Building Chinese Sense Annotated Corpus with the Help of Software Tools

Yunfang Wu
School of Electronic Engineering and Computer Science, Peking University, Beijing 100871
wuyf@pku.edu.cn

Peng Jin
School of Electronic Engineering and Computer Science, Peking University, Beijing 100871
jandp@pku.edu.cn

Tao Guo
School of Electronic Engineering and Computer Science, Peking University, Beijing 100871
gtwcq@pku.edu.cn

Shiwen Yu
School of Electronic Engineering and Computer Science, Peking University, Beijing 100871
yusw@pku.edu.cn

Abstract

This paper presents the building procedure of a Chinese sense annotated corpus. A set of software tools is designed to help human annotator to accelerate the annotation speed and keep the consistency. The software tools include 1) a tagger for word segmentation and POS tagging, 2) an annotating interface responsible for the sense describing in the lexicon and sense annotating in the corpus, 3) a checker for consistency keeping, 4) a transformer responsible for the transforming from text file to XML format, and 5) a counter for sense frequency distribution calculating.

1 Introduction

There is a strong need for a large-scale Chinese corpus annotated with word senses both for word sense disambiguation (WSD) and linguistic research. Although much research has been carried out, there is still a long way to go for WSD techniques to meet the requirements of practical NLP programs such as machine translation and information retrieval. It was argued that no fundamental progress in WSD could be made until large-scale lexical resources were built (Veronis, 2003). In English a word sense annotated corpus SEMCOR (Semantic Concordances) (Landes et al., 1999) has been built, which was later trained and tested by many WSD systems and stimulated large amounts of WSD work. In Japanese the Hinoki Sensebank is constructed (Tanaka et al., 2006). In the field of Chinese corpus construction, plenty of attention has been paid to POS tagging and syntactic structures bracketing, for instance the Penn Chinese Treebank (Xue et al., 2002) and Sinica Corpus (Huang et al., 1992), but very limited work has been done with semantic knowledge annotation. Huang et al. (2004) introduced the Sinica sense-based lexical knowledge base, but as is well known, Chinese pervasive in Taiwan is not the same as mandarin Chinese. SENSEVAL-3 provides a Chinese word sense annotated corpus, which contains 20 words and 15 sentences per meaning for most words, but obviously the data is too limited to achieve wide coverage, high accuracy WSD systems.

This paper is devoted to building a large-scale Chinese corpus annotated with word senses. A small part of the Chinese sense annotated corpus has been adopted as one of the SemEval-2007 tasks namely “Multilingual Chinese-English Lexical Sample Task” This paper concentrates on the description of the manually annotating schemes

Proceedings of the Linguistic Annotation Workshop, pages 125–131,
Prague, June 2007.

with the help of software tools. The software tools will help human annotators mainly in the two aspects: 1) Reduce the labor time and accelerate the speed; 2) Keep the inter-annotator agreement. The overall procedure along with the software tools is illustrated in figure 1.

