Design and Prototype of a Large-scale and Fully Sense-tagged Corpus

Sue-jin Ker[1], Chu-Ren Huang[2], Jia-Fei Hong[3], Shi-Yin Liu[1], Hui-Ling Jian[1], I-Li Su[2] and Shu-Kai Hsieh[4]

[1] Department of Computer and Information Science, Soochow University, Taiwan
[2]Institute of Linguistics, Academia Sinica, Taiwan
[3] Graduate Institute of Linguistics, National Taiwan University, Taiwan
[4] Department of English, National Taiwan Normal University, Taiwan

Abstract. Sense tagged corpus plays a very crucial role to Natural Language Processing, especially on the research of word sense disambiguation and natural language understanding. Having a large-scale Chinese sense tagged corpus seems to be very essential, but in fact, such large-scale corpus is the critical deficiency at the current stage. This paper is aimed to design a large-scale Chinese full text sense tagged Corpus, which contains over 110,000 words. The Academia Sinica Balanced Corpus of Modern Chinese (also named Sinica Corpus) is treated as the tagging object, and there are 56 full texts extracted from this corpus. By using the N-gram statistics and the information of collocation, the preparation work for automatic sense tagging is planned by combining the techniques and methods of machine learning and the probability model. In order to achieve a highly precise result, the result of automatic sense tagging needs the touch of manual revising.

Keywords: Word Sense Disambiguation, Sense Tagged Corpus, Natural Language Processing, Bootstrap Method.

1 Introduction

The availability of large-scale sense tagged corpus is crucial for many Natural Language Processing systems, because such a corpus usually contains a lot of rich semantic knowledge resources that can be applied as the basis for the representation and processing of meaning. Due to the popularization of digital documents, there are more and more various types of corpora appearing and the content in the corpora are very abundant. Basically, a corpus with the complete tagging information is more helpful to researches. Some corpora only simply display the content of original texts and some cover the relevant information, such as part-of-speech (POS) and senses. At present, a few corpora having the POS tagging have already existed, such as Sinica Corpus with 10 million words, Chinese Gigaword Corpus and so on. However, the corpora with sense tagging are fairly few no matter in Chinese or English. To the

research in theoretical linguistics, the resources on sense tagging or semantics are useful in providing many rich materials or the basic structures. To the research in computational linguistics, those resources play the crucial breakthrough applying on the core works in Natural Language Processing, for instance, the multiple senses analysis, such as WSD, and Natural Language Understanding. Moreover, the statistic data extracted from Sense tagged corpus can be implemented in the research issues such as Information Retrieval, Information Extraction, Text Summarization, Automatic Question Answering and so on.

Massive accurate sense tagged data does provide rich resources to different relevant researches in computational linguistics. However, the main bottleneck of making a Chinese sense tagging corpus is the deficiency of reference material for automatic tagging. In addition, manually tagging is very expensive and time-consuming. Those reasons make the sense tagging for corpus become more difficult. In the recent years, in many researches, the need of large-scale sense tagged corpus is growing. The completeness of a sense tagged corpus often affects the research direction and the accuracy of the research result. In certain languages, there has existed some representative sense tagged corpora, such as the SemCor [3] for English language materials and SENSEVAL [4], providing a multi-lingual full text tagged corpus in Czech, Dutch, Italian and English. When we look back to the corpus development in Chinese, a large-scale corpus is still the critical deficiency at the current stage. There only exists several small Chinese Sense tagged corpora, for example, the SENSEVAL-2, covering the Chinese sense tagging for 15 Chinese words, and SENSEVAL -3 for 20 Chinese words. There is a huge gap between the scale of the corpus and the real language environment. Cost is the main issue in constructing a massive corpus. Manual tagging does acquire a higher accuracy rate, but the cost of manual tagging is too expensive. Besides, finding the proper experts for doing sense tagging manually is another issue. In order to overcome such difficulties, we propose a method of semi-automatic tagging as the preparation work for doing manual sense tagging and then the result from the semiautomatic tagging can be revised manually by the professionals. Here, the Academia Sinica Balanced Corpus of Modern Chinese (also named Sinica Corpus) is treated as the tagging object. We extract the words in the articles from Sinica Corpus and design a large-scale Chinese sense tagged corpus for the researches in Natural Language Processing.

2 Dictionary of Sense Discrimination

The Dictionary of Sense Discrimination in the third edition [5] used in this experiment is developed by the CKIP (Chinese Knowledge and Information Processing) group. CKIP group is a joint research team formed by the Institute of Information Science and the Institute of Linguistics of Academia Sinica in Taiwan. The entries are limited in the middle frequency term of Modern Chinese words. Each entry has a rich information list, including the sense definition, the phonetic symbols, meaning facets, part-of-speech (POS), example sentences and the synset number corresponding to Princeton English WordNet 1.6 (http://wordnet.princeton.edu/). As shown in Fig. 1, the entry “feng1 kuang2” has two senses. The first sense corresponds

to the English WN synset, “crazy”; and the second sense corresponds to the WN synset, “madly”. In this example, each sense can be divided into two meaning facets respectively.

