A Dynamic POS-Tagging-Learned Approach of Chinese WSD

1.Introduction

1.1 What is word sense disambiguation (WSD)

Word sense tagging is essentially a word sense disambiguation (WSD) process. Its main task is to assign a proper word sense code for every word in the input sentences according to word sense tagging set and the context. This sense coding can take four forms: 1) The corresponding code of a word in a semantic domain dictionary. 2) The sequence number of an explanation text for a word in an ordinary dictionary. 3) The corresponding target word of a word in a translation dictionary. 4) The concept definition item in a concept dictionary (e.g. the concept definition in How-Net). Word sense tagging includes two processes which are tagging assignment and tagging disambiguation. Tagging assignment is to assign a set of possible or potential word sense tagging to every word. Tagging disambiguation is to choose a proper word sense that satisfies the context.

Word sense tagging is considered to be the biggest unsolved issue at word level in natural language processing. Although word sense tagging is similar to POS tagging, it is more abstract and difficult because of its focus on the aggregation relation of word meanings. As early as 1950s, people started to work on the research of WSD. A large variety of methodologies have been put forwarded. However, at present stage, most of the current systems can handle the WSD for only a very limited number of multi-sense words. There is still long way to go for WSD to reach the present level of POS tagging. In Chinese NLP research area, WSD has been a rarely touched topic and relatively little effort has been paid to this important issue. This paper is going to propose a technically feasible approach to improve the WSD of Chinese based on the relative maturity of POS tagging system.

1.2 What is the research problem

In order to make the computer have the same word sense discrimination ability as human, we must store in computer grammars and semantic and pragmatic rules as well. WSD methodologies can be grouped into four categories in terms of the methodologies used to acquire knowledge.

1)WSD based on dictionary knowledge. Machine readable dictionaries and semantic domain dictionaries provide sufficient knowledge on word usages and word senses and therefore can be used as a resource for WSD. Machine readable dictionary has become the major knowledge resource for WSD since 80s. Typical machine readable dictionary-based methodology for WSD works this way: It first computes the coverage of each word sense for ambiguous words and the senses of words in the context and then it chooses the sense with maximum overlapping as the current word sense.

Unfortunately, the correctness rate for this methodology is only 50% to 70%. There are several reasons for this comparatively low accuracy: Firstly, traditional machine readable dictionary-based methodology does not make full use of the useful information of the phrases and examples in the dictionary. Secondly, the word sense definition sentences in machine readable dictionaries are normally very short, and it is hence difficult to calculate the overlapping of those two kinds of definitions. In many cases, there are zero overlapping between the word sense definitions of the ambiguous word and context words. Thirdly, in practice, the inevitable combination explosion also greatly restricts the application of this method. Finally. the dictionary is for human use and not for machine use. In addition, there is also inconsistence in the dictionary itself , all these raise difficulty for knowledge retrieval.

Knowledge in semantic domain dictionary is also widely used for WSD. Unlike the methodology used by machine readable dictionary, semantic domain dictionary organizes the words into hierarchy levels in terms of word senses, providing connections between words. WordNet (Miller 1990) and "Roget's Thesaurus" are the most famous English semantic domain dictionaries; "Chinese Thesaurus"(J.Mei,1996) and "Knowledge Net" (Z.Dong, 1999) are most commonly used Chinese semantic resources.

2) WSD based on rules: This methodology depends on the language knowledge of language experts. It constructs rule sets, analyzing the ambiguous words and their contexts, and chooses the word senses that satisfy the restriction rules. Generally speaking, the rules describe the elements that are licensed to modify the ambiguous words and those disallowed to modify the ambiguous words.

