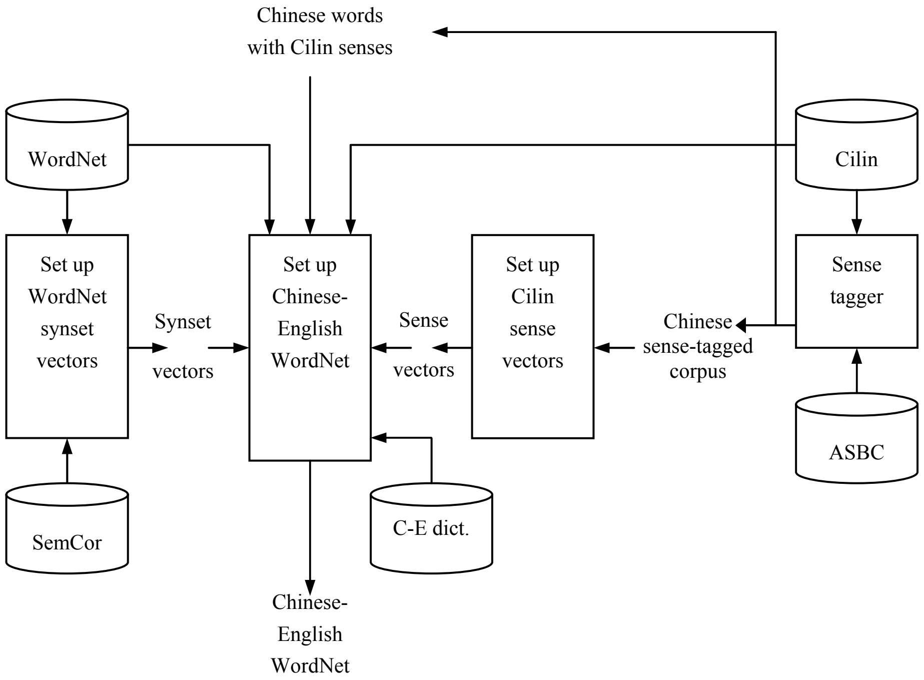

Fig. 1. Overview for building a Chinese-English WordNet.

2. OVERVIEW

The basic idea is to construct a Chinese word net from an available Chinese linguistic resource (Cilin) and following the structure of WordNet. A Chinese-English WordNet is set up at the same time. Figure 1 shows the architecture of our design. It is composed of four major components, and integrates five major linguistic resources. The linguistic resources are introduced below.

(1) ***Cilin*** gathers 65,464 Chinese word entries.[1] Cilin senses are decomposed to 12 large categories, 94 middle categories, 1,428 small categories, and 3,925 word clusters. Some Cilin senses follow:

A people (人)
Aa a collective name (泛稱)
01 human being (人), the people (人民), everybody (眾人)
02 I (我), we (我們)
03 you (你), you (你們)
04 he/she (他), they (他們)
05 myself (自己), others (別人), someone (某人)
06 who (誰)

[1] A word with *n* senses corresponds to *n* entries, so that the number of word types is less than 65,464. In reality, there are 53,644 words, including 45,586 unambiguous words and 8,058 ambiguous words.

Ab people of all ages and both sexes (男女老少)
 01 man (男人), woman, (女人), men and women (男女)
 02 old person (老人), adult (成年人), old and young (老小)
 03 teenager (青少年)
 04 infant (嬰兒), child (兒童)
Ac posture (體態)
 01 tall person (高個兒), dwarf (矮子)
 02 fat person (胖子), thin person (瘦子)
 03 beautiful woman (美女), handsome man (美男子)
…
B body (物)
Ba generally called (統稱)
 01 body (物), object (物體)
 02 being (生物)
 03 article (物品), object (物件)
 04 goods (貨物), products (產品)
 05 appliances (器具), facilities (設備)
 06 goods and materials (物資), daily information (生活資料)
 07 gifts (禮品), dowry (嫁妝), sacrificial offering (祭品)
 08 treasure (寶物), waste material (廢物)
 09 burden (擔子), pack (馱子), bag (包裹)
 10 it (它), what (什麼)
…
L respectful words (敬語)
 01 good morning (早安), good night (晚安), good bye (再見)
 02 excuse me (請問), advise (指教), ask advice (領教)
 03 be obliged to (承蒙), ask somebody to come to (賞光)
 04 disturb (打擾), excuse me (勞駕), sorry (抱歉)
 05 been looking forward to seeing you (久仰)
 haven't seen you for ages (久違), neglect (怠慢)
 to do as you please (請便)
 06 congratulations (恭喜), good fortune (托福)
 thank you (謝謝), you are welcome (不謝)

Symbols A, B, ..., L denote large categories; symbols Aa, Ab, Ac, ..., Ba, ... denote middle categories; symbols Aa01, ..., Aa06, Ab01, ..., Ab04, Ac01, ..., Ac03 ... denote small categories. Besides semantic classification, Cilin also lists words on the basis of lexical categories, e.g., A-D (nouns), E (adjectives), F-J (verbs), K (auxiliary), and L (others). In this article , 1,428 small categories form a Cilin sense tagging set.

(2) ***ASBC*** [Huang and Chen 1995]. Academic Sinica Balanced Corpus (abbreviated ASBC corpus) is a POS-tagged Chinese balanced corpus. The major topics include philosophy (10%), science (10%), society (35%), art (5%), life (20%), and literature (20%). This corpus is composed of five million words.

(3) ***WordNet*** is an English knowledge base. Words that have the same meanings form a synset in WordNet. Special relational pointers like hypernym, hyponym, antonym, and so on, link synsets to a network. WordNet is also classified into four structures (nouns, verbs, adjectives, and adverbs) based on POS. There are 99,685 synsets

.

(4) ***SemCor*** is a synset-tagged corpus created at Princeton University. The material is selected from POS-tagged Brown Corpus. SemCor includes three concordances, i.e., brown1 concordance (103 files); brown2 concordance (83 files); and brownv concordance (166 files). In total, SemCor is composed of 198,597 words.

.

(5) ***Chinese-English dictionary***. The bilingual dictionary is integrated from four sources, including the LDC Chinese-English dictionary; Denisowski's CEDICT[2], BDC Chinese-English dictionary v2.2,[3] and a dictionary used for query translation in the MTIR project [Bian and Chen 2000]. The dictionary has a collection of 200,037 words, where a word may have more than one translation.

