International Journal of Knowledge and Language Processing
Volume **2**, Number 4, October **2011**

www.ijklp.org
KLP International ©2011 ISSN 2191-2734
pp. **39-45**

New Word Recognition Based on Word Formation

Xu Yan-hua
School of Chinese Language and Literature
Ludong University
Yantai, Shandong, 264025, China
Xu Yanhua @gmail.com

Received September 2011; revised October 2