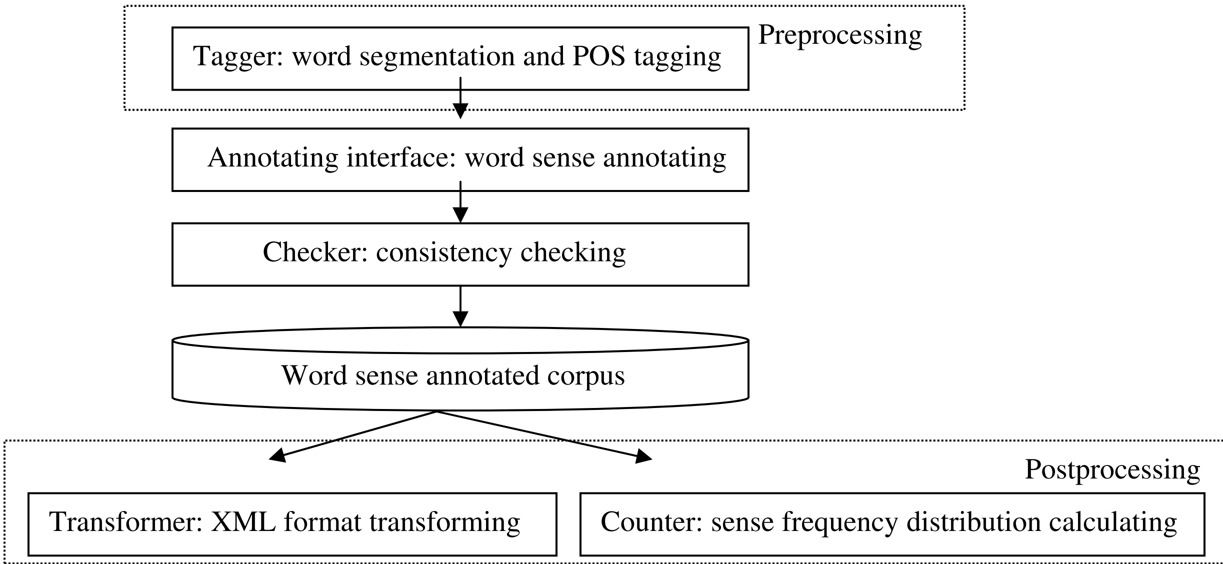

Fig.1.The overall procedure along with the software tools

This paper is so organized as follows. In section 2 the preprocessing stage (word segmentation and POS tagging) is discussed. Then in section 3 the annotating scheme and the annotating interface are demonstrated in detail. The strategy to keep consistency is addressed in section 4. And then in section 5 and 6 the two postprocessing stages are respectively presented. Finally in section 7 conclusions are drawn and future works are presented.

2 Word segmentation and POS tagging

The input data for word sense annotating is firstly word segmented and POS tagged using Peking University's POS tagger (Yu et al., 2003). The POS tagging precision is up to 97.5%, which lays a sound foundation for researches on sense annotating. This is actually to make use of the full-fledged syntactic processing techniques to deal with the semantic annotation problems. Different senses of one ambiguous word sometimes behave so differently that they bear different POS tags. Take "把握/hold" in sentence (1) as an example. The noun of "把握/hold" means "confidence", but the verb means "grasp".

(1) a 有(have) 把握/n(confidence)
b 把握/v(grasp) 住(ZHU) 机会(chance)

Due to the unique characteristic of Chinese language that lacks word inflection, the ambiguous words with different POSs are very common. According to the research of Li (1999), after POS tagging the ratio of ambiguous word occurrences in the text of People's Daily is reduced from 42% to 26%. Therefore the emphasis of manually sense annotating in this paper falls on the ambiguous words with the same part of speech. This will in turn save 16% of the annotation effort compared with the sense annotating before the preprocessing of POS tagging.

3 Word sense annotating

The resulting lexical knowledge base in this project will contain three major components: 1) a corpus annotated with Chinese word senses namely Chinese Senses Pool (CSP); 2) a lexicon containing sense distinction and description namely Chinese Semantic Dictionary (CSD); 3) the linking between the CSD and the Chinese Concept Dictionary (CCD) (Liu et al., 2002). The corpus CSP, the lexicon CSD and CCD constitute a highly relational and tightly integrated system: 1) In CSD the sense distinctions are described relying on the corpus; 2) In CSP the word occurrences are assigned sense tags according to the sense entry specified in CSD; 3) The linking between the sense entry in CSD and CCD synsets are established. The dynamic model is shown in figure 2. A software tool is developed in Java to be used as the word sense annotating interface (figure 3), which embodies the spirit of the dynamic model properly.