3) WSD based on language corpus: The advent of this methodology marks a new age in NLP and gradually becomes indispensable for WSD research. It can be categorized into supervised and unsupervised WSD. The former collects the WSD knowledge from the processed and tagged material; The latter collects WSD knowledge from the raw and tagged material . The essence of the corpus-based WSD is to determine different meanings of words in a context by automatic or semi-automatic learning of the corpus. It can be either statistics-based or example-based. The former collects statistical context proof that support the specific sense of an ambiguous word in a certain context, and the proof is used for WSD of the new input sentences. This method usually collects statistical data of collocation between words and word senses. Example-based method uses statistics almost without turning to any language structural knowledge. What it needs is large quantity of regular parallel sentences. It's easy to implement but hard to obtain a large number of parallel sentences.

4) WSD based on combination of different methods: After several years' efforts, more and more researchers now tend to combine multiple methods to realize WSD. The method combines a variety of methods to get better WSD performance. The combination of knowledge resources will extend the potentially useful knowledge for

WSD, such as dictionary information, collocation information, domain knowledge, syntactic restriction and all of the other potential heuristic information . Based on the word sense system in "Chinese Thesaurus", C.Huang and J.Li adopted the unsupervised learning method to process large scale corpus and constructed the "categorizing machine" for WSD use.

1.3 Who might be interested in a proposal to improve WSD performance

The WSD is itself an intermediate process and hence an indispensable middle layer for most NLP tasks. Any improvement of WSD would interest people employed in the following fields:

1.3.1 Machine translation
The WSD in machine translation has its own characteristics in that it uses target word to distinguish different word senses. How to find the target word in machine translation is an important issue and how the WSD is solved directly affects the quality of the translated text. For example, in the English-Chinese translation system, the word "interest" has two target words in Chinese "□□" and "□□". When translating this word, the system should use the context to decide which to choose in a specific context.

1.3.2 Information search and information processing
In information search and processing, due to the multiple meanings of a word, the system sometimes would come up with the texts that containing the same word with different meaning. For example, someone may use "□□" as key word if he is searching for some reference related to a file. If the search is solely based on these characters, then it will give you all the papers containing this word most of which are related to "manufacturing material" and do not actually mean "files". Therefore, the word sense is important in information search.

1.3.3 Sentence analysis
The word sense also plays a big role when analyzing the grammatical structure of sentences. Grammatical ambiguity is common in many languages and solution to the problem is to introduce word sense. For example, "□□□□□□□□(visiting library's hall)" and "□□□□□□□□(the people who are visiting library)" share the POS sequence of "verb+noun+de+noun", but they have very different sentence structures.

1.3.4 Natural language understanding
When defining the semantic structure of a sentence, we have to consider the word sense of every word in the sentence. The semantic structure of a sentence (e.g Case structure) can be obtained only when the word sense of every word in the sentence is known.

1.3.5 Speech recognition and pronunciation-character transformation
The word-based N-yuan model takes into consideration only the continuous relations between words, thus there exists sentences with no meaning connections between

words in the recognition result. With the introduction of word sense, the continuous relations on the word sense level are provided; therefore it avoids this kind of error mentioned above in some sense.

In summary, as an important process in NLP, the WSD research has great significance both theoretically and practically. The result of WSD research can be directly applied in many aspects of NLP.

2. An Overview of Chinese word sense tagging system and ambiguity distribution of Chinese words

2.1 "Chinese Thesauras" as a popular word sense dictionary

Like POS tagging, word sense tagging also need a tagging system which can provide a description of the word sense categorization principles. Since the dictionary is the basis of word sense definition, a perfect dictionary should be the precondition of WSD. "Chinese Thesaurus" is the only machine-readable semantic dictionary in current Chinese information processing. The compiler of "Chinese Thesaurus" defines the word categorizing rules based on characteristics of usages of Chinese words. Technically, it adopts the coding system of "The Longman Lexicon of Contemporary English". "Thesaurus" divides word senses into large, medium, and small categories. There are 12 large categories, 94 medium categories and 1428 small categories. Uppercase letters represent large categories; Lowercase letters and numbers represent the medium and small categories respectively. The twelve large categories include: people(A), object(B), time and space(C), abstract object(D), characteristics(E), action(F), psychological activity(G), activity(H), phenomenon and state(I), connection(J), auxillaries(K) and honorific(L). It describe a word sense categorizing system from up to down, from broad concept to concrete word senses. The organization of the word sense coding system is as follows:

