*Journal of Chinese Language and Computing, 15 (3)*:(149-160)

# Some Suggestions on How to Improve the Lexical Semantic Knowledge-Base

Su Xinchun [1] Wang Hui [2] Lu Weiqing [1] Qin Shaokang [1]
1.Xiamen University，suxch@jingxian.xmu.edu.cn
2.National University of Singapore，chswh@nus.edu.sg

**Abstract**

Disambiguation, particularly that of lexical meanings is the key problem involved in natural language processing (NLP), and among many means of achieving this end, is to construct a language knowledge base. Nowadays, the knowledge bases established tend to be more and more advanced and fine-grained, with the function of providing the lexical information improved a lot. This paper tries to argue for the importance of constructing the lexical semantic knowledge base, as a means of increasing the amount of information stored in and the practical value of the language knowledge base. The higher aims of the language knowledge base are to identify accurately the lexical meanings in the context, to calculate the sense frequencies displayed in the text and to improve the function of updating the language information stored in the knowledge base. The issue addressed in this paper is an attempt at natural language disambiguation, which, as an essential problem, has a wide range of applications.

**Keyword:** lexical meanings, language knowledge base, disambiguation, sense frequencies

## 1. Introduction：the current tendencies in the knowledge-base technology

1.1 The two types of lexical knowledge-base

The initial efforts to establish the language knowledge base began more than a decade ago, since many scholars realized its essential important role in NLP (Zhu and Yu, 1996). Wang, Zhan and Yu (2003) gives a brief summary of the recent developments in this area: “Since the mid 1980s, many countries began to invest to develop the machine semantic dictionaries, for example, Wordnet (Fellbaum, 1998) by U.S.A., Mindnet (Richardson, 1998), Framenet (Fillmore, 1998), the conceptual dictionary (EDR) by Japan, SenseWeb by Singapore. China has also developed some semantic dictionaries, for example, Chinese Semantic Dictionary for Information Processing(905) (Chen and Yuan, 1995), the Machine Dictionary of Modern Chinese Verbs, Hownet (Dong, 1999), Chinese Conceptual Dictionary (CCD) (Yu and Yu, 2002) and so on. Besides, many scholars have also attempted to extract the lexical semantic information from the machine dictionary (Chodorow, 1985; Ide, 1993; Huang, 1998).”

There are two classical knowledge bases which are designed to process Chinese lexical meanings, and this technology is the essential step in the knowledge base construction. These are: Dictionary of Modern Chinese Grammatical Information (abbreviated as GD) and Hownet. These two products represent two different descriptive approaches to the language knowledge base: while the former presents a static, individual and uni-dimensional description of the grammatical information, the latter focuses on such aspects as semantic components, semantic characteristics, semantic relevance and semantic network. Based on the methodology employed in the former one, the discussion in this paper will urge the establishment of a more advanced Chinese lexical knowledge base which will provide more lexical information in a more effective way.

## 2. How to make the language knowledge base more fine-grained

### 2.1 Word segging

When we try to establish the language knowledge base, the problem of word division is a fundamental one, because its aim is to segment the language chunk into a group of appropriate lexical units according to the syntactical structures and basic and frequently used word senses. In the computer-aided processing of natural languages, all the words which share the same orthographical system are clustered together, so in this process, some lexical phenomena with no clear lexical boundaries, for example, *erhua*, *qingsheng*, and verb phrases, and so on, are ignored or not dealt with properly. Because the character-based word division, in fact, is a great step backward and is short of accuracy and appropriateness scholars have managed to make much improvement in this aspect.