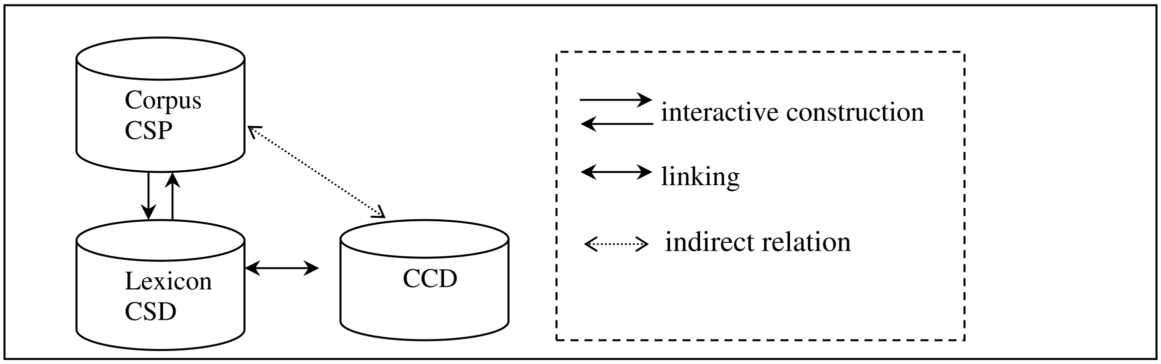

Fig 2. The dynamic model between the CSP, CSD and CCD

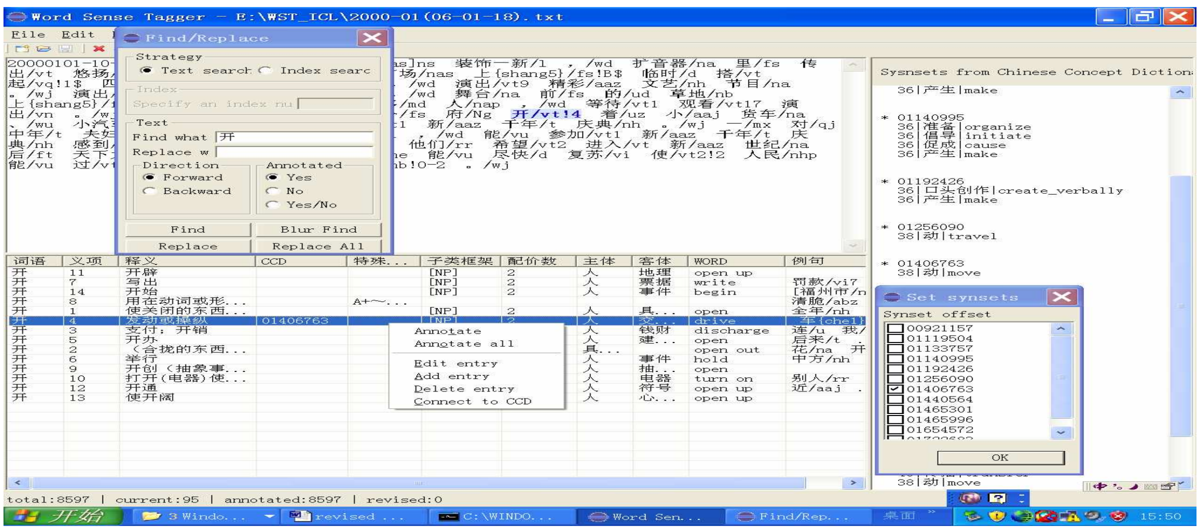

Fig3. The word sense annotating interface

3.1 Sense describing in the lexicon and sense annotating in the corpus

In this project the lexicon CSD containing sense descriptions and the corpus CSP annotated with senses are built interactively, simultaneously and dynamically. On one hand, the sense distinctions in the lexicon are made relying heavily on the corpus usage. On the other hand, using the sense information specified in the lexicon the human annotators assign semantic tags to all the instances of the word in a corpus.

In the word sense annotating interface, the sentences from CSP containing the target ambiguous words are displayed in the upper section, and the word senses with feature-based description from CSD are displayed in the bottom section.

Through reading the context in the corpus, the human annotator decides to add or delete or edit a sense entry in the lexicon. The default value of the range of the context is within a sentence, and the surrounding characters in the left and right of the target word can be specified by the annotator. Annotators can do four kinds of operations in CSD: 1) Add a sense entry and then fill in all the features; 2) Delete a sense entry along with all its feature description; 3) Edit a sense entry and change any of the features; 4) Select a sample sentence form the CSP and add it to the lexicon in the corresponding sense entry.

According to the sense specification in CSD the human annotator assigns semantic tags to the word occurrences in CSP. The operation is quite easy. When the annotator double clicks the appropriate sense entry in CSD the sense tag is automatically added to the target word.