<word sense coding>::=<large category><medium category><small category>
<large category>::=<uppercase English letter>
<medium category>::=<lowercase English letter>
<small category>::=<digit><digit>

In the small category, the words are further grouped into smaller word clusters synonyms. Every word cluster has a title word. There are totally 3925 title words altogether and word clusters are further represented by a two digit number. For example, the coding of the word "□□" is "Ga15". It is represented in "Chinese Thesaurus" as:

Ga15 □□ □□

□□ □□ □□ □□ □□ □ □□□□ □□□□ .....

□□ □□□ □□ □□

There are two word clusters in "Ga15". One is composed of the words representing "conscious" and the other is composed of words representing "know something". Therefore, the further word sense coding for "□□" is "Ga1501" which can be called sub small category coding. Multi-sense words are grouped into different word clusters in terms of their word senses. For example, "□□" has three word senses in "Thesaurus": (1) stuff that can be made into product. (2) stuff provided for the content of a work or for references. (3) a person that has talent to do something. The corresponding coding for each word senses are "Ba06","Dk17" and "A103". In this article, the word sense coding in "Thesaurus" is directly used to represent word sense. Multi-sense word has multiple word sense codings and the length of the coding is 4.

2.2 Chinese Word Ambiguity Degree and Distribution

"Chinese Thesaurus" has a vocabulary of more than 50,000 words. The following table shows the distribution of multi-sense words in Chinese:

Table1. Multi-sense word distribution in "Chinese Thesaurus"

	One word sense	Two word senses	Three word senses	More than three word senses	Total	Ambiguity percentage
One character word	1973	833	397	571	3774	48.0%
Two character word	28154	3837	572	118	32681	16.0%
Three character word	12597	999	96	6	13698	9.0%
Total	42724	5669	1065	695	50154	14.8%

The statistics above shows that "Thesaurus" has 42724 mono-sense words and 7370 multi-sense words. Besides that, it also shows that 14.8% of the words are multi-sense words. We note also that shorter words have higher chance to be ambiguous and also tend to be more ambiguous.

According to the word sense and ambiguity type of the word sense Chinese word can be categorized into mono-sense word, category ambiguous word, non-category ambiguous word and blending ambiguous word. Among these types, the WSD of the non-category ambiguous word is most difficult because it needs more contexts. The research on the distribution of these different types helps to adopt different tagging strategies to different words. In order to have a better understanding of how the words distribute with respect to word senses, we collect the statistical data on static distribution of word sense based on "Chinese Thesaurus". The results are in table2 as follows:

Table2 word distribution with respect to word sense in "Chinese Thesaurus"

Sample	TagSet	Word distribution(%)	Average number of senses

		Mono-sense word	Multi-sense word	
Chinese Thesaurus	Large category	87.69	12.31	2.30
	Medium category	83.47	16.53	2.48
	Small category	81.46	18.54	2.57