.

In Figure 1, a sense tagger labels Cilin senses to words in the ASBC corpus and generates a sense-tagged ASBC corpus. Sense/synset vectors for Chinese/English are trained, respectively, from sense-tagged ASBC/SemCor. The Chinese-English dictionary acts as a bridge linking Chinese words to their proper positions in WordNet and so, finally, a Chinese-English WordNet is built. It supports translingual applications like cross-lingual information retrieval [Chen et al. 2000]; multilingual news summarization [Chen and Lin 2000b]; multilingual topic detection and tracking [Chen and Ku 2002], and so on.

3. A SENSE TAGGER

3.1 Tagging Unambiguous Words

A tagging task assigns sense tags to words in a sentence. The difficulty of the task for a given language can be measured in terms of the average number of word senses. The higher the degree of polysemy, the more challenging the sense tagging. The paper by Chen and Lin [2000a] studies this problem by using the ASBC corpus and the Cilin sense tags. A total of 28,321 word types[4] appear both in Cilin and in the ASBC corpus, Of these, 5,922 words are polysemous, i.e., they have more than one sense. The statistics show that 93.77% of polysemous word types have 2-4 senses, but they only occupy 58.52% of word tokens[5] in the ASBC corpus. In comparison, highly ambiguous words (i.e., number of senses > 4) tend to be highly frequent words; that is, 6.23% of polysemous word types occupy 41.48% of tokens. This shows that semantic tagging is a challenging problem in Mandarin Chinese. Tagging accuracy depends on several issues [Manning and Schutze 1999], e.g., the amount of training data, the granularity of the tagging set, the occurrences of unknown words, and so on. Because a sense-tagged Chinese corpus was unavailable, the Chinese sense tagger was developed incrementally.

[2] The dictionary is available at http://www.mandarintools.com/cedict.html

[3] The BDC dictionary was developed by the Behavior Design Corporation (http://www.bdc.com.tw).

[4] A word type corresponds to a dictionary entry.

[5] A word token is an occurrence of a type in a corpus.

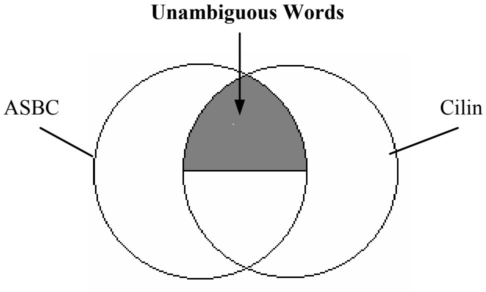

Fig. 2. Tagging unambiguous words.

The small categories were adopted in sense tagging. We postulate that the sense definition for each word in Cilin is complete. That is, a word that has only one sense in Cilin is called an *unambiguous* word or a *monosemous* word. If POS information is also considered, a word may be unambiguous under a specific POS. Since we did not have a sense-tagged corpus for training, we tried to acquire the context for each sense tag beginning with the unambiguous words.

The ASBC corpus is the target we studied. At the first stage, only those words that are unambiguous in Cilin, and also appear in the ASBC corpus were tagged. Figure 2 shows this case.

An unambiguous word (and hence its sense tag) is characterized by the words surrounding it. The window size was set to 6 and the stop words were removed. A list of stop words was trained from the ASBC corpus. Words with POSes such as Neu (數詞, numeral); DE (的, 之, 得, 地, de); SHI (是, be); FW (foreign word); C (conjunction); T (particle); and I (interjection/exclamation) are regarded as stop words. A sense tag *Ctag* is used in terms of a vector (w_1, w_2, ..., w_n), where n is the vocabulary size and w_i is a weight of word *cw*. The weight can be determined in the following two ways:

(1) ***MI metric*** [Church et al. 1989]:

$$MI(Ctag, cw) = \log_2 \frac{P(Ctag, cw)}{P(Ctag)P(cw)} \approx \log_2 \frac{f(Ctag, cw)}{f(Ctag)f(cw)} \times N \tag{1}$$

where $P(Ctag)$ is the probability of *Ctag*;
$P(cw)$ is the probability of *cw*;
$P(Ctag, cw)$ is the co-occurrence probability of *Ctag* and *cw*;
$f(Ctag)$ is the frequency of *Ctag*;
$f(cw)$ is the frequency of *cw*;
$f(Ctag, cw)$ is the co-occurrence frequency of *Ctag* and *cw*; and
N is total number of words in the corpus.

(2) ***EM metric*** [Ballesteros and Croft 1998]:

$$em(Ctag, cw) = \max\left(\frac{f(Ctag, cw) - En(Ctag, cw)}{f(Ctag) + f(cw)}, 0 \right) \tag{2}$$

$$En(Ctag, cw) = \frac{f(Ctag)f(cw)}{N} \tag{3}$$

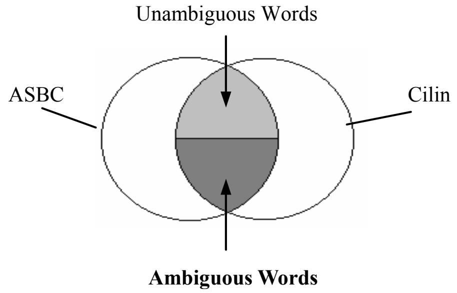

Fig. 3. Tagging ambiguous words.

Take sense tag Fa01 as an example. Sense tag Fa01 means 打 (hit), 拍 (hit lightly), 撫摩 (stroke), 搔 (scratch) and 摸 (touch). The sense vectors are trained from the sense-tagged ASBC corpus after unambiguous words are tagged. The sense vector of Fa01 is < 踢 (kick, play), 叫好 (applaud, cheer, shout bravo), 罵 (scold, abuse), 樂團(band), 安打 (bingle, safety hit), 自信心(conviction, belief, confidence), ...>.

3.2 Tagging Ambiguous Words

At the second stage we dealt with those words that have more than one sense in Cilin. Figure 3 shows the words we made use of.