2.2 Part of speech (POS) tagging

To assign the appropriate part of speech is the first step towards an accurate description of Chinese lexicon, since some weaknesses involved in the character-based word division can be overcome. And its another advantage is to distinguish the senses of the polysemous word. For example, 比 (bi3) has two entries in *Modern Chinese Dictionary* (abbreviated as MCD):

| | |
|---|---|
| 比 | ⑴比较；较量：～干劲丨学先进，～先进。⑵能够相比：近邻～亲丨坚～金石丨演讲不～自言自语。⑶比画：连说带～。⑷〈方〉对着；向着：别拿枪～着人，小心走火。⑸仿照：～着葫芦画瓢（比喻模仿着做事）。⑹比方；比喻：打～丨人们常把聪明的人～做诸葛亮。⑺比较两个同类数量之间的倍数关系，叫做他们的比，其中一数是另一数的几倍或几分之几：这里的小麦年产量和水稻年产量约为一与四之～。⑻表示比赛双方得分的对比：甲队以二～一胜乙队。⑼介词，用来比较性状和程度的差别：今天的风～昨天更加大了丨许多同志都～我强。【注意】A）“一”加量词在“比”的前后重复，可以表示程度的累进：人民的生活一年～一年富裕了。B）比较高下的时候用“比”，表示异同时候用“跟”或“同”。 |
| 比 | 〈书〉⑴紧靠；挨着：～肩丨林次栉～。⑵依附；勾结：朋～为奸。⑶近来：～来。⑷等到：～及。 |

In GD, however, there are five senses attached to this word.

| **POS** | **Homograph** | **Sense** | **Example** |
|---|---|---|---|
| Verb | 1 | 比较 | ~谁的力气大 |
| Verb | 2 | 比画 | ~着葫芦画瓢 |
| Verb | 3 | 数量对比 | 二~一 |
| Preposition | | 比 | 你去~他去更合适 |
| Noun | | 比数 | 3 与 4 之~ |

A comparison between these two dictionaries can show that in MCD, there are as many as 13 senses of 比 (bi3), only five of which are listed in GD, with 8 senses omitted.

| **MCD** | **GD** | | | |
|---|---|---|---|---|
| **Sense** | **POS** | **Homograph** | **Sense** | **Example** |
| (1)比较；较量 | Verb | 1 | 比较 | ~谁的力气大 |
| (3)比画： | Verb | 2 | 比画 | ~着葫芦画瓢 |
| (7)比较两个同类数量之间的倍数关系 | Verb | 3 | 数量对比 | 二~一 |
| (8)表示比赛双方得分的对比 | Noun | | 比数 | 3 与 4 之~ |
| (9)介词 | Preposition | | 比 | 你去~他去更合适 |

There are several possible reasons for the absence of some senses in GD.

Some senses are omitted deliberately, for example, ⑵能够相比 and ⑹比方；比喻 of 比 1.

Some senses in MCD are integrated into the listed senses in GD as is in the case wherethe two senses of 比 1, ⑶比画 and ⑸仿照 are merged.

And some discrepancies between the two are due to the morphological or dialectal reasons. For example, (4)对着；向着 of 比 1and such senses of 比 2 as (1)紧靠；挨着, (2)依附；勾结, (3)近来 and (4)等到. In fact, it does not matter much whether these morphological or dialectal senses are listed in the dictionary or not.

Form the above examples, we can see that the grammatical function based POS assignment has several advantages: to identify the lexical sense in the context, to reduce some redundant entries in MCD, and to make a finer distinction between different senses. There are, however, some weaknesses, that is, although it can identify the senses of the polysemous words, it can not distinguish the senses that belong to different POSs. For example, the two senses of 比 1 as the verb are not listed in GD, partly due to the deliberate omission, partly because the distinction between these two senses are too fine to be spelled out.

The merger of (3)比画 and (5)仿照 of 比 1, however, are not appropriate in fact. 比画 is usually used as a verb, but in 连说带比, 连……带 is a connective construction, which connects two verbs. It seems that 仿照 is a verb, but in the phrase, 比着葫芦画瓢, 比 is a preposition.

The POS assignment, which is based on the grammatical properties displayed by the word in its colligational environment, still falls within the traditional theoretical confines. This method should be complemented with the more sense-oriented method which

considers collocational sense relations among a cluster of words. In a word, the sense relations, instead of the grammatical relations, among the words, should be the focus of the POS assignment method.