Word sense distribution is in close relation with the technical principles of a dictionary and specific word sense category system. The data in Table 2 and Table 3 may not perfectly reflect the relations among POS, word sense and word distribution, we can still see clearly the following general trends: Firstly, in Chinese, most words are mono-sense words, with a dominating percentage of 81.46-87.69%. Secondly, multi-sense words, although occupying only around a quarter of the whole vocabulary, are used much more frequently than mono-sense words. More commonly used words usually have a higher degree of ambiguity.[1] Thirdly, multi-sense words are not evenly distributed with respect to their POS. Thirdly, the granularity of word sense categorizing has great effects on WSD. The smaller the granularity, the higher percentage of the multi-sense words, and also the bigger average word senses. On the other hand, the smaller the granularity, the higher percentage of non-category multi-sense words and the lower percentage of category multi-sense words. There is little effect on the percentage of blending multi-sense words though. In a word, word sense ambiguity becomes more serious if we adopt a comparatively smaller granularity when define the word sense tagging system. Finally, one conspicuous characteristic of Chinese word is that its POS is morphologically unmarked. And there is always the extreme claim that Chinese words do not even have POS, it only gets POS when it comes into a specific sentence("□□□□□□□□□") . This makes WSD more important and complicated for Chinese words than for English words. And this same fact leads us to believe that the approach we are proposing in this article should work much better for Chinese than for English.

3. Our proposal

Compared with WSD in English, the research of WSD in Chinese has been drawn much less attention. The method we put forward here is based on the characteristic of Chinese words, we see that POS information is relatively not closely related to lexicon items in dictionary, but once recognized in a given context, brings much information which will be useful for WSD. Our basic idea is that given the high accuracy of present POS tagging system, we are going to count on the output of POS tagging of a text we are processing. And we are going to let the POS information take over the WSD task in the first place. Given the characteristics of Chinese words we

[1] This trend can be clearly seen if we have a look at the word distribution in a dynamic corpus. For example, the PFR corpus of "People's Daily"(Jan,1998) at <http://www.icl.pku.edu.cn/Introduction/corpustagging.htm>

discussed above, we believe it a feasible and effective way of improving the WSD of Chinese text processing. We propose the following steps of this approach.

3.1 Text segmentation and POS tagging

In current Chinese tagging systems, these two steps can be realized at the same time. Among currently existent systems, the one developed by the Institute of Computational Linguistics of Peking University has reached the accuracy of 96.8%. And in our final evaluation part of this paper, we are also using this system.

3.2 Word sense label assignment base on "Chinese Thesaurus"

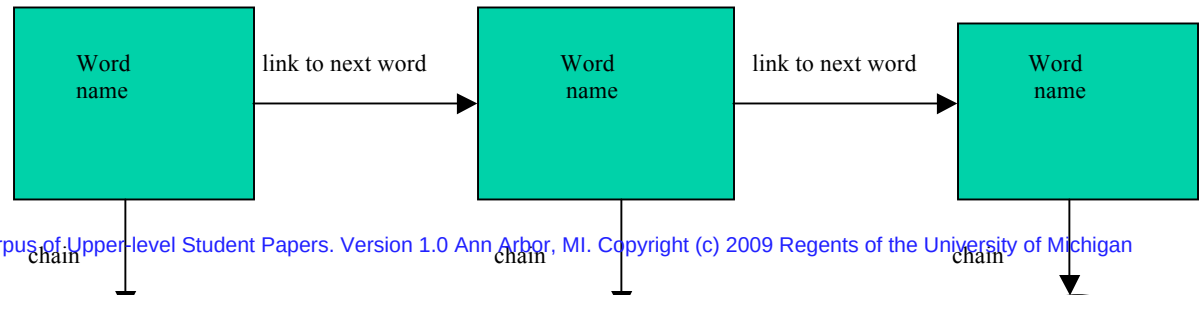
The label assignment of word sense is to assign a set of possible or potential word senses to each word. Since we have tagged the POS for the input sentences before word sense tagging. Therefore, the word sense label assigning process can be described as follows: Firstly, to input the word sequence with POS tagging as $WP = w_1 / p_1 w_2 / p_2 wi / pi \cdots w_m / p_m$ (w is Chinese word, p is its POS in this specific context). Secondly, to assign a set of potential word sense codings for each word $w_i (1 <= i <= m)$ in the input according to "Chinese Thesaurus". Hence we get a tagging cluster: $WPS = *w_1 / p_1 / s_1 * w_2 / p_2 / s_2 * wi / pi / si \cdots * w_m / p_m / s_m$ as output.