Because an ambiguous word has more than one sense tag, we had to select the best one. The information trained at the first stage is used to select the best sense tag. At first, we employed a method similar to the one specified in Section 3.1 to identify the context vector of an ambiguous word. Recall that a vector corresponds to a sense tag. The context vector is then compared to the sense vector of each candidate sense tag. The sense tag with the highest similarity score is chosen. A cosine formula shown below measures the similarity between a sense vector *w* and a context vector *v*.

$$\cos\left(w, v\right) = \frac{w \cdot v}{|w|\,|v|} \tag{4}$$

We retrained the sense vector for each sense tag after the unambiguous words were resolved.

3.3 Tagging Unknown Words

Words that appear in the ASBC corpus, but are not collected in Cilin are called *unknown words*. All 1,428 sense tags are possible candidates. Intuitively, the algorithm in Section 3.2 can be applied directly to selecting a sense tag from the 1,428 candidates. However, the candidate set is very large. Here we adopted outside evidence from the mappings among WordNet synsets and Cilin sense tags to narrow the candidate set. Figure 4 summarizes the flow of our algorithm, illustrated as follows:

(1) Find all the English translations of an unknown Chinese word by looking up a Chinese-English dictionary.

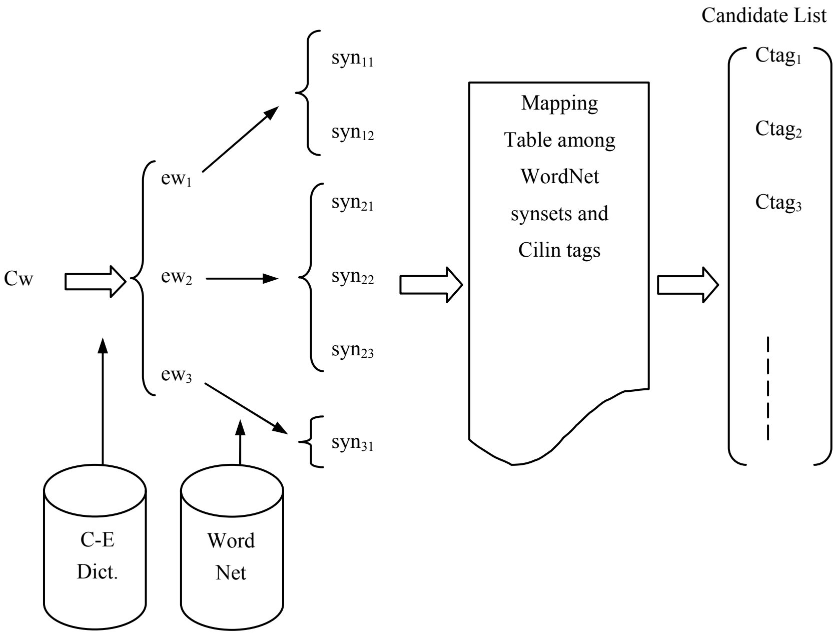

Fig. 4. Sense tagging flow.

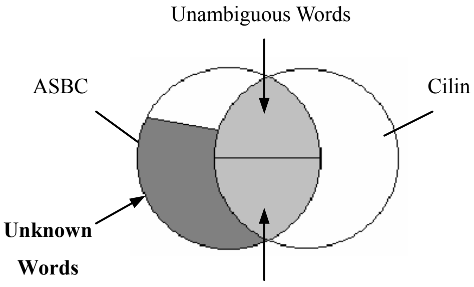

Fig. 5. Tagging unknown words.

(2) Find all the synsets of the English translations by looking up WordNet. We did not resolve translation ambiguity and target polysemy at these two steps, thus the retrieved synsets may cover more senses than that of the original Chinese word.

(3) Transform the synsets back to Cilin sense tags by looking up a mapping table [Chen and Lin 2000a].

(4) Select a sense tag from the candidates proposed in step (3) by using the method specified in Section 3.2.

Table I. Tagging Ambiguous Words: Performance

Ambiguity Word Tokens	Low	Middle	High	Summary
Total Tokens	6,601	3,511	989	11,101
Correct Rate (MI)	62.60%	31.36%	27.00%	49.55%
Correct Rate (EM)	63.98%	37.99%	31.34%	52.85%

Table II. Tagging Using the First-n and EM: Performance

Ambiguity First-n	Low	Middle	High	Middle and High
1	**63.98%**	37.99%	31.34%	36.53%
2		**60.92%**	53.99%	59.40%
3		71.35%	**67.95%**	70.60%

Table III. Tagging Unknown Words: Performance

Categories	#Tokens		Baseline	M	P	M (POS)
All	1633	Correct	20	443	438	561
		Precision	1.22%	27.13%	26.82%	34.35%
N	858	Correct	11	255	255	320
		Precision	1.28%	29.72%	29.72%	37.30%
V	619	Correct	5	144	137	167
		Precision	0.81%	23.26%	22.13%	26.98%
A	58	Correct	0	5	5	28
		Precision	0	8.62%	8.62%	48.28%
F	4	Correct	1	1	1	4
		Precision	25.00%	25.00%	25.00%	100.00%
K	94	Correct	3	38	40	42
		Precision	3.19%	40.43%	42.55	44.68%

Figure 5 shows the unknown words we dealt with at this stage. Words not in our Chinese-English dictionary were not considered, so that only parts of unknown words were resolved. In other words, words without sense tags remain.

3.4 Sense-Tagging Experiments

We sampled documents from different categories of the ASBC corpus, including philosophy (10%), science (10%), society (35%), art (5%), life (20%) and literature (20%). There were 35,921 words in the test corpus. Research associates tagged this corpus manually. They first marked the ambiguous words by checking Cilin, and then tagged the unknown words. A list of candidate words was developed by checking against the mapping table. Because the mapping table may have contained errors, the annotators assigned a "none" tag when they could not choose a solution from the proposed candidates. After manual tagging, 17.48% of the unknown words were tagged with "none."

Table I shows the performance for tagging ambiguous words. A total of 11,101 words were tagged. When the MI metric was used, the performance for tagging low (2-4 senses), middle (5-8 senses), and highly ambiguous words (>8 senses) were 62.60%, 31.36%, and 27.00%, respectively. When the EM metric was used, performance improved, in particular for the classes with middle- and high- ambiguity. The overall correct rate increased from 49.55% to 52.85%.