2.3 Semantic category assignment

Word division and POS assignment are the initial steps in Chinese lexicon processing, which will pave the way for the subsequent work. The distinction made between words solely according to POS is only a very rough one, which applies only to those words or lexical meanings which possess the similar grammatical functions. So the grammatical property based distinction of lexical meanings should be complemented with the meaning based approach. Realizing this problem, the authors of GD have begun to establish Modern Chinese Semantic Dictionary. Wang, Zhan and Yu (2003) observe: "The semantic knowledge base should be based on the grammatical knowledge base. In this regard, we should conduct a systematical semantic classification and a comprehensive description of the information realized in the semantic collocations. The aim of our efforts is to improve the dynamic research and summary of the semantic combination and to establish an appropriate descriptive framework for the semantic information, which is closely related to the language engineering. " With feasibility as its primary aim, this semantic knowledge base is based on GD and conducts a semantic classification of the content words. Under the classes of nouns, verbs, adjectives and adverbs, there are 23 first degree classes, and 41 second degree classes, 37 third degree classes, 29 fourth degree classes and 19 fifth degree classes. A caveat should be added that the five hierarchical classification is not necessarily applicable to all the words to be investigated. For some nouns, the distinction is much finer, and the hierarchical chain is much longer such as that of concrete things---non-living things---man-made things---tools---transport facilities. There are some cases where this hierarchical chain lasts only one layer, for example, the verbs which indicate the psychological movements, static relations, those adjectives which indicate time value, event value and seven types of adverbs, three types of numericals and so on. In these cases, the short hierarchical chain is enough to distinguish the sense relations.

With the semantic information incorporated, the ability of the lexical processing is enhanced considerably, because only grammar-based method is not applicable to the cases where the senses of a particular word belong to the same POS, and so much finer semantic categories must be introduced. (Wang, 2004)The author has also conducted an investigation into 3989 polysemous nouns, and finds that solely grammar-based method can distinguish only 23% of the senses, with the majority of cases unsolved where the senses of a particular

polysemous words belong to different POSs.

The aim of the semantic dictionary is to use the semantic information to address the problems which are left in GD which employs the POS-based method. Take the word 把握 (ba3wo4) as an example. It has three senses:

(1)握；拿(to hold)：司机～着方向盘。

(2)抓住(抽象的东西) (to grasp)：～时机丨透过现象，～本质。

(3)成功的可靠性(多用于‘有’或‘没’后) (probability)：球赛获胜是有～的。

(1) and (2) are verbs, which can take objects. There are two methods to distinguish these two senses. The first one is to see their collocational behaviors respectively. For example, the objects that 把握 (to hold) can take include: gun, spear and so on while the objects that 把握 (to grasp) can take include: activity, characteristics, situation and so on. This method is very accurate but can not exhaust all the possible cases where the word is used. Another method is to employ the semantic information and sense relations displayed by them: 把握 (to hold) usually take concrete objects and 把握 (to grasp) abstract objects.

Since method of constructing the semantic knowledge base is based on the most essential sense relations between words, it can tackle the fluid variations in the contextualized usage of the word. For example, the basic sense of 吃(chi1) is to eat, which is followed by the nouns indicating food, but sometimes, this word can be followed by other elements, for example, the place where eating takes place, the manner of eating, or something associated with the food.

### 2.4 The sense-based method

In order to achieve the best result in language processing, the sense-based method is the more effective one among the two possible means. In this way, the basic analytical unit is lexical meaning, instead of the lexis. Many excellent dictionaries have followed this method. For example, *Dictionary of Synonymous Words* incorporates more than 64,000 lexical items, with more than 7,000 arranged according to their senses, among which, 打 (da3) has as many as more than 20 senses. Many scholars have realized the advantage involved in this method, and have tried to get this method improved.

Huang (2004) observes: “There are two important elements in the lexical net. The first one is the sense-based lexical group or synset and the second the sense relation connecting the word sets. The semantic network is formed on the basis of the connectors of synonymous words, which are connected according to the sense relations.” And subsequently, he proposes several principles in dealing with the semantic relations among

words: one sense for one item, one thing for one sense, one event for one sense, to make the sense independent of the context, to correlate the sense to the context. Yu (2004) proposes to establish a comprehensive language knowledge bank, which, considered as a general treatment of Chinese lexical meanings, is based on the chain relation between words, POSs and homographs. Yu, furthermore, proposes a sense-based comprehensive language knowledge bank, and this method is a more detailed framework since it forms a chain relation between word, POSs, homographs and senses, by taking the lexical meaning into consideration.