In this paper, the process of automatic label assignment is actually the process of link search and dictionary search. To facilitate the search, we put all of the words searched in the dictionary together with their word sense codings in a link. The structure of the link is as follows:

```
Struct word{ // structure used to store the word sense codings
    Char cla[10];// char array to store the word sense codings
    Struct word *chain; // pointing to the next node
};

struct node{ //structure used to store words
  char nam[20]; //to store Chinese words
  struct node * link; // pointing to the next node
  struct word * next;//pointing to the word sense node of the word
};
```

We can further illustrate this process as follows:



The technical process of this word sense label assignment is made up of three steps:

1) Read in a Chinese word
2) Search in the link, if information about that word is found there, then use the information do the tagging and get the output. When this is over, go back to the first step. If not found, then go to step 3.
3) Search the word sense information in the machine readable dictionary "Thesaurus", if found, then assign all of its word senses at the output. If not found, set the word sense tagging to null and then go to step 1.

3.2 WSD based on the dynamic corresponding information between POS and word senses

We mentioned earlier that the twelve large upper-case letters represent twelve semantic fields that word senses may fall into. “Thesaurus” does not group words under the name of POS, but by investigating into “Thesaurus” and comparing it with “Modern Chinese Dictionary”, we can see the correspondence between word sense and POS as shown in Table 3:

Table 3 The correspondence between word sense and POS

词性	词义标记	词性	词义标记
名词n	A,B,D(Dn04-05,Dn08-10除外)	方位词f	Cb
动词v	F, G, H, I, J, L	处所词s	Cb
形容词a	E	副词d	Ka
状态词z	E	介词p	Kb
区别词b	E	连词c	Kc
量词q	Dn08, Dn09, Dn10	助词u	Kd01, Kd02
数词m	Dn04, Dn05	语气词y	Kd03, Kd04
代词r	A, C, E	叹词e	Ke
时间词t	Ca	拟声词o	Kf

We can see that there are strong relation between word sense tagging and POS tagging . For example, A, B and D in “thesaurus” are mostly noun. This correspondence relation is very useful for WSD improvement.

Given the correspondence relation between the POS and word sense, the last step of your method is: Cross out all of the word senses that are not included in the correspondence list of the specific tagged POS *pi* of that word. And for some words, if the results are null, then just keep the original word sense tagging in order to guarantee the reliability of the final accuracy of the WSD. By this final step, we are descreasing the average sensed of the words in the text to a noticeable extent and hence improve the WSD performance.

4. Evaluation of this approach

We are expecting that the application of our approach at the first step of WSD will decrease the average sense numbers of the words in a given text. For the sake of time, we tested our approach manually on one news article from “bbs.mit.edu”. If the ambiguities are well decreased by our approach for this article, we have good reasons to believe that this approach will improve Chinese WSD to a considerable extent. And the advantage of this approach is that it is easy to apply and can be freely combined with any other WSD algorithms. We segmented and tagged the news article using segmentation and tagging system developed by the Institute of Computational Linguistics of Peking University, flowing is the output of the tagging:

中国/ns 领事馆/n 放宽/v 来自/v 中国/ns 大陆/n 新/d 移民/v 的/u 护照/n 申请/v 资格/n 的/u 谣言/n ，/w 连日来/l 在/p 纽约/ns 华人/n 社区/n 传/v 得/u 沸沸扬扬/i ，/w 以讹传讹/i ，/w 连续/a 几/m 天/q 都/d 有/v 上千/m 人/n 新/d 移民/v 涌/v 向/p 位于/v 曼哈顿/ns 的/u 中国/ns 驻/v 纽约/ns 总/v 领/v 馆/Ng ，/w 在/p 中/f 领/v 馆/Ng 的/u 签证/n 组/n 门前/s 大/a 排长/n 龙/n 。/w 中/f 领/v 馆/Ng 官员/n 对外/v 声明/v ，/w 目前/t 中/f 领/v 馆/Ng 对/p 纽约/ns 地区/n 新/d 移民/v 申请/v 中国/ns 护照/n 的/u 需要/v 进行/v 最/d 新/a 的/u 意见/n 调查/v ，/w 听取/v 纽约/ns 地区/n 侨/Ng 团/v 及/c