In the previous experiments, only one sense was reported for each word. When we reported more than one sense for the middle and highly ambiguous words, performance improved. Table II shows that the first 2 and 3 candidates were selected. From the diagonal of this table, the performance for tagging low ambiguity (2-4), middle ambiguity (5-8), and high ambiguity (>8) is similar (i.e., 63.98%, 60.92%, and 67.95%) when, respectively, 1 candidate, 2 candidates, and 3 candidates were proposed. In this case, 7,034 of 11,101 words were tagged correctly; i.e., the correct rate is 63.36%.

There were 1,979 unknown words in our test corpus. A total of 1,633 words were tagged manually. Table III shows the performance. Experiment M used the training results in Section 3.1 (i.e., unambiguous words), while experiment P utilized the training results in Section 3.2 (i.e., unambiguous and ambiguous words). In the baseline model, all 1,428 Cilin tags are the candidates of unknown words, and the performance is worse. On the average, the precision is 1.22%. Both M and P were better than the baseline model. This table also lists the performance of each category. It meets our expectation, i.e., tagging verbs is harder than tagging the other categories. Next we used POS to improve performance. POS narrowed the number of candidates, so that the overall correct rate was enhanced from 27.13% to 34.35%.

Finally, we considered the overall performance of tagging our sample data. Recall that there were 35,921 words in the test corpus. Except for the stop words that were not tagged by the sense tagger, there remained 13,586 unambiguous words, 11,101 ambiguous words, and 1,633 unknown words for tagging. From Tables I and III, we know that 5,867 ambiguous words and 561 unknown words were tagged correctly. The sense tagger achieved a performance of 76.04%.

4. A TRANSLINGUAL CONTEXT MAPPER

A semantic tag (e.g., a Cilin sense tag and a WordNet synset) is characterized by the words surrounding it in a sense-tagged corpus (e.g., a sense-tagged ASBC corpus and SemCor). The window size is set to 6 and stop words are removed. The weight of a vector term is determined by mutual information.

Using the above schemes, we can derive two sets of semantic vectors, i.e., Cilin sense vectors and WordNet synset vectors. The former are Chinese vectors and the latter English vectors. To compare the similarities of these two types of semantic vectors during construction of a word net, we translated Chinese into English. Because the sense vector was very large, we only translated the top 200 Chinese words for each sense vector in order to reduce complexity. For a Cilin sense vector, we found all the English translation equivalents for the top 200 Chinese words in the sense vector by checking a Chinese-English dictionary. We selected a set of English translations that are semantically coherent by using the word co-occurrence model. The word co-occurrence model was used as follows.

Table IV. The English Version of Sense Vector for Sense Tag Fa01

Fa01	打 (hit), 拍 (hit lightly), 撫摩 (stroke), 搔 (scratch), 摸 (touch)
Sense vector of Fa01	踢 (kick, play), 叫好 (applaud, cheer, shout bravo), 罵 (scold, abuse), 樂團 (band), 安打 (bingle, safety hit) , 自信心 (conviction, belief, confidence), ...
English version of sense vector of Fa01	play, applaud, abuse, band, bingle, confidence, ...

(1) From the 200 Chinese words, those words that do not have translation ambiguity, i.e., they have only one English translation, are identified, and the corresponding English translations are regarded as *a seed* to find the other English translations.

(2) If the *seed* is not large enough, we consider the first *k* (*k* ≤ 10) Chinese words that have two English translations. A set of English translations was selected from 2^k combinations according to the following formula and added to the *seed*.

$$\arg\max_{ew_1, ew_2, \ldots, ew_k} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} MI\,(ew_i, ew_j) \tag{5}$$

(3) We sorted the remaining Chinese words with respect to the number of English translation equivalents. The Chinese words were considered in sequence based on their degree of ambiguity. That is, the less ambiguous Chinese words were examined first. We assumed there were (*i*-1) English translations in the *seed* up to now. We let the English translations for the next Chinese word cw_i be ew_{i1}, ew_{i2}, …, and ew_{in}. We selected an English translation equivalent based on the following formula and inserted it into the *seed*.

$$\arg\max_{j} \sum_{k=1}^{i-1} MI\,(ew_{ij}, ew_k) \tag{6}$$

This formula tries to find an English translation equivalent that has strong relationships with the words in the current *seed*.

An English version for a Cilin sense vector was generated with the algorithm above. In the following, we refer to the English version of Cilin sense vectors when computing similarity. Table IV shows the English version of a sense vector for sense tag Fa01.

5. A CHINESE-ENGLISH WORDNET CONSTRUCTOR

At first, the constructor puts unambiguous words (specified in Section 3.1) into WordNet by looking in a Chinese-English dictionary. Although these words do not have translation ambiguity, the corresponding English translation may have target polysemy problem. In other words, the English translation may cover irrelevant senses besides the correct one. The following algorithm finds the closest synset to the Chinese sense tag.

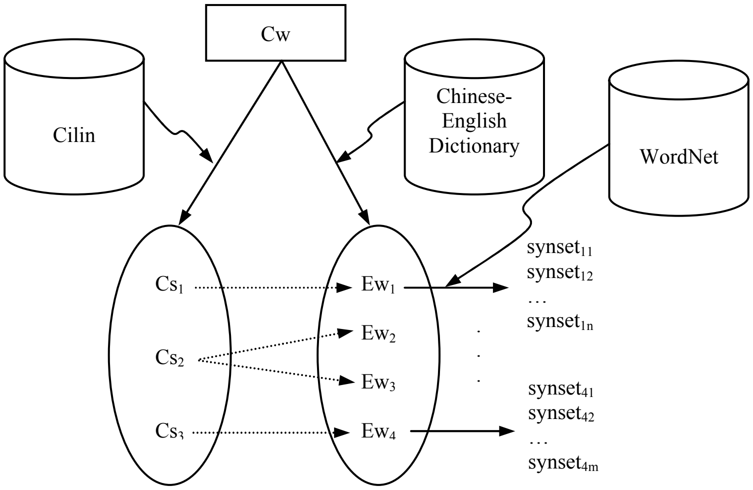

Fig. 6. Chinese senses and English translation.