Since the second half of 2004, the project team of Chinese Sense Frequency Bank at Xiamen University and Department of Chinese Studies of National University of Singapore have cooperated to develop the Chinese Lexical Semantic Knowledge Base Based on Grammatical Description, which is abbreviated as XHK. Based on the previous corpus and all the lexical items in MCD, XHK conducts a comprehensive assignment of the grammatical and semantic properties to the lexical items in the knowledge base. And in this process, the essential step is to present a detailed description and a finer distinction of the semantic knowledge and the relevant details of this step will be introduced in *XHK: Grammar-based Semantic Descriptions of the Chinese Lexicon* by Wang Hui.

**3. Why to describe and distinguish lexical meanings and the difficulties involved**

We can observe that implied in the course of Chinese lexical processing, from word division, POS assignment, semantic categorization to sense identification is a tendency in which the consideration is becoming more and more detailed and oriented to the lexical meanings. And there has been a scholarly consensus that the knowledge base should be established on the description of the lexical senses and their relations. Besides, there have emerged many new problems which need to be reinvestigated and some new areas left to be improved. And the ultimate aim is to obtain a more comprehensive portrayal of the lexical senses and the lexical components. After the analytical unit has been shifted from the word to the lexical meaning, a set of problems underneath the previous framework will emerge:

( 1). The relationship between the semantic categories is so rigid that a variety of the possible sense relations are concealed.

(2). The criteria in the sense classification are so homogeneous that some other sense categories than objective sense, basic sense, main sense are ignored.

(3). Although the static sense elements are accorded due attention, some variations along the dimensions of style, register, context are ignored.

(4). The comparison of the synonymous senses is absent in XHK. The comparison of

the lexical meanings can be conducted within different domains, for example, within that of the senses of a particular polysemous words, within that of the synonymous words, or within that of the hypernymycial words.

Moreover, the distinction between the associative meanings should also be made among the words, which will contribute to the finer understanding of the nuance of the different lexical meanings. The present understanding of the lexical senses are based on the comprehensive understanding of all the senses of the words.

In *Dictionary of Synonymous Words* edited by the author, more than 800 groups of synonymous words are distinguished mainly in terms of the basic sense, the collocational pattern and the associative meaning.

The following are three examples, in which three groups of synonymous words are distinguished along the above three dimensions.

颤抖 cha4ndo3u　　发抖 fa1do3u

(Both mean to tremble.)

[辨析] <动> 都指连续、无规律的抖动。【颤抖】可指人或事物，抖动的程度较轻，常用于书面语，如：他的心在～丨那棵落光了叶子的小柳树在风中～着丨老人用～的声音向路人乞求施舍。【发抖】多指身体的抖动，抖动的程度较大，带口语色彩。如；妈妈气得全身直～丨孩子一会儿发高烧，一会冷得～丨小狗水淋淋地爬上岸，冻得全身～。

In this example, the meanings of these two words are compared in terms both of the basic meaning and of the style. The degree of the act of trembling indicated by 颤抖 is much less intense than that indicated by 发抖. Besides, the former is less formal than the latter.

爱护 a4ihu4　爱惜 a4ixi1　　珍惜 zhe1nxi1

(All mean to treasure something.)

[辨析]〈动〉都有不使受损害的意思。【爱护】语意重在保护，使用范围较广，如：～眼睛丨～动物丨～公共设施。【爱惜】强调因重视而不糟蹋、不浪费，对象多指易消耗的具体事物，如：～粮食丨～身体丨～时间/光阴。也可是友谊、名誉、人才等比较抽象的事物，如：他是个～人才的好厂长。【珍惜】强调因宝贵而珍重与爱惜，语意较重，且有书面语色彩，多用于表示特别有价值或具有特殊意义的事物。如：那双布鞋是母亲亲手做的，他非常～，一直舍不得穿丨要像～生命那样～自己的尊严。

In this example, the three words are compared in terms both of the basic meaning and of the stylistic meaning. In terms of the amount of attention accorded, 珍惜 is more intense than the other two. And 珍惜 tends to be used in more formal contexts.