侨胞/n 的/u 有关/p 护照/n 侨/Ng 团/v 的/u 反映/v 。/w 希望/v 作为/p 中国/ns 护照/n 申请/v 有关/vn 部门/n 的/u 研究/v 参考/v ，/w 但/d 从未/d 改变/v 护照/n 申请/v 发放/v 政策/n ，/w 希望/v 侨胞/n 们/k 不/d 要/v 轻信/v 谣言/n 。/w

The following is the word sense list of the 118 words in the article according to "Chinese Thesaurus", those with "*" after them indicate that it is crossed out after the WSD:

□□ Da19
□□ Gb14
□ Ed28* Ed23* Ee19* Ee38* Ie01 Ka18
□ Ag04* Dk10* Ed28* Gb04 Gc03 Hi25
Jc05 Kc01* Kc08*
□□ Ad01
□____
□□ Df08* Gb04
□□ Di09
□□ Hc07
□□ Hc15
□□_____
□□ Ih02
□□_____
□ Ka07 Kc03*
□□ Hg12
□□ Gb01 Hg14
□ Bo29* Ed01* Kd01
□□ Di09 Dm01
□□ Je01* Kb04
□□ Hc15
□□ ____
□□ Hc15
□□ Di02 Ja01(both of these two codings are not included in the list of POS, and we keep these two)
□□ Df08* Gb04
□□ Hc15 Ja03
□ Bo29* Ed01* Kd01
□ Bb04* Br09* Di09* Di10* Dn08* Ea13* Fa34
□ Ad01
□□ ____
□□ Je01* Kb04
□ Bo29* Ed01* Kd01
□□ Ad01
□ Je12* Kb02* Kc01
□ Ad01
□ Bb04* Br09* Di09* Di10* Dn08* Ea13* Fa34
□□ Cb08 (this coding is not included in the list of POS, and we keep it)
□□____
□□ Fc05

□□ Hc18
□□ Df14
□ Bo29* Ed01* Kd01
□ Eb22* Eb28* Ka12
□ Ka02
□□ Ig03
□□ Df07* Jc05
□ Bo29* Ed01* Kd01
□□ ____
□□ Hc15
□□ Hc15
□□ ____
□ Eb22* Eb28* Ka12
□□ Cb08
□□ ____
□ Dk17* Dn08* Eb02* Ed12* Fa26* Hc18* Hi08* Hi18* Jc02* Kb01 Kb04 Kb07
□ Bk05* Bq04* Dn08* Hf04 Hi27 Je14
□ Dm05
□ Aj14* Ca06* Ca08* Cb04 Cb05 Da05* Ea03* Ed06* Ed47* Ie11* Je13*
□□ Ca10
□□ Dk14* Hi13
□□ _____
□□ ____
□ Dm05
□ Bk05* Bq04* Dn08* Hf04 Hi27 Je14
□ Aj14* Ca06* Ca08* Cb04 Cb05 Da05* Ea03* Ed06* Ed47* Ie11* Je13*
□ Dd15 Dh03 Ed57*
□□ Ae10
□ Ah04* Ah05* Bf02* Dj05* Dn04* Ea03 Eb04 Ec05 Ed26 Ed38 Ka01*
□□ ____
□ Di09 Di10
□□ Hc16 (this coding is not included in the list of POS, and we keep it)
□ Bo29* Ed01* Kd01
□ Dm05
□ Bk05* Bq04* Dn08* Hf04 Hi27 Je14
□ Aj14* Ca06* Ca08* Cb04 Cb05 Da05* Ea03* Ed06* Ed47* Ie11* Je13*