(1) If the English translation corresponds to only one synset, this synset is the solution.
(2) If the English translation corresponds to more than one synset, POS is considered:

(a) If the Chinese sense tag belongs to one of categories A-D in Cilin (i.e., a noun sense), and there is only one noun synset, then the synset is adopted. Otherwise, we compare the English version of the Chinese sense vector with vectors of the synsets, and select the closest synset.
(b) If the Chinese sense tag belongs to one of categories F-J in Cilin (i.e., a verb sense), we try to find a verb synset similarly to (a). If this fails, we try noun and adjective synsets instead.
(c) If the Chinese sense tag belongs to category E in Cilin (i.e., an adjective sense), we try adjective, adverb, noun, and verb synsets in sequence.
(d) If the Chinese sense tag belongs to category K in Cilin (i.e., an adverb sense), only adverb synsets are considered.

We next considered the ambiguous words. That is, words that have more than one Cilin sense. All the English translations were found by looking in the Chinese-English dictionary. WordNet search collected the synset candidates for the translations, and some synsets were selected. Here problems of translation ambiguity and target polysemy had to be faced. In other words, not all English translations covered all the Cilin senses. Figure 6 shows an example. A Chinese word Cw has three senses – say, Cs_1, Cs_2 and Cs_3, and four English translations – say, Ew_1, Ew_2, Ew_3 and Ew_4. Assume Ew_1 covers sense Cs_1, Ew_4 covers sense Cs_3, and Ew_2 and Ew_3 have a common sense Cs_2. After looking inWordNet, the senses expand drastically.

Because our Chinese-English dictionary does not distinguish Chinese senses and English translations, their relationships in Figure 6 are marked by dotted lines. A problem also appears when translation ambiguity and target polysemy are integrated.

Here we took a conservative approach. At first, the synsets, e.g., synset$_{11}$, synset$_{12}$, …, synset$_{1n}$, …, synset$_{41}$, synset$_{42}$, …, synset$_{4m}$ in Figure 6, are collected and partitioned by POS. We then selected a synset from the synset candidates for each sense of the Chinese word by using a similar approach to mapping unambiguous words.

During mapping, some Chinese words may not be mapped to WordNet synsets because they could not be found in the Chinese-English dictionary, or WordNet could not collect the English translations even when dictionary look-up was successful. We employed the semantic relationship in Cilin to deal with such problem words. Recall that Cilin is divided into four layers, consisting of large categories, middle categories, small categories, and word clusters, from rough to fine grained. We collected the unambiguous words in the same word cluster as the problem word and mapped the problem word to the synsets that most of the unambiguous words map to.

6. AN APPLICATION TO CHINESE-ENGLISH INFORMATION RETRIEVAL

6.1 Test Materials

We next experimented with Chinese-English information retrieval based on the Chinese-English WordNet. We employed the TREC-6 text collection, TREC topics 301-350 [Harman 1997], and the Smart information retrieval system [Salton and Buckley 1998]. Both the document weight and term weight were determined by *ntc* mode in the Smart system. The text collection contained 556,077 English documents, and was about 2.2G bytes. Because the goal was to evaluate the performance of Chinese-English information retrieval on different models, the 50 English queries were translated into Chinese by a person who was independent of the conductors of the query experiments.

Topic 301 is used as an example in the following: The original English version and the human-translated Chinese version are shown. A TREC topic is composed of several fields. Tags <num>, <title>, <des>, and <narr> denote topic number, title, description, and narrative fields. The narrative provides a complete description of document relevance for the assessors. In our experiments, only the title and description fields were used to generate queries.

<top>
<num> Number: 301
<E-title> International Organized Crime
<C-title> 國際組織犯罪
<E-desc> Description:
Identify organizations that participate in international criminal activity, the activity, and if possible, collaborating organizations and the countries involved.
<C-desc> Description:
辨識參與國際犯罪活動的組織。可能的話，找出活動，共事的組織與牽涉的國家。
<E-narr> Narrative:
A relevant document must as a minimum identify the organization and the type of illegal activity (e.g., Columbian cartel exporting cocaine). Vague references to the international drug trade without identification of the organization(s) involved would not be relevant.
<C-narr> Narrative:
最少必須有一份相關的文件來辨識其組織和非法活動的型態（如：哥倫比亞聯合出口古柯鹼）。提到國際販毒而沒有指出牽涉的組織的文件是不相關的。</top>

Table V. MI Values of Any Two Synsets in the Example Query

		國際		組織		犯罪								
		syn$_{11}$	syn$_{12}$	syn$_{21}$	syn$_{22}$	syn$_{31}$	syn$_{32}$	syn$_{33}$	syn$_{34}$	syn$_{35}$	syn$_{36}$	syn$_{37}$	syn$_{38}$	syn$_{39}$
國際	**syn$_{11}$**			1.517	**4.394**	1.233	0.444				1.583		1.451	-0.076
	syn$_{12}$													
組織	syn$_{21}$	1.517				-0.061	0.028				-0.536		0.601	1.468
	syn$_{22}$	4.394				**3.899**					0.417		0.010	
犯罪	**syn$_{31}$**	1.233		-0.061	**3.899**									
	syn$_{32}$	0.444		0.028										
	syn$_{33}$													
	syn$_{34}$													
	syn$_{35}$													
	syn$_{36}$	1.583		-0.536	0.417									
	syn$_{37}$													
	syn$_{38}$	1.451		0.601	0.010									
	syn$_{39}$	-0.076		1.486										

6.2 Query Translation

A Chinese query is segmented by a word recognition system and then tagged by a POS tagger [Chen et al.1998]. After removing stop words, we looked up the Chinese-English WordNet for the remaining Chinese words. A set of synsets was retrieved for each Chinese query term. We computed the mutual information for the sets of synsets and selected a synset for each Chinese query term. The result forms an English query and was submitted to the Smart system. The mutual information of two synsets is defined as follows. Let synset$_1$ and synset$_2$ be synsets for two query terms. Assume synset$_1$ and synset$_2$ are composed of *m* and *n* English words, respectively.

$$MI(synset_1, synset_2) = \sum_{i=1}^{m}\sum_{j=1}^{n} MI(t_{1i}, t_{2j})/(m \times n) \qquad (7)$$

The MI values of any two English words were trained from the TREC-6 corpus.