The notable feature in this word sense annotating interface is that it provides flexible searching schemes. 1) Search sequentially (forward or backward) all the instances of an ambiguous words regardless of the annotating state; 2) Search sequentially (forward or backward) the already annotated instances; 3) Search sequentially (forward or backward) the yet un-annotated instances and 4) Search the instances of a specific ambiguous word (the window named Find/Replace in figure3, and again is shown in figure 4 for clearness).

The tool of Find/Replace is widely used in this project and has proven to be effective in annotating word senses. It allows the annotator to search for a specific word to finish tagging all its occurrences in the same period of time rather than move sequentially through the text. The consistency is more easily kept when the annotator manages many different instances of the same word than handle a few occurrences of many different words in a specific time frame, because the former method enables the annotator to establish an integrative knowledge system about a specific word and its sense distinction. Also the tool of Find/Replace provides flexible searching schemes for a specific ambiguous word. For instance, search in the corpus with different directions (forward/backward) and search with different annotating states (annotated/un-annotated/both). Using the tool the annotator can also replace some specific word occurrences in the corpus (often with special POS tags) with a sense tag, thus can finish annotating the corpus quickly and with a batch method. For instance the POS tag of "vq" (means verb complement) often uniquely corresponds to a specific verb sense such as "开/vq→开/vq!8".

There is the status bar in the bottom line of the word sense annotating interface, and there clearly show the annotating status: the total word occurrences, the serial number of the current processing instance and the number of the already annotated instances.

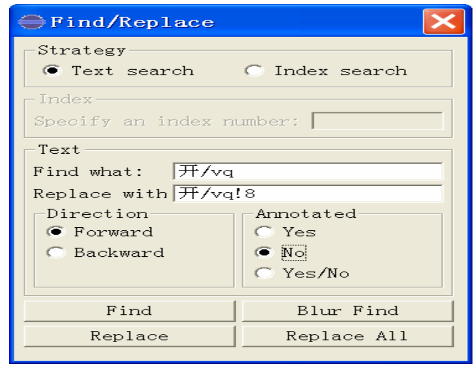

Fig.4 The tool of Find/Replace

3.2 Linking between CSD and CCD

The feature-based description of word meanings in CSD describes mainly the syntagmatic information, such as the subcategory frames of verbs, the semantic categories of the head noun of adjectives, but cannot include the paradigmatic relations. WordNet is a popular open resource and has been widely experimented in WSD researches. Chinese Concept Dictionary (CCD) is a WordNet-like Chinese lexicon (Liu et al., 2002), which carries the main relations defined in WordNet and can be seen as a bilingual concept lexicon with the parallel Chinese-English concepts to be simultaneously included. So the linking between the sense entries in CSD and the synsets in CCD is tried to establish in this project. After the linking has been established, the paradigmatic relations (such as hypernym / hyponym, meronym / holonym) expressed in CCD can map automatically to the sense entry in CSD. What's more, the many existing WSD approaches based on WordNet can be trained and tested on the Chinese sense tagged corpus.

In the right section of the word sense annotating interface there displays the synset information from CCD. When coping with a specific ambiguous word (such as "开/open") in CSD, the linking between CSD and CCD is automatically established with the word itself ("开/open") as the primary key. And then all the synsets of the word ("开/open") in CCD, along with the hypernyms of each sense (expressed by the first word in a synset), are displayed in the right section. A synset selection window (namely Set synsets) containing the offset numbers of the synsets then appears in the right section. The annotator clicks on the appropriate box(es) before the corresponding offset number and then the offset number is automatically added to the feature "CCD" in the currently selected sense entry in CSD.

The linking is now done manually. Unfortunately some of the ambiguous words existing in CSD are not included in CCD. This also provides a good way to improve the coverage and quality of CCD.

4 Consistency Checking

Consistency is always an important concern for hand-annotated corpus, and is even critical for the sense tagged corpus due to the subtle meanings to handle. A software tool namely Sense Consistency Checker is developed in the checking procedure. The checker extracts all the instances of a specific ambiguous word into a checking file with the format of the sense concordances (as shown in figure 5). The checking file enables the checker to have a closer examination of how the senses are used and distributed, and to form a general view of how the sense distinctions are made. The inter-annotator in-agreement thus can be reached quickly and correctly. As illustrated in figure 5, it is obviously an error to assign the same semantic tag to "开/drive 倒车/car" and "会议/meeting 开/held". Simply as it is the checker greatly accelerates the checking speed and improve the consistency.