□ Hj19* Ib03* Ja05* Jd01* Jd02* Ka12* Kb01
□ Dm05
□ Bk05* Bq04* Dn08* Hf04 Hi27 Je14
□ Ae10* Ed56* Ka11* Ka15* Ka29*(all of these codings are not

included in the list of POS, and we keep them all)
□□ ____
□ Hb04
□□ Di02
□ Bo29* Ed01* Kd01
□□□ ____
□□ Jd02
□ Cb01* Ka10* Kb01*(all of these codings are not
included in the list of POS, and we keep them all)
□ ____
□□ ____
□ Eb22* Eb28* Ka12
□ Aa01 Ab02 Dd17 De01 Dn03
□□ _____
□ Eb02* Ed61* Jd01 Jd04 Jd07
□ Cb25* Di03* Ka07 Ka12 Ka28
□ Ca23* Cb07* Da24* Dh01* Ed51* Ed57* Bp26* Dn05
Ka27
□ Bp26* Dn05 Ka27*
□□ Ig03* Ka11*(both of these two codings are not included in the list of POS,
and we keep these two)
□□□□ Ie01
□□□□ Ef03
□ Gc02* Gc03* Ie14* Jc05* Je12* Ka15* Kd01
□ Dk26* Hg01 Hi15 Ie01 Je03
□□ ____
□□ ____
□□ ____
□ Hj19* Ib03* Ja05* Jd01* Jd02* Ka12* Kb01
□□□ ____
□□ Da19
□ Bo29* Ed01* Kd01
□□ Dd16
□□ Hc15
□□ ____
□ Bo29* Ed01* Kd01
□□ ____
□ Eb22* Eb28* Ka12
□□ Be01
□□ Di02
□□ Ja05
□□ ____
□□□ _____
□□ Di02

We can see that before this dynamic POS-tagging-learned WSD, there are 118 words in the text with 225 large semantic categories represented by the 12 upper-case letters, 266 medium categories and 296 small categories. Before WSD, there are 37 mono-big-sense words, 34 mono-medium sense words and 34 mono-small-sense words . After WSD, there are 166 large semantic categories, 191 medium categories and 207 small categories. And after WSD, there are 63 mono-large-sense words, 57 mono-medium-sense words and 62 mono-small-sense words. We can see from the following table that the average sense number is greatly decreased because of the WSD.

Table 4 Average Word Sense Before and After WSD

		Word distribution(%)		
		Mono-sense word	Multi-sense word	Average number of senses
Before	Large	31.3	68.7	1.9
	Medium	20.3	79.7	2.2
	Small	20.3	79.7	2.5
After	large	43.3	56.7	1.3
	Medium	48.3	51.7	1.5
	Small	52.5	47.5	1.7

Bibliography:

1.Ji Donghong□Huang Changning□Sense Tagging of <TongYiCiCiLin> Using Itself□Laboratoryof Intelligent Technology and Systems, Tsinghua University
2.D.J. Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys and Louisa Sadler Machine Translation: an Introductory Guide, Blackwells-NCC, London, 1994, ISBN: 1855542-17x. http://clwww.essex.ac.uk/MTbook/, chapter9
3.Martin Volk. 1997. Probing the lexicon in evaluating commercial MT systems. In Proc. of ACL/EACL Joint Conference, pages 112--119, Madrid. EAGLES Evaluation of Natural Language Processing Systems, http://www.ilc.pi.cnr.it/EAGLES96/browse.html#wg3
4.Nancy Ide, Jean Veronis, Introduction to the Special Issue on Word Sense Disambiguation: The State of Art.
5. □□□□<<□□□□□>>□□□□□□□□□1983

6□□□□□□□一个人机互助的汉语语料库多级加工处理系统，北大计算语言所学术研讨会论文