Take "國際組織犯罪" (international organized crime) as an example. Table V shows the MI values of two synsets. There are 2, 2, and 9 synsets corresponding to "國際" (international), "組織" (organized), and "犯罪" (crime), respectively. These synsets are as follows (words enclosed by parentheses are the English translations of the Chinese words):

(1) 國際

syn11	(international) international
syn12	(internationality) internationality internationalism

(2) 組織

syn21	(group) group grouping
syn22	(community) community

(3) 犯罪

syn31	(crime) crime
syn32	(offence) misdemeanor misdemeanour infraction offence offense violation infringement
syn33	(criminality) criminalism criminality criminalness
syn34	(misdeed) misbehavior misbehaviour misdeed
syn35	(misdoing) blunder blooper bungle foul-up flub botch boner boo-boo misdoing
syn36	(perpetration) perpetration commission
syn37	(transgression) evildoing transgression
syn38	(criminal) criminal felon crook outlaw malefactor
syn39	(illegal) illegal

Consider the term "國際" (international) first. The largest MI, which is between synset$_{11}$ and synset$_{22}$, is 4.394, so that synset$_{11}$ is selected for query term "國際". We then selected the best synsets for "組織" (organized) and "犯罪"(crime). Synset$_{22}$ and synset$_{31}$, respectively, were selected similarly.

Figure 7 shows seven alternatives in query construction. Their differences are the weighting scheme and the resource used. The four variations (i.e., AS, ASW, AST, and ASWTW) shown on the left-hand side employ Chinese-English WordNet. The three variations (i.e., ALL, CO, U1WCO) shown on the right-hand side utilize a Chinese-English dictionary. In the AS model (meaning *A*ll words in *S*ynset), if a query term cannot be found in the word net, then it is not translated and is not put into the final query either. Also, all the English words in a synset have the same weight. ASW (meaning *A*ll words in *S*ynset with *W*eight) is similar to AS, except that the English translations of query terms are given higher weights than the other English words in the synsets. The adjusted weights of the English translations of query terms are three times their original weights. AST (meaning *A*ll words in *S*ynset plus *T*ranslation from dictionary) and ASWTW also consider query terms that are not found in the Chinese-English WordNet. If a query term is absent from the Chinese-English WordNet, we find its English translation by checking a Chinese-English dictionary and then putting all these translations into a final query. Thus, some query terms are the results of synset co-occurrence and the others are the results of dictionary look-up. All the query terms are of equal weight in AST. In ASWTW, English translations of query terms from the Chinese-English WordNet and the Chinese-English dictionary are assigned larger weights. Similar to model ASW, the adjusted weights of the English translations are three times their original weight.

The following shows the queries constructed using AS and AST. Topic 301 is taken as an example. The query terms are composed of words selected from title and description fields. The words in boldface are from the bilingual dictionary.

(1) AS：	international		國際 (international)
	community		組織 (organized)
	crime		犯罪 (crime)
	intervention	intercession	參與 (participate in)
	international		國際 (international)

	crime				犯罪 (crime)
	community				組織 (organization)
	conversation				話 (an expletive)
	active				活動 (activity)
	travel_rapidly	speed	hurry	zip	共事[6] (work together)
	community				組織 (organization)
	dredge drag				牽 (involve)
	involve	affect	regard		涉 (involve)
	national	subject			國家 (country)
(2) AST：	international				國際 (international)
	community				組織 (organized)
	crime				犯罪 (crime)
	recognize	**recognition**			**辨識 (identify)**
	intervention	intercession			參與 (participate in)
	international				國際 (international)
	crime				犯罪 (crime)
	active astir	**volitant**			**活動的 (active)**
	community				組織 (organization)
	possible	**feasible**	**imaginable**	**probable**	**可能的 (if possible)**
	conversation				話 (an expletive)
	find_up	**spot**			**找出 (find out)**
	active				活動 (activity)
	travel_rapidly	speed	hurry	zip	共事 (work together)
	community				組織 (organization)
	dredge drag				牽 (involve)
	involve	affect	regard		涉 (involve)
	national	subject			國家 (country)

There are 1,017 words in the title and description fields of the 50 TREC topics. After removing stop words, there remain 703 words. A total of 484 words can be found in synsets and 219 words are from the Chinese-English dictionary.

On the right-hand side of Figure 7, three models, i.e., ALL, CO, and U1WCO [Chen et al. 1999], are shown for comparison with our methods. In the ALL model, all the English translation equivalents for the Chinese query terms are selected. The CO model considers the content around the English translations in deciding the best target equivalent. The translation of a query term can be disambiguated using the co-occurrence of the translation equivalents of the given term and other terms. We employ mutual information to measure the strength of word co-occurrence. The selection strategy is the same as that for selecting synsets.

[6] The Chinese words "共事" (work together) and "趕工" (hurry) belong to the same sense category Hj11. Because its English translation "work together" cannot be found in WordNet, we assign the Synset for "趕工" (hurry) to "共事" (work together).

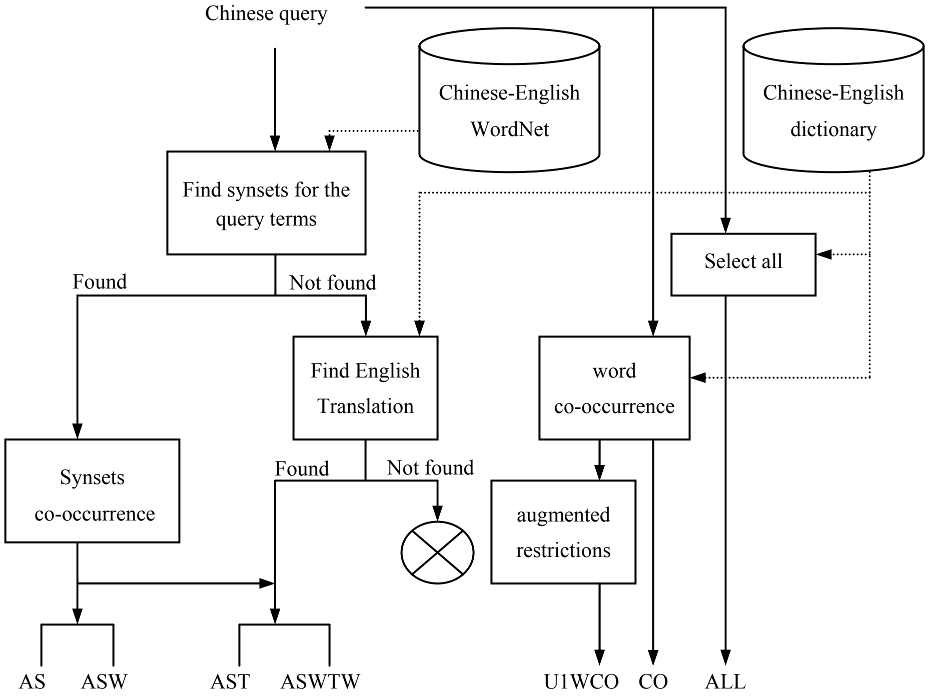

Fig. 7. Query construction.