瘋狂 feng1 kuang2 ㄈㄥ ㄎㄨㄤˊ
詞義1：【不及物動詞，VH；名詞，nom】形容人因精神錯亂而舉止失常。｛crazy, 00872382A｝
義面1：【不及物動詞，VH】形容人因精神錯亂而舉止失常。
例句：片中一名〈瘋狂〉殺手，拿著剃刀。
例句：石門五子命案的父母與其說是迷信，不如說是〈瘋狂〉。
義面2：【名詞，nom】形容人因精神錯亂而舉止失常。
例句：因此，石門五子命案的〈瘋狂〉，其實也正是我們社會瘋狂的一粒種籽啊！
詞義2：【不及物動詞，VH；名詞，nom】形容行爲或事物無節制，超乎平常的程度。通常用於人的情感或事件的程度。｛madly, 00045197R｝
義面1：【不及物動詞，VH】形容行爲或事物無節制，超乎平常的程度。通常用於人的情感或事件的程度。
例句：他〈瘋狂〉的愛上一個女孩子。
例句：每年都有不計其數的台灣客前往香港〈瘋狂〉大採購。
例句：當時少棒青少棒在台灣很〈瘋狂〉，連我們城市的小孩子也愛打棒球。
義面2：【名詞，nom】形容行爲或事物無節制，超乎平常的程度。通常用於人的情感或事件的程度。
例句：經過一陣〈瘋狂〉後，大家都累了，個個都喊著喉嚨痛、腳痛。
例句：死了七九百餘人的人民教室案，也使人想到愈來愈多的宗教〈瘋狂〉事件。
例句：只要幅度不超過，則多頭仍然大有可爲，但仍切忌一味追高的〈瘋狂〉舉動。

Fig. 1. The sense discrimination for “feng1 kuang2 (crazy/ madly)

3 The Data Source for Sense Tagging

In the experiment, we use the Academia Sinica Balanced Corpus of Modern Chinese (also named Sinica Corpus) as the basis for semi-automatic sense tagging. The content of this corpus has been segmented and marked with the POS tags. In order to preserve the context completeness, `text’ is chosen as the processing unit of tagging materials. The basic principle of sense tagging here is to do a full text tagging. However, the construction of the Dictionary of Sense Discrimination is still under processing, so not all the words in the corpus are included in the dictionary at the moment. For those words that cannot be found in the dictionary, what we do is to use the POS tags to mark them. Tagging by POS tags is able to disambiguate the word senses, so the POS tagging here is fairly general.

Based on coverage rate for the words in the dictionary appearing in the article content and the length of article, there are 56 articles, containing 114,066 words and 148,863 characters, extracted from the corpus. The statistics for the distribution of article subjects is shown in Table 1. There are many articles in the subject for *Literature*, but the type, *Life*, has longer article length. Analysis of the statistics shows that the subject *Literature* has the most articles and the subject *Life* contains the most words.

Overall, among the target words for sense tagging, there are 863 words having only one POS (one lemma to one POS) and 650 words having multiple POS. The frequency of the appearance for the ones having one POS is 12,124 times and 23,521 times for multi POS ones. The statistic of POS distribution is shown in Table 2. The sense amount for multi POS words are from 2 (e.g. 自然 D "natural"、堆 Nf "a pile" and 喜 VK "like") to 27 (e.g. 吃 VC "eat") and the average sense number is 2.97. If the data is distinguished by POS, particles have the most senses by reaching the average of 4.83 and the lowest one is interjections, where the average of sense amount is 1.32. If the statistics is based on lemmas, without considering the difference of the POS, there are 598 lemmas from the extracted articles that have been included in the Dictionary of Sense Discrimination, so the average sense number of each lemma is 4.53.

Table 1. The distribution of article subjects in the tagged corpus.

Subject	no. of Article	Article Length	
		in words	in characters
Philosophy	4	1451	1976
Society	5	27385	35918
Life	12	57605	74710
Literature	35	27625	36259
Total	56	114066	148863

Table 2. The distribution of POS for the tagged corpus.

