Sense-Tagging Chinese Corpus

Hsin-Hsi Chen
Department of Computer Science and
Information Engineering
National Taiwan University
Taipei, TAIWAN
hh_chen@csie.ntu.edu.tw

Chi-Ching Lin
Department of Computer Science and
Information Engineering
National Taiwan University
Taipei, TAIWAN
cclin@nlg2.csie.ntu.edu.tw

Abstract

Contextual information and the mapping from WordNet synsets to Cilin sense tags deal with word sense disambiguation. The average performance is 63.36% when small categories are used, and 1, 2 and 3 candidates are proposed for low, middle and high ambiguous words. The performance of tagging unknown words is 34.35%, which is much better than that of baseline mode. The sense tagger achieves the performance of 76.04%, when unambiguous, ambiguous, and unknown words are tagged.

1 Introduction

Tagging task, which adds lexical, syntactic or semantic information to raw text, makes materials more valuable. The researches on part of speech (POS) tagging have been a long history, and achieve very good results. Many POS-tagged corpora are available. The accuracy for POS-tagging is in the range of 95% to 97%[1]. In contrast, although the researches on word sense disambiguation (WSD) are also very early (Kelly and Stone, 1975), large-scale sense-tagged corpus is relatively few. In English, only some sense-tagged corpora such as HECTOR (Atkins, 1993), DSO (Ng and Lee, 1996), SEMCOR (Fellbaum, 1997), and SENSEVAL (Kilgarriff, 1998) are available. For evaluating word sense disambiguation systems, the first SENSEVAL (Kilgarriff and Rosenzweig, 2000) reports that the performance for a fine-grained word sense disambiguation task is at around 75%.

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. Three approaches have been proposed for WSD, including dictionary/thesaurus-based approach, supervised learning, and unsupervised learning. The major differences are what kinds of resources are used, i.e., dictionary versus text corpus, and sense-tagged corpus versus untagged corpus. A good survey refers to the paper (Ide and Veronis, 1998). Compared with English, Chinese does not have large-scale sense-tagged corpus. The widely available corpus is Academic Sinica Balanced Corpus abbreviated as ASBC hereafter (Huang and Chen, 1995), which is a POS-tagged corpus. Thus, a computer-aided tool to sense-tag Chinese corpus is indispensable.

This paper presents a sense tagger for Mandarin Chinese. It is organized as follows. Section 2 discusses the degree of polysemy in Mandarin Chinese from several viewpoints. Section 3 presents WSD algorithms for tagging ambiguous words and unknown words. Section 4 shows our experimental results. Finally, Section 5 concludes the remarks.

2 Degree of Polysemy in Mandarin Chinese

The degree of polysemy is defined as the average number of senses of words. We adopt tagging set from tong2yi4ci2ci2lin2 (同義詞詞林) abbreviated as Cilin (Mei, et al., 1982). It is composed of 12 large categories, 94 middle categories, and 1,428 small categories.

Small categories (more fine granularity) are used to compute the distribution of word senses. Besides Cilin, ASBC is employed to count frequency of a word. Total 28,321 word types appear both in Cilin and in ASBC corpus. Here a word type corresponds to a dictionary

---

Table 1. The Distribution of Word Senses

Low Ambiguity		Middle Ambiguity		High Ambiguity			
Degree	#Word Types	Degree	#Word Types	Degree	#Word Types	Degree	#Word Types
2	4261 (71.95%)	5	186 (3.14%)	9	14 (0.24%)	14	1 (0.02%)
3	948 (16.01%)	6	77 (1.30%)	10	8 (0.14%)	15	1 (0.02%)
4	344 (5.81%)	7	42 (0.71%)	11	3 (0.05%)	17	1 (0.02%)
		8	25 (0.42%)	12	4 (0.07%)	18	1 (0.02%)
				13	5 (0.08%)	20	1 (0.02%)
Sum	5553 (93.77%)	Sum	330 (5.57%)	Sum			39 (0.66%)
Total Word Types		5922					