美观 me3igua1n　　漂亮 pia4olia4ng

(Both mean “very beautiful”.)

［辨析］<形>都有好看的意思。【美观】多指服饰、用具、建筑物、工具等的外在形式好看，多用于书面语，不可重叠。如：整套房子装修得既简朴又～｜这件衣服穿上显得～大方。【漂亮】使用范围比较广，可用来形容人的容貌、风景、交通工具等，多用于口语，可重叠。如：海边的风景实在～｜这个～的姑娘是你女儿吗？｜“六一”节那天每个孩子都打扮得漂漂亮亮。还有引申义，表示好、精彩、出色的意思。如：这件事干得～｜这场球踢得真～｜我军又打了一个～战。

In this example, the two words are compared in such aspects as the collocational pattern, the style and the orthographical variation. While the former can not modify as many things as the latter does, the former is more formal than the latter. The orthographical form of the latter can be extended to 漂漂亮亮(pia4opia4olia4nglia4ng).

The above three examples illustrate the fact that there are many factors contributing to the difference in the senses used in the specific context, for example, the basic meaning, the collocational pattern, the register, and the associative meanings which include the affective meaning, style, origin, the times and the context in which the word is usually used. So the distinction between the lexical meanings should be made along these numerous dimensions, and it is because of the complexity of the factors involved in the sense distinction that we need to establish an efficient and accurate lexical knowledge base that can make an accurate identification of the different senses of the word.

Several difficulties must be solved in the construction of a lexical semantic knowledge base which incorporates the basic senses and the associative meanings of the word:

(1) What lexical components should and are able to be stored into the knowledge base?
(2) How to make the various lexical senses formalized and symbolized?
(3) How to identify the target lexical aspect to be compared?
(4) How to define the effective domain in which the lexical meanings can be compared?

**4. Some suggestions on how to describe the lexical knowledge**

The potential problems listed above will be addressed in the process of designing the present knowledge base. The author has envisaged several suggestions on how to solve these problems:

(1) Priority should be given to the various aspects of the lexical meanings.

There exist a cluster of multi-dimensional criteria along which the lexical senses can be distinguished. And which criterion should be incorporated and may be formalized with success into the knowledge base is a very important question to be considered, since this question will influence the way that the lexical meanings will be compared. In fact, it is very difficult to get the lexical meanings formalized with symbols because the senses and characteristics of the word will vary a great deal from context to context. In contrast, the associative meanings of the word are much easier to distinguish since the comparison is casually based on such dichotomies as foreign versus native, dialectal versus standard, archaic versus currently used, more intense versus less intense, colloquial versus written, newly coined versus borrowed or loaned, general versus register-specific, complimentary versus derogatory or neutral, and so on. If a generalized method could be designed to distinguish all the lexical meanings in *Dictionary of Synonymous Words*, we would obtain a much better understanding of the differences and method involved in the distinction. The objective meaning should be distinguished with reference to the semantic classification.

(2). How to identify the relativity of the lexical meanings and to use the semantic differential scales to distinguish the lexical components

When we try to identify and distinguish the lexical components, it is important to get the connotative meanings formalized with the symbols. It is equally difficult to indicate the minute differences in the lexical meanings. Here we can employ the semantic differential scales to establish a continuum between the two lexical meanings so that there exists a transitory state set by the upper bound and the lower bound. The method of imagining a lexical continuum to identify the lexical difference will be much more effective if it is based on the lexical groups or categories.