POS	no. of word	no. of instances	Example
Intransitive Verb	231	3,317	對$_{VH}$, 跑$_{VA}$, 走$_{VA}$
Preposition	51	1,854	在$_{P}$, 跟$_{P}$, 到$_{P}$
Transitive Verb	373	5,733	說$_{VE}$, 沒有$_{VJ}$, 開始$_{VL}$
Noun	321	5,070	人家$_{Nh}$, 感覺$_{Na}$, 下$_{Ncd}$
Adjective	21	45	一般$_{A}$, 原$_{A}$, 定期$_{A}$
Determinatives	55	3,175	那$_{Nep}$, 前$_{Nes}$, 多$_{Neqa}$
Postposition	31	455	上$_{Ng}$, 裡$_{Ng}$, 當中$_{Ng}$
Adverb	287	8,892	就$_{D}$, 又$_{D}$, 起來$_{Di}$
Conjunction	69	1,554	就是$_{Cbb}$, 而$_{Cbb}$, 或$_{Caa}$
Measure	81	976	回$_{Nf}$, 份$_{Nf}$, 間$_{Nf}$
Particle	47	4,574	啊$_{T}$, 喔$_{T}$, 哇$_{I}$
Total	1567	35,645	

4 Sense Tagged Corpus

The file type of our sense tagged corpus is encoded in XML format. The usage of the tags is specified in Table 3. Each document is segmented by using the tag <doc> and </doc>. The content for each article can be segmented by sentences. Each sentence is marked by the tag <sent>. The tag <w> is used to segment words and it can be further divided into three respective tags: “word” for the information of lexicon, “pos” for the information of part-of-speech and “tag1” for the information of sense tagging.

For part of the sense tagging, there are three tagging types. The first type is based on the definition in Huang et al. [6]. It is a four-digit code. The first two digits are for the sense, which are used to indicate the sequence of senses in the dictionary. The third and fourth digits are for lemma and meaning facets. The second type is to deal with the tagging for punctuation. Basically, there is no point in doing the sense tagging for the punctuation, so we use those punctuations symbols as the tagging codes directly. The third type is for the unknown words, i.e. the words have not been included in the dictionary yet or have not analyzed through the sense discrimination. The POS tags of those unknown words are used as the sense tagging codes temporarily.

The whole tagged corpus contains a total of 114,066 words. The statistic result based on the tagging types is shown in Table 4. A total of 27,530 words are punctuation symbols and a total of 35,645 words are successfully tagged by assigning the sense ids. There are 50,891 unknown words are tagged by using their POS tags. We further analyze those unknown words and realize some of those unknown words are actually the English abbreviation, numbers or proper nouns, for example, （二）, CPU, National Tsing Hua University and so on. Such examples in our tagged corpus are a total of 4,258 words. In addition, due to the Dictionary of Sense Discrimination adopted in this experiment is still under construction, some words have not yet been included in the dictionary. There are 4,541 words that have not been included in the dictionary and the frequency for appearing in the corpus is 31,730 times. As for the remaining 14,903 words, they are the words that have already been included in the dictionary but they are not within the scope of tagging for this time. Therefore, we mark those words by using the POS tags because tagging by the POS tags is able to disambiguate the polysemy.

Table 3. The instruction of the tags used in the corpus.

Tag Name	Meaning	Example
<corpus>	The beginning of the corpus	<corpus>
<doc id=>	The beginning of the document and its id number	<doc id="100863">
<sent id=>	The beginning of the sentence and its id number	<sent id="1">
<w id=>	The information of the word and its id number	<w id="1">
<word>	Word	<word>人家</word>
<pos>	Part-Of-Speech	<pos>Nh</pos>
<tag1>	Sense Tagging	<tag1>0122</tag1>

Table 4. The statistic for the tagging types in the corpus.

Tagging Type	no. of instances		Meaning
Punctuation	27530		No need to tag the sense
Sense code	35645		Complete the sense tagging
POS tagging	50891	4258	No need to tag the sense (English abbreviation, numbers or proper nouns)
		31730	4541 words have not included in the dictionary yet
		14903	Not finished yet
Total	114066		

5 The Method of Sense Tagging

Generally speaking, doing manual tagging relies on a lot of people's efforts. Therefore, in order to save the costs, we design a method for semi-automatic tagging [7]. This method can complete the initial tagging task, so it can be treated as a preparation work for doing the manual tagging.

The semi-automatic tagging in the experiment has implemented the bootstrap method to gradually loosen the tagging conditions and enlarge the tagging materials. The example sentences of the dictionary is treated as the training set and the 56 articles randomly extracted from the Sinica Corpus as the testing set.

The N-gram model is used for automatic sense tagging at the first stage. Using N-gram to process the sense tagging is based on the following assumption: For the given target word, if there exist two sentences in which their N surrounding words are all the same, we infer they should be assigned the same sense tag [8]. Using N-gram here has two main purposes. The first is to enlarge the training set because it is quite often to see the similar clauses in the corpus. The second purpose is to filter the noise in the training data set and to use this to examine the inconsistency of manual tagging.