Table 2. The Distribution of Word Senses with Consideration of POS

Degree \ POS		N	V	A	F	K
Low	2	1441 (81.05%)	1056 (71.79%)	580 (79.67%)	14 (77.78%)	101 (73.72%)
	3	238 (13.39%)	238 (16.18%)	115 (15.80%)	4 (22.22%)	25 (18.25%)
	4	55 (3.09%)	99 (6.73%)	20 (2.75%)		7 (5.11%)
Middle	5	26 (1.46%)	41 (2.79%)	9 (1.24%)		3 (2.19%)
	6	12 (0.67%)	13 (0.88%)	2 (0.27%)		1 (0.73%)
	7	3 (0.17%)	13 (0.88%)	2 (0.27%)		
	8	2 (0.11%)	6 (0.40%)			
High	9	1 (0.06%)	1 (0.07%)			
	11		1 (0.07%)			
	12		1 (0.07%)			
	13		1 (0.07%)			
	19		1 (0.07%)			
Total Word Types		1778	1471	728	18	137

entry. Of these, 5,922 words are polysemous, i.e., they have more than one sense. Table 1 lists the statistics. We divide the ambiguity degree into three levels according to the number of senses of a word. It includes low (2-4), middle (5-8), and high ambiguity (>8). The statistics shows that 93.77% of word types belong to the class of low ambiguity.

We further consider POS when computing the distribution of word senses. Table 2 shows the statistics. N, V, A, F, and K denote nouns, verbs, adjectives, numerals, and auxiliaries (adverbs), respectively. We can find most of words belong to the class of low ambiguity no matter which POSes they are. Besides, the ambiguity is decreased when POS is considered. The number of polysemous words is down to 4,132. For A and K, the number of senses is no more than 7, and the percentages in the class of low degrees are 98.22% and 97.08%, respectively. For N and V, there are some high ambiguous words. In particular, the verb (打, da3) has 19 senses[2]. The percentages in the class of low degrees are 97.53% and 94.70%, respectively.

Then, the frequency of word types is considered. ASBC corpus is used to compute the occurrences of word types. Table 3 lists the statistics. A word token is an occurrence of a type in the corpus. On the average, the words of low, middle and high ambiguity appear 205.96, 1926.65, and 4480.28 times, respectively. Table 1 shows 93.77% of polysemous words belong to the class of low ambiguity, but Table 3 illustrates they only

---

[2] The word (打, da3) has 20 senses. Besides verb usage, it also functions as an auxiliary.

Table 3. The Distribution of Word Senses with Consideration of Frequencies

Low Ambiguity			Middle Ambiguity			High Ambiguity		
Types	Tokens	#Tokens/ #Types	Types	Tokens	#Tokens/ #Types	Types	Tokens	#Tokens/ #Types
5553	1143686	205.96	330	635796	1926.65	39	174731	4480.28
93.77%	58.52%		5.57%	32.53%		0.66%	8.94%	

Table 4. The Distribution of Word Senses and Frequencies with Consideration of POS

Ambiguity	Frequency	Low	Middle	High	Sum	Percentage
Low	Types (C)	3112	734	147	3993	96.64%
	Tokens (A)	70131	230955	735819	1036905	85.52%
	A/C	22.54	314.65	5005.57	259.68	
Middle	Types (C)	42	62	29	133	3.22%
	Tokens (A)	1905	14667	153307	169879	14.01%
	A/C	45.36	236.56	5286.45	1277.29	
High	Types (C)	0	2	4	6	0.15%
	Tokens (A)	0	843	4847	5690	0.47%
	A/C	0	421.5	1211.75	948.33	
Sum	Types (C)	3154	798	180	4132	
	Tokens (A)	72036	246465	893973	1212474	
	A/C	22.84	308.85	4966.52		
%	Types (C)	76.33%	19.31%	4.36%		
	Tokens (A)	5.94%	20.33%	73.73%		

occupy 58.52% of tokens in ASBC corpus.

Table 4 summarizes the distribution of word senses and frequencies. Low frequency denotes the number of occurrences less than 100, middle frequency denotes the number of occurrences between 100 and 1000, and high frequency denotes the number of occurrences more than 1000. Rows C and A in Table 4 denote number of word types and word tokens, respectively. The last column denotes percentage for each ambiguity degree. For example, the percentage of word types with low ambiguity is 96.64% (i.e., 3993/4132). This table shows the following two phenomena:

(1) POS information reduces the degree of ambiguities. Total 8.94% of word tokens are high ambiguous in Table 3. It decreases to 0.47% in Table 4.