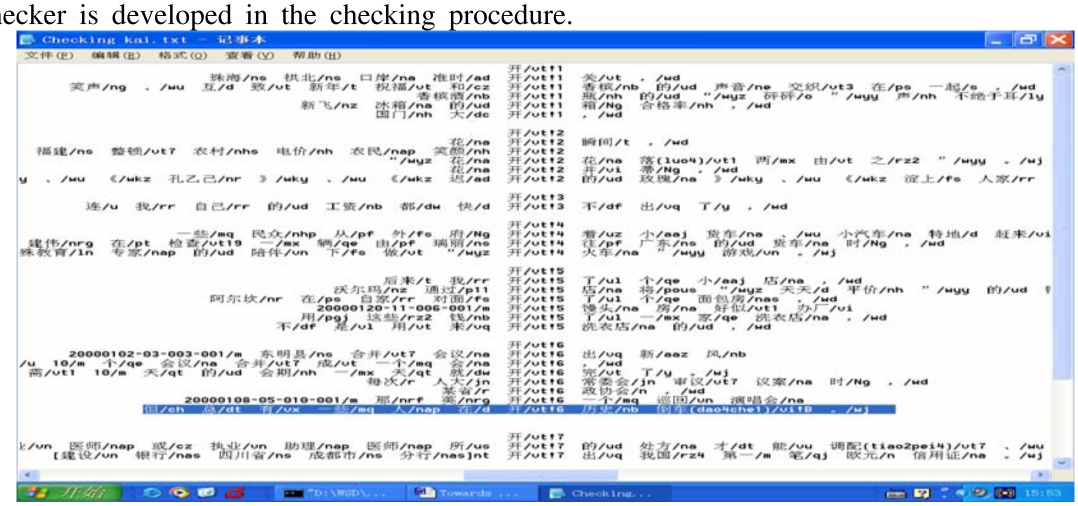

Fig. 5. Some example sentences in the checking file of "开/open"

Together five researchers took part in the annotation, of which three are majored in linguistics and two are majored in computational linguistics. In this project the annotators are also checkers, who check other annotators' work. A text generally is first tagged by one annotator and then verified by two checkers.

After the preprocessing of word segmentation and Pos tagging, the word sense annotating and the consistency checking, the Chinese word sense annotated corpus is constructed. And then other software tools are needed to do further processing in the sense annotated corpus.

5 XML format transforming

The original format of the Chinese sense annotated corpus is in text file as shown in figure 6. In the text file the sign following "/" denotes the POS tag, and the number following "!" indicates the sense ID. The text file complies with the other language resources at the Institute of Computational Linguistics, Peking University, which provides a quite easy way to make full use of the existing resources and techniques at ICL/PKU when constructing the sense annotated corpus.

At the same time in order to exchange and share information easily with other language resources in the world, a software tool namely Text-to-XML Transformer is developed to change the text to XML format (as shown in figure 7). In the XML file, the item "pos" denotes the POS tag of the word, and the item "senseid" denotes sense ID of the ambiguous word.

Thus there are two kinds of format for the Chinese sense annotated corpus, each of which has its advantages and can be adopted to meet different requirements in different situations.

严格/a 的/u 管理/vn 使/vt!2 整个/b 企业/n 像/p 一/m 架/q!1 各/r2 部位/n 零件/n 咬合/vi 得/u 十分/d 紧密/a 的/u 机器/n ，/w 生产/vn 成本/n 逐年/d 下降/vt 。/w 去年/t 电解铝/n 每/r 吨/q 制造/vn 成本/n 已/d 降/vt!3 到 /v 9000/m 多/m 元/q 。/w

Fig. 6. The sense annotated corpus in text file

```
<head date="20000201" page="01" articleno="003" passageno="019">
<passage>
严格的管理使整个企业像一架各部位零件咬合得十分紧密的机器，生产成本逐年下降。去年电解铝每吨制造成本
已降到 9000 多元
</passage>
<postagging>
<word id="0" pos="a" senseid="">
<token>严格</token>
</word>
<word id="1" pos="u" senseid="">
<token>的</token>
</word>
<word id="2" pos="vn" senseid="">
<token>管理</token>
</word>
<word id="3" pos="vt" senseid="2">
<token>使</token>
</word>
……  ……
```