In the second stage, we use the information of collocations to increase the amount of tagged set. The information of collocations is a very powerful linguistic relation for determining the sense meaning for the target word [9]. We start by using some conditions, such word frequency, collocation words and the distance variation to the target word, as the preliminary basis. Then use the MI values to examine the association between target word and its collocation word.

After the previous stages, we do extend the tagged amount of the training data. Then, through the calculation of the probability model, we try to mark the most words with the sense information. Finally, in order to get the high accuracy of tagged data, we send the automatic tagging result back to CKIP/CWN group for manual reviewing.

The experimental results for the automatic tagging are shown in Table 5, the overall accuracy rate is 64.5%.

Table 5. The experimental result of sense tagging.

POS	no. of word	no. of instances	no. of corrected instances	Accuracy rate%
A	6	22	14	63.7%
Caa	2	38	37	97.4%
Cbb	7	231	37	16.0%
D	64	3454	2158	62.5%
Da	7	22	21	95.5%
Dfa	5	202	200	99.0%
Dfb	1	1	1	100.0%
Di	7	1146	934	81.5%
Dk	2	5	5	100.0%
I	15	693	307	44.3%
Na	98	648	570	88.0%
Nb	5	18	18	100.0%
Nc	8	29	27	93.1%
Ncd	13	283	209	73.9%
Nep	6	2227	642	28.8%
Neqa	2	38	32	84.2%
Nes	6	128	118	92.2%
Neu	3	127	114	89.8%
Nf	40	228	212	93.0%
Ng	13	147	102	69.4%
Nh	8	1668	1549	92.9%
P	33	1659	1143	68.9%
T	13	2838	1660	58.5%
VA	28	451	347	76.9%
VAC	1	4	3	75.0%
VB	9	14	14	100.0%
VC	76	1177	1065	90.5%
VCL	5	174	107	61.5%
VD	19	170	128	75.3%
VE	26	1703	1475	86.6%
VF	5	20	11	55.0%
VG	9	170	103	60.6%
VH	66	1940	664	34.2%
VHC	2	13	13	100.0%
VI	3	4	4	100.0%
VJ	19	326	206	63.2%
VK	11	63	55	87.3%
VL	5	160	140	87.5%
V_2	1	823	433	52.6%
Nom	1	1	1	100.0%
Total	650	23065	14879	64.5%

6 Conclusion

Sense Tagged Corpus plays a very important part to Natural Language Processing, especially for creating successful WSD systems. At the current stage, such large-scale Chinese Sense Tagged Corpus is only few and for between, so we design a large-scale Chinese sense tagged corpus, containing about 100,000 words, for research on natural language processing. The automatic sense tagging is a preparation work for manual tagging. The automatic tagging uses the information provided by the peripheral words, the N-gram method, the information of collocations, and the probability model to calculate the most likely sense meaning. We sincerely hope that through the completeness of the Dictionary of Sense Discrimination, it is possible to complete the task for doing the full text tagging for whole five million words contained in the Sinica Corpus

Acknowledgments. This work was partly supported by National Science Council, the ROC, under contract no. NSC-94-2213-E-031-003-, 96-2221-E-031-002- and 96-2422-H-001-002. The authors would like to thank the reviewers for their valuable comments.

References

1. Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. Sinica Corpus: Design Methodology for Balanced Corpora. In. B.-S. Park and J.B. Kim. Eds. *Proceeding of the 11 th Pacific Asia Conference on Language, Information and Computation.* Seoul:Kyung Hee University. pp.167--176 (1996)
2. Wei-yun Ma, and Chu-Ren Huang: Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. Presented at the 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy. pp. 24--28 May (2006)
3. SemCor, http://multisemcor.itc.it/semcor.php.
4. Senseval, http://www.senseval.org/.
5. Huang, Chu-Ren. (ed): Meaning and Sense. Technical Report 06-03, Chinese Wordnet Group, Academia Sinica, Taiwan (in Chinese) (2006)
6. Huang, Chu-Ren**,** Chun-Ling Chen, Cui-Xia Weng, Hsiang-Ping Lee, Yong-Xiang Chen and Keh-jiann Chen: The Sinica Sense Management System: Design and Implementation, *Computational Linguistics and Chinese Language Processing.* Vol. 10, No. 4, pp. 417-- 430 (2005)
7. Ker, Sur-Jin, Chu-Ren Huang and Jen-Nan Chen: A Preliminary Study of Large-scale Corpus Sense Tagging. The 5th Lexical Semantics Workshop, Beijing. (in Chinese) (2004)

8 Ker, Sur-Jin and Jen-Nan Chen: Large-scale Sense Tagging by Combining Machine Learning and Linguistic Knowledge. The 8th Lexical Semantics Workshop, Hong Kong. (in Chinese) (2007)

9 Yarowsky: One Sense Per Collocation, In Proceedings of ARPA Human Language Technology Workshop, Princeton (1993)