The U1WCO model augments translation restrictions to query translation. The frequently accompanying nouns and verbs are collected for each word in the ASBC corpus. Those words that co-occur with a given word within a window are selected. The word association strength of a word and its accompanying words are measured based on mutual information. Assume that a Chinese query Q is composed of *n* words C_1, C_2, ..., C_n. Each word C_i is augmented with a sequence of Chinese words trained in the above way. The accompanying words may be translated into more than one English word. An augmented translation restriction may add erroneous patterns when a word in a restriction has more than one sense. In this model, the terms without ambiguity, i.e., those Chinese and English words that have a one-to-one correspondent in an Chinese-English bilingual dictionary, are added. The translations of the accompanying words constitute a pseudo-translation context. We applied the CO model again to disambiguate the pseudo-translation contexts. Thus, at most one restriction word was added for each query term. The translations of query terms and augmented translation restrictions were assigned different weights, which were determined by the following formulas.

$$\text{weight}(\text{E}_\text{i}) = \frac{1}{n+1} \tag{8}$$

$$\text{weight}(\text{EW}_\text{j}) = \frac{1}{(n+1)*m} \tag{9}$$

Table VI. Experimental Results

	baseline	ALL	CO	U1WCO	AS	ASW	AST	ASWTW
11-pt Avg. Precision	0.1459	0.0652	0.0831	0.0918	0.0159	0.0194	0.0970	0.1010
%baseline		44.69%	56.96%	62.92%	10.90%	13.30%	66.48%	69.23%
%change			+27.46%	+40.79%	-75.61%	-70.24%	+48.76%	+54.91%

where E_i is the translation of a query term C_i;
EW_j is an augmented translation restriction term;
n is number of words in Q; and
m is number of augmented translation restriction terms.

6.3 Experimental Results

Table VI shows the overall performance of the models for 50 topics. Eleven-point average precision on the top 1,000 retrieved documents is adopted to measure the performance of all the experiments. The monolingual information retrieval, i.e., the original English queries to a English text collection, is regarded as a baseline model. The performance is 0.1459 under the specified environment. The results, i.e., ALL, CO, and U1WCO, from Chen et al. [1999] are listed for comparison. The ALL model, in which all the translation equivalents are passed without disambiguation, has 0.0652 average precision. About 44.69% of the performance of the monolingual information retrieval is achieved. A comparison of the performance with monolingual information retrieval is shown in the third row. When the co-occurrence model is employed to resolve translation ambiguity, 0.0831 average precision (i.e., 56.96% of monolingual information retrieval) is reported. Besides co-occurrence to treat translation ambiguity, U1WCO employs augmented restriction to deal with the target polysemy problem. It gains 62.92% of monolingual information retrieval.

We now discuss the performance of our experimental results. The AS model achieves only 10.90% of monolingual IR model. When reweighing the query terms, the performance increased a little in ASW model. The errors resulted from unknown words in queries. Recall that the terms in Chinese-English WordNet come from Cilin and WordNet. Besides Cilin's coverage problem, WordNet may not gather the English words in the Chinese-English dictionary, particularly proper names. Thus we included English translations from the bilingual dictionary when checking Chinese-English WordNet failed. AST, which demonstrates a large improvement, achieves 0.0970 average precision. It is better than the CO model, which resolves translation ambiguity only. When reweighing query terms, ASWTW has 0.1010 average precision, i.e., 69.23% of the monolingual IR model. Compared to U1WCO, which resolves translation ambiguity and target polysemy together, ASWTW gains a 10.02% increase.

Table VI also lists the performance change (%) of model *A* relative to model *B*, which is defined by (11 point average precision of model *A* – 11 point average precision of model *B*) / 11 point average precision of model *B*. The fourth row of Table VI summarizes the performance changes relative to the ALL model. ASWTW achieves 54.91% change.

7. CONCLUDING REMARKS

A WordNet-like linguistic resource is useful but difficult to construct. This article proposes a method to integrate five linguistic resources, including English/Chinese sense-tagged corpora, English/Chinese thesauruses, and a bilingual dictionary. By following the structures of WordNet, we derived a Chinese-English WordNet. Experiments on Chinese-English information retrieval were developed to evaluate the applicability of the Chinese-English WordNet. The best model achieves 0.1010 average precision, 69.23% of monolingual information retrieval. It also gains 10.02% increase relative to a model that resolves translation ambiguity and target polysemy together.

There is still much room to improve the Chinese-English WordNet; coverage of the bilingual WordNet is not good enough. Using the approach in this article, 52,302 of 65,464 word entries in Cilin were entered in WordNet. This is because English translations of a Chinese word may not be found in the Chinese-English dictionary, and WordNet may not have collected the English translations even when dictionary look-up is successful. In particular, WordNet contains only a few proper names, and the bilingual dictionary, which acts as a bridge linking the various resources, is not a sense-based dictionary. That is, we do not know the exact sense of each dictionary entry. Thus the method we proposed serves as a computer-aided tool to construct a knowledge base in a more effective way. Human examination is necessary to guarantee the quality of the Chinese-English WordNet and make it fit for public use.

REFERENCES

ARTALE, A., MAGNINII, B., AND STRAPPARAVA, C. 1997. Lexical discrimination with the Italian version of WordNet. In *Proceedings of ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (Madrid, Spain, July 1997), P. VOSSEN ET AL. Eds. Association for Computational Linguistics.

ATKINS, S. 1993. Tools for computer-aided lexicography: the Hector project. *Acta Linguistica Hungarica 41*, 5-72.

BALLESTEROS, L. AND CROFT, W. B. 1998. Resolving ambiguity for cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia, Aug. 1998), W. B. CROFT ET AL. Eds. ACM Press, New York, NY, 64-71.