(3).How to enhance the independence of the lexical meaning in order to get the lexical description more simplified

We should establish an independent data base so that a word can be identified with reference both to the lexical orthography and to the lexical meaning. In this respect, the method used in MCD to deal with the homograph is a very effective example. When MCD was being compiled, the analytical linguistic unit was shifted from character to word. Due to some reasons, for example, the extensive use of the word as the analytical unit, solely synchronic but not diachronic consideration, and too much attention given to the phonological and morphological differences, the number of the homographs is surprisingly large. (Su, 2000a, 2000b) Such a method, however, is not used very frequently in synonymous dictionaries. The revised edition of MCD, which is under way now, will make a lot of changes, by dividing the lexical items according only to the sense relations among words as stated in Notes on the Revised Edition of *Modern Chinese Dictionary* (2005). One

possible solution is to get the lexical components tagged with numbers after their meanings are identified. This practice, which is not due to the limitations of the computer technologies, will save a lot of human efforts so that the main attention will be spent on the description of the lexical components.

(4).How to establish the polysemy-synonymy-homography three-dimensional system in order to determine the effective domain where the comparison of the lexical meanings can be made since it is necessary to determine the effective space before comparing the lexical meanings.

A. The polysemous dimensional

The different lexical meanings of a particular word are often compared along this dimension, for example, the more concrete meaning of 把 versus its more abstract meaning. This dichotomy between the abstract and the concrete should be generalized to a higher inclusive one so that many other dichotomies, for example, far versus near, dormant versus dominant and wide versus narrow, and so on, can be incorporated into this more general dichotomous category. And this part is the most tricky one when we try to establish the reference system to compare the polysemous words.

B.The synonymous dimension

It is a very familiar practice to compare the synonymous words, for example, the minute difference between 颤抖 and 发抖. There exist as many as more than ten thousand synonymous dictionaries, among which *Dictionary of Synonymous Words* is the most important one, with its value lying in the systematical semantic classification. Along the hierarchical system in the thematic categories, there are 12 first degree ones, 94 second degree ones， 1428 third degree ones, 3927 fourth degree ones. In fact, there still exist about 11,000 fifth degree categories, which fall into the synonymous domain. For example,

First degree category: A/human

Second degree category: Aa/general refernce

Third degree category: Aa01/ human, the people, many people

Fourth degree category: human

Fifth degree category: hand, staff, population, man, mouth, index figure, etc.

Some dichotomies, for example, standard versus dialectal, modern versus archaic, written versus colloquial, frequent versus infrequent can be observed among the seven words in the fifth category. And in fact, the distinction among complimentary, neutral and derogatory is also applicable to this category.

C. The hypernymical dimension

Generally, those dictionaries which include the semantic classifications deal with the hypernymical words which occupy different positions along the hierarchy of the same

semantic category. In the current dictionaries, there are usually several hundreds of semantic categories, with each one including dozens of or hundreds of words. Take *Dictionary of Synonymous Words* as an example, on average, there are 676 words in the second degree category, 45 ones in the third degree, 16 ones in the fourth degree. It is, however, a wise policy not to subsume too many words into each category and to make the number of the included words vary from category to category. There are still some weaknesses with the current synonymous dictionaries, since the number of the words included in each category varies only a little from case to case, from five or six thousand to ten thousand.

**References**

1. 朱学锋、俞士汶。1996。自然语言处理与语言知识库，见罗振声，袁毓林主编，刊《计算机时代的汉语汉字研究》，清华大学出版社
2. 王惠，詹卫东，俞士汶。2003。《现代汉语语义词典规范》。 Journal of Chinese Language and Computing (Singapore). Vol 13, No.2. pp 159-176
3. 王惠，2004。《现代汉语名词词义组合分析》，北京大学出版社。第 220 页。
4. 黄居仁，2004。《中文的意义与词义》，台北 • 南港，中央研究院资讯科学研究所中文词知识库小组，中央研究院语言学研究所筹备处。
5. 俞士汶、段慧明、朱学锋、张化瑞. 2004。《综合型语言知识库的建设与利用》，《中文信息学报》，第 18 卷第 5 期。
6. 苏新春，2000A，《同形词与“词”的意义范围——析〈现代汉语词典〉的同形词词目》，《辞书研究》，第 5 期。
7. 苏新春，2000B，《同形词与多义词的区分及其对词典编纂的影响》，刊《世纪之交的应用语言学》，北京广播学院出版社。