(2) High ambiguous words tend to be high frequent. From the row of low ambiguity, there are 3,112 low-frequent words. They occur 70,131 times in ASBC corpus. Comparatively, there are only 881 middle- or high-frequent words, but they occur 966,774 times. That is, 23.67% of word types are middle- or high-frequent words, and they occupy 94.06% of word tokens. From the row of high ambiguity, there are only a few words, but they occur frequently in the ASBC corpus. It shows that semantic tagging is a challengeable problem in Mandarin Chinese.

3 Semantic Tagging

3.1 Tagging Unambiguous Words

In the semantic tagging, the small categories are selected. 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. Because we do not have a semantically tagged corpus for training, we try to acquire the context

for each semantic tag starting from the unambiguous words.

ASBC corpus is the target we study. At the first stage, only those words that are unambiguous in Cilin, and also appear in ASBC corpus are tagged. Figure 1 shows this case.
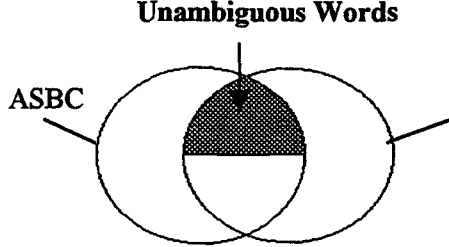
Unambiguous Words

Figure 1. Tagging Unambiguous Words

An unambiguous word (and hence its sense tag) is characterized by the words surrounding it. The window size is set to 6, and stop words are removed. A list of stop words is trained from ASBC corpus. The words of POSes Neu (數詞), DE (的，之，得，地), SHI (是), FW (外文標記), C (連接詞), T (語助詞), and I (感嘆詞) are regarded as stop words. A sense tag $Ctag$ is 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 by the following two ways.

(1) MI metric (Church, *et al.*, 1989)

$$MI\ (Ctag, cw) =$$

$$\log_2 \frac{P(Ctag, cw)}{P(Ctag)P(cw)} =$$

$$\log_2 \frac{f(Ctag, cw)}{f(Ctag)f(cw)} \times N$$

where $P(Ctag)$ is the probability of $Ctag$,
$P(cw)$ is the probability of cw,
$P(Ctag, cw)$ is the cooccurrence probability of $Ctag$ and cw,
$f(Ctag)$ is the frequency of $Ctag$,
$f(cw)$ is the frequency of cw,
$f(Ctag, cw)$ is the cooccurrence frequency of $Ctag$ and cw, and
N is total number of words in the corpus.

(2) EM metric (Ballesteros and Croft, 1998)

$$em(Ctag, cw) =$$

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

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

3.2 Tagging Ambiguous Words

At the second stage, we deal with those words that have more than one sense in the Cilin. Figure 2 shows the words we consider.
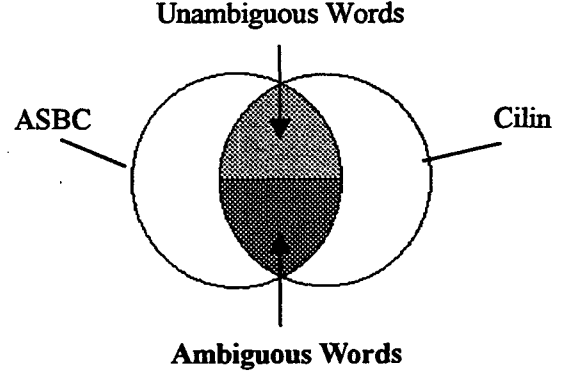
Unambiguous Words

Ambiguous Words
Figure 2. Tagging Ambiguous Words

The approach we adopted on semantic tagging rests on an underlying assumption: each sense has a characteristic context that is different from the context of all the other senses. In addition, all words expressing the same sense share the same characteristic context. We will apply the information trained at the first stage to selecting the best sense tag from the candidates of each ambiguous word. Recall that a vector corresponds to a sense tag. We employ the similar way specified in Section 3.1 to identify the context vector of an ambiguous word. A cosine formula shown as follows measures the similarity between a sense vector and a context vector, where w and v are a sense vector and a context vector, respectively. The sense tag of the highest similarity score is chosen.

$$\cos\ (w, v) = \frac{w \cdot v}{|w|\,|v|}$$

We retrain the sense vector for each sense tag after the unambiguous words are resolved.