Fig. 7. The sense annotated corpus in XML format

6 Sense frequency calculating

Word sense frequency distribution in the real texts is a vital kind of information both for the algorithms of word sense disambiguation and for the research on lexical semantics. In the postprocessing stage a software tool namely Sense Frequency Counter is developed to make statistics on the sense frequency distribution. Quite valuable information can be acquired through the counter based on the sense annotated corpus: 1) The amount of all the instances of an ambiguous word; 2) The number of the already annotated instances; 3) The occurrence of each sense of an ambiguous word and 4) The sense frequency. Table 1 illustrates the sense frequency distribution of ambiguous verb “开/open” in 10 day’s People’s Daily.

7 Conclusions

This paper describes the overall building procedure of a Chinese sense annotated corpus. The corpus is firstly word segmented and POS tagging using Peking University’s tagger in the preprocessing stage. Then the lexicon Chinese Semantic Dictionary (CSD) containing sense descriptions and the corpus Chinese Senses Pool (CSP) annotated with senses are built interactively, simultaneously and dynamically using the word sense annotating interface. At the same time the linking between the sense entries in CSD and the synsets in Chinese Concept Dictionary (CCD) are manually established. And then the Sense Consistency Checker is used to keep the inter-annotator agreement. Finally two software tools are developed to do further processing based on the sense annotated corpus. A software tool namely Text-to-XML Transformer is developed to change the text to XML format, and the Sense Frequency Counter is developed to make statistics on the sense frequency distribution. The annotation schemes and all the software tools have been experimented in building the SemEval-2007 task 5 “Multilingual Chinese-English Lexical Sample Task”, and have proven to be effective.

Table 1 the sense frequency distribution of ambiguous verb “开/open”

Ambiguous verbs	Sense ID	Occurrences	Frequency(%)
开	8	30	32.26
开	4	13	13.98
开	6	12	12.90
开	7	8	8.60
开	0	6	6.45
开	1	6	6.45
开	9	4	4.30
开	12	4	4.30
开	11	3	3.23
开	2	3	3.23
开	10	3	3.23
开	14	1	1.08
开	15	0	0.00
开	3	0	0.00
开	5	0	0.00
开	13	0	0.00

Acknowledgments. This research is supported by Humanity and Social Science Research Project of China State Education Ministry (No. 06JC740001) and National Basic Research Program of China (No. 2004CB318102).

References

Huang, Ch. R and Chen, K. J. 1992. A Chinese Corpus for Linguistics Research. In Proceedings of COLING-1992.

Huang, Ch. R., Chen, Ch. L., Weng C. X. and Chen. K. J. 2004. The Sinica Sense Management System: Design and Implementation. In Recent advancement in Chinese lexical semantics.

Landes, S., Leacock, C. and Tengi, R. 1999. Building Semantic Concordances. In Christiane Fellbaum (Ed.) WordNet: an Electronic Lexical Database. MIT Press, Cambridge.

Li, J. 1999. The research on Chinese word sense disambiguation. Doctoral dissertation in computer science department of Tsinghua University.

Liu, Y., Yu, S. W. and Yu, J.S. 2002. Building a Bilingual WordNet-like Lexicon: the New Approach and Algorithms. In Proceedings of COLING 2002.

Tanaka, T., Bond F. and Fujita, S. 2006. The Hinoki Sensebank----A large-scale word sense tagged corpus of Japanese. In Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006.

Veronis, J. 2003. Sense Tagging: Does It Make Sense? In Wilson et al. (Eds). Corpus Linguistics by the Rule: a Festschrift for Geoffrey Leech.

Xue, N., Chiou, F. D. and Palmer, M. 2002. Building a Large-Scale Annotated Chinese Corpus. In Proceedings of COLING 2002.

Yu, S. W., Duan, H. M., Zhu, X. F., Swen, B. and Chang, B. B. 2003. Specification for Corpus Processing at Peking University: Word Segmentation, POS tagging and Phonetic Notation. Journal of Chinese Language and Computing.