BIAN, G. W. AND CHEN, H. H. 2000. Cross language information access to multilingual collections on the Internet. *J. Am. Soc. Inf. Sci. 51, 3*, 281-296.

CHEN, H. H., BIAN, G. W., AND LIN, W. C. 1999. Resolving translation ambiguity and target polysemy in cross-language information retrieval. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics* (June 1999). Association for Computational Linguistics, 215-222.

CHEN, H. H., DING, Y. W., AND TSAI, S. C. 1998. Named entity extraction for information retrieval. *Comput. Process. Oriental Lang. Special Issue on Information Retrieval of Oriental Languages 12, 1,* 75-85.

CHEN, H. H. AND KU, L. W. 2002. An NLP&IR approach to topic detection. In *Topic Detection and Tracking: Event-Based Information Organization*, J. ALLAN ET AL. Eds. Kluwer, Boston, MA, 243-264.

CHEN, H. H. AND LIN, C. C. 2000a. Sense-tagging Chinese corpus. In *Proceedings of ACL Workshop on Chinese Language Processing* (Hong Kong, Oct. 2000), M. PALMER ET AL. Eds. Association for Computational Linguistics, 7-14.

CHEN, H H. AND LIN, C. J. 2000b. A multilingual news summarizer. In *Proceedings of 18th International Conference on Computational Linguistics* (Saarbrücken, Germany, Aug. 2000), 159-165.

CHEN, H .H., LIN, C. C., AND LIN, W. C. 2000. Construction of a Chinese-English WordNet and its application to CLIR. In *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages* (Hong Kong, Sept.-Oct. 2000). ACM Press, New York, NY, 189-196.

CHURCH, K., GALE, W., HANKS, P., AND HINDLE, D. 1989. Parsing, word associations and typical predicate-argument relations. In *Proceedings of International Workshop on Parsing Technologies* (Pittsburgh, PA, Aug. 1989). Carnegie Mellon Univ., Pittsburgh, PA, 389-398.

DONG, Z. AND DONG, Q. 2000. *HowNet* [online]. Available at http://www.keenage.com/zhiwang/e_zhiwang.html

DORR, B. J., LEVOW, G. A., AND LIN, D. 2000. Building a Chinese-English mapping between verb concepts for multilingual applications. In *Proceedings of 4th Conference of the Association for Machine Translation in the Americas, AMTA-2000* (Cuernavaca, Mexico, Oct. 2000), J. S. WHITE, Ed. Springer, New York, NY, 1-12.

FARRERES, X., RIGAU, G., AND RODRIGUEZ, H. 1998. Using WordNet for building WordNets. In *Proceedings of the ACL Workshop on the Usage of WordNet in Natural Language Processing Systems* (Montréal, Canada, Aug. 1998), S. HARABAGIU, Ed. Association for Computational Linguistics, 65-72.

FELLBAUM, C., Ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

HAMP, B. AND FELDWEG, H. 1997. GermanNet: A lexical-semantic net for German. In *Proceedings of ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, (Madrid, Spain, July 199), P. VOSSEN ET AL. Eds. Association for Computational Linguistics, 9-15.

HARABAGIU, S., Ed. 1998. Usage of WordNet in natural language processing systems. In *Proceedings of the Workshop* (Montréal, Canada). Association for Computational Linguistics.

HARMAN, D. K. 1997. *TREC-6 Proceedings*. National Institute of Standards and Technology. Gaithersburg, MD.

HUANG, C. R. AND CHEN, K. J. 1995. Academic Sinica Balanced Corpus. Tech. Rep. 95-02/98-04. Academic Sinica, Taipei, Taiwan.

IDE, N. AND VERONIS, J. 1998. Word sense disambiguation: The state of art. *Comput. Linguist.* 24, 1, 1-40.

KELLY, E. AND STONE, P. 1975. *Computer Recognition of English Word Senses*. North-Holland, Amsterdam, The Netherlands.

KILGARRIFF, A. 1998. SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proceedings of First International Conference on Language Resources and Evaluation* (Granada, Spain, May 1998), 581-588.

KILGARRIFF, A. AND ROSENZWEIG, J. 2000. English SENSEVAL: Report and results. In *Proceedings of Second International Conference on Language Resources and Evaluation* (Athens, Greece, May-June 2000).

LANDES, S., LEACOCK, C., AND TENGI, R. I. 1998. Building semantic concordances. In *WordNet: An Electronic Lexical Database*, C. FELLBAUM, Ed. MIT Press, Cambridge, MA, 199-216.

LUA, K. T. 1997. An efficient inductive unsupervised semantic tagger. *Comput. Process. Oriental Lang*. 11, 1, 35-47.

MANNING, C. D. AND SCHUTZE, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

MARSHALL, I. 1987. Tag selection using probabilistic methods. In *The Computational Analysis of English*, R. GARSIDE ET AL. Eds. Longman, London, 42-56.

MEI, J., ZHU, Y., GAO, Y., AND YIN, H. 1982. *tong2yi4ci2ci2lin2*. Shanghai Dictionary Press.

NG, H. T. AND LEE, H. B. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of 34th Annual Meeting of Association for Computational Linguistics* (Santa Cruz, CA, June 1996). Association for Computational Linguistics, 40-47.

RILA, M. 1998. The use of WordNet in information retrieval. In *Proceedings of ACL Workshop on the Usage of WordNet in Natural Language Processing Systems* (Montréal, Canada, Aug. 1998), S. HARABAGIU, Ed. Association for Computational Linguistics, 31-37.

RUIZ, M., DIEKEMA, A., AND SHERIDAN, P. 1999. CINDOR conceptual interlingua document retrieval: TREC-8 Evaluation. In *Proceedings of Eighth Text Retrieval Conference* (Gaithersburg, MD, Nov. 1999). National Institute of Standards and Technology, 597-606.

SALTON, G. AND BUCKLEY, C. 1998. Term weighting approaches in automatic text retrieval. *Inf. Process. Manage.* 5, 24, 513-523.

VOSSEN, P. 1998. EuroWordNet: Building a multilingual database with wordnets for European languages. *ELRA Newsl.* 3, 1, 7-10.

YAROWSKY, D. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the Fifteenth International Conference on Computational Linguistics* (Nantes, France, Aug. 1992) C. BOITET, Ed. 454-460.

Received January 2001; accepted May 2002.