3.3 Tagging Unknown Words

Those words that appear in ASBC corpus, but are not gathered in Cilin are called *unknown words*. All the 1,428 sense tags are the possible candidates. Intuitively, the algorithm in Section 3.2 can be applied directly to select a sense tag from the 1,428 candidates. However, the candidate set is very large. Here we adopt outside evidences from the mapping among WordNet synsets (Fellbaum, 1998) and Cilin
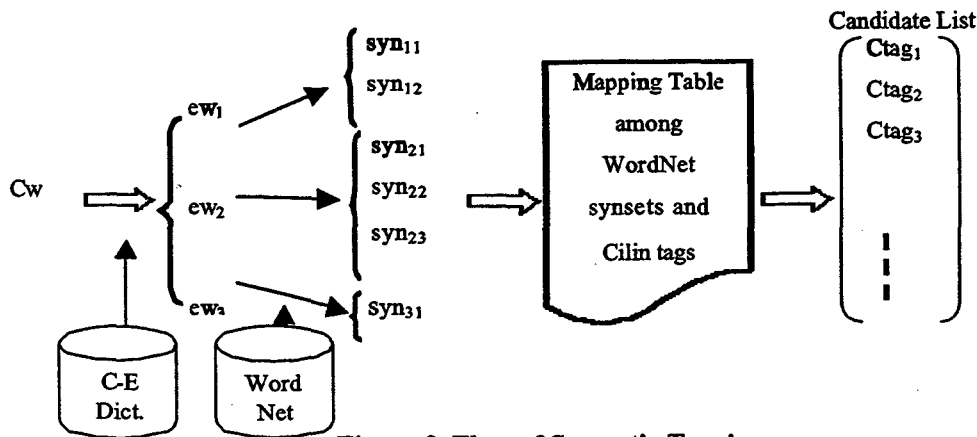
Figure 3. Flow of Semantic Tagging

sense tags to narrow down the candidate set. Figure 3 summarizes the flow of our algorithm. It is illustrated as follows.

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

(2) Find all the synsets of the English translations by looking up WordNet. We do 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. How the mapping table is set up will be discussed in Section 3.3.1.

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

Figure 4 shows the unknown words we deal with at this stage. Those words that are not gathered in our Chinese-English dictionary are not considered, so that only parts of unknown words are resolved. In other words, there remain words without sense tags.
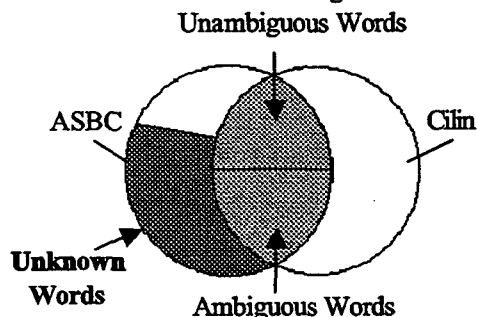

Figure 4. Tagging Unknown Words

3.3.1 Mapping SynSets to Cilin Sense Tags

At first, we put unambiguous words (specified in Section 3.1) into WordNet by looking up 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 will find the most similar synset with Chinese sense tag.

(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 translate the context vector of the Chinese sense into English, compare it with vectors of the synsets, and select the most similar 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 in the similar way as (a). If it 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.

Next, we consider the ambiguous words. Chinese-English dictionary lookup finds all the English translations. WordNet search collects

11

the synset candidates for the translations. Some synsets are selected and regarded as the mapping of the Cilin sense tag. Here the problems of translation ambiguity and target polysemy must be faced. In other words, not all English translations cover the Cilin sense. Because the goal is to find a mapping table between WordNet synsets and Cilin sense tags, we neglect the problem of translation ambiguity and follow the method in the previous paragraph to choose the most similar synsets.

During mapping, English translations of a word may not be found in the Chinese-English dictionary, and WordNet may not gather the English translations even dictionary look-up is successful. Thus, only 1,328 of 1,428 Cilin tags are mapped to WordNet synsets. From the other view, there remains some WordNet synsets that do not correspond to any Cilin sense tags. Let such a synset be S_i. We follow the relational pointers like hypernym, hyponym, similar, derived, antonym, or participle to collect the neighboring synsets denoted by S_j. The following method selects suitable Cilin tag(s) for S_i.

    (1) If S_j is the only one synset that has been mapped to Cilin tags, we choose a Cilin tag and map S_i to it.

    (2) If there exists more than one S_j (say, S_{j1}, S_{j2}, ..., S_{jn}) that has been mapped to Cilin tags, we choose the Cilin tags that more synsets map to.

The above method is called a *more restrictive* scheme. An alternative method (called *less restrictive* method) is: all the Cilin tags that the neighboring synsets map to are selected. If Cilin tags cannot be found from neighboring synsets, we extend the range one more, and repeat the selection procedure again until all the synsets are considered.

4 Experiments
4.1 Test Materials
We sample documents of different categories from ASBC corpus, including philosophy (10%), science (10%), society (35%), art (5%), life (20%) and literary (20%). There are 35,921 words in the test corpus. Research associates tag this corpus manually. At first, they mark up the ambiguous words by looking up the Cilin dictionary. Next, they tag the unknown words. A list of candidates is proposed by looking up the mapping table. Because the mapping table may have errors, the annotators assign a tag "none" when they cannot choose a solution from the proposed candidates. Total 435 of 1,979 words are tagged with "none" with the more restrictive method. In contrast, only 346 words are labeled with "none" with the less restrictive method. The tag mapper achieves 82.52% of performance approximately.

4.2 Tagging Ambiguous Words
Table 5 shows the performance of tagging ambiguous words. MI defined in Section 3.1 is used. Total 11,101 words are tagged. The performance of tagging low, middle, and high ambiguous words are 62.60%, 31.36%, and 27.00%, respectively. Table 6 shows that the performance is improved, in particular, the classes of middle- and high- ambiguity, when EM (defined in Section 3.1) is used. The overall performance is increased from 49.55% to 52.85%.

In the previous experiments, only one sense is reported for each word. If we report more than one sense for middle and high ambiguous words, the performance is improved. Table 7 shows that the first 2 and 3 candidates are 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 1 candidate, 2 candidates, and 3 candidates are proposed, respectively. In this case, 7,034 of 11,101 words are tagged correctly. That is, the performance is 63.36%.

In the next experiment, we adopt middle categories (i.e., 94 categories) rather than the above small categories (i.e., 1428 categories). Table 8 shows that the overall performance is improved by 11.05%. It also lists the results with the combinations of first-n and middle categories. Under the middle categories and 1-3 proposed candidates, the performance for tagging low, middle and high ambiguous words are 71.02%, 73.88%, and 75.94%, respectively. Total 8,033 of 11,101 words are tagged correctly. In other words, the performance is 72.36%.

Table 5. Performance of Tagging Ambiguous Words using MI

Ambiguity / Word Tokens	Low	Middle	High	Summary
Total Tokens	6601	3511	989	11101
Correct Tokens	4132	1101	267	5500
Correct Rate	62.60%	31.36%	27.00%	49.55%

Table 6. Performance of Tagging Ambiguous Words using EM

Ambiguity / Word Tokens	Low	Middle	High	Summary
Total Tokens	6601	3511	989	11101
Correct Tokens	4223	1334	310	5867
Correct Rate	63.98%	37.99%	31.34%	52.85%

Table 7. Performance of Tagging using the First-n and EM

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 8. Performance of Tagging using First-n and Middle Categories

Ambiguity / First-n/Categories		Low	Middle	High	Middle and High
1	Small	63.98%	37.99%	31.34%	36.53%
	Middle	71.02%	56.19%	43.78%	53.47%
2	Small		60.92%	53.99%	59.40%
	Middle		73.88%	65.72%	72.09%
3	Small		71.35%	67.95%	70.60%
	Middle		79.27%	75.94%	78.53%

4.3 Tagging Unknown Words

There are 1,979 unknown words in our test corpus. Total 1,663 words have been tagged manually. In the experiments, we consider the effects from training corpus and mapping table. Table 9 shows the performance. M1 and P1 employ more restrictive mapping table, while M2 and P2 adopt less restrictive mapping table. M1 and M2 use the training result in Section 3.1 (i.e., unambiguous words), while P1 and P2 utilize the training result in Section 3.2 (i.e., unambiguous and ambiguous words). In the baseline model, all 1428 Cilin tags are the candidates of unknown words. The performance is worse. On the average, the precision is 1.22%. M1 is the best because more restrictive mapping table reduces the possibility of mapping errors. This table also lists the performance of each category. It meets our expectation, i.e., tagging verb is harder than tagging other categories. Next we

use POS to improve the performance. POS narrows down the number of candidates, so that the overall performance is enhanced from 27.13%% to 34.35%%.

In summary, we consider the overall performance of tagging our sample data. Recall that there are 35,921 words in the test corpus. Except the stop words that are not tagged by the sense tagger, there remain 13,586 unambiguous words, 11,101 ambiguous words, and 1,633 unknown words for tagging. From Tables 6 and 9, we know 5,867 unambiguous words and 561 unknown words are tagged correctly. The sense tagger achieves the performance of 76.04%.

5. Conclusion

This paper analyzes the polysemy degree in Mandarin Chinese. We consider the distribution of word senses from POS and frequency. Under the Cilin small categories, 23.67% of word types in ASBC corpus are

Table 9. Performance of Tagging Unknown Words

Categories	#Tokens		Baseline	M1	M2	P1	P2	M1(POS)
All	1633	Correct	20	443	395	438	396	561
		Precision	1.22%	27.13%	24.19%	26.82%	24.25%	34.35%
N	858	Correct	11	255	228	255	231	320
		Precision	1.28%	29.72%	26.57%	29.72%	26.92%	37.30%
V	619	Correct	5	144	124	137	120	167
		Precision	0.81%	23.26%	20.03%	22.13%	19.39%	26.98%
A	58	Correct	0	5	5	5	5	28
		Precision	0	8.62%	8.62%	8.62%	8.62%	48.28%
F	4	Correct	1	1	1	1	1	4
		Precision	25.00%	25.00%	25.00%	25.00%	25.00%	100.00%
K	94	Correct	3	38	37	40	39	42
		Precision	3.19%	40.43%	39.36	42.55	41.49	44.68%

middle or high frequent words, but they occupy 94.06% of word tokens. We adopt contextual information and mapping from WordNet synsets to Cilin sense tags to deal with this challengeable problem. The performances for tagging low, middle and high ambiguous words are 63.98%, 60.92%, and 67.95% when small proposed. Comparatively, the performances categories are used and 1-3 candidates are 71.02%, 73.88%, and 75.94% by using middle categories. The performance of tagging unknown words is 34.35%. It is worse than that of tagging ambiguous words, but is much better than that of the baseline mode. The overall performance is the sense tagger is 76.04%. Although sense tagging does not achieve the performance of POS tagging, the sense tagger proposed in this paper is still a useful computer-aided tool to reduce the human cost on tagging a large-scale corpus.

References

Atkins, S. (1993) "Tools for Computer-Aided Lexicography: the Hector Project," *Acta Linguistica Hungarica*, 41, pp. 5-72.

Ballesteros, L. and Croft, W.B. (1998) "Resolving Ambiguity for Cross-Language Information Retrieval," *Proceedings of the 21st Annual International ACM SIGIR Conference*, pp. 64-71.

Church, K.W., et al. (1989) "Parsing, Word Associations and Typical Predicate-Argument Relations." *Proceedings of International Workshop on Parsing Technologies*, pp. 389-398.

Huang, C.R. and Chen, K.J. (1995) "Academic Sinica Balanced Corpus," Technical Report 95-02/98-04, Academic Sinica, Taipei, Taiwan.

Fellbaum, C. editor (1998) *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, Mass.

Ide, N. and Veronis, J. (1998) "Word Sense Disambiguation: The State of Art," *Computational Linguistics*, 24(1), pp. 1-40.

Kelly, E. and Stone, P. (1975) *Computer Recognition of English Word Senses*, North-Holland, Amsterdam.

Kilgarriff, A. (1998) "SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs," *Proceedings of First International Conference on Language Resources and Evaluation*, Granada, pp. 581-588.

Kilgarriff, A. and Rosenzweig, J. (2000) "English SENSEVAL: Report and Results," *Proceedings of Second International Conference on Language Resources and Evaluation*.

Manning, C.D. and Schutze, H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Mass.

Marshall, I. (1987) "Tag Selection using Probabilistic Methods," in Roger Garside, Geoffrey Leech and Geoffrey Sampson (editors), *The Computational Analysis of English*, Longman, pp. 42-56.

Mei, J.; et al. (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," *Proceedings of 34th Annual Meeting of Association for Computational Linguistics*, pp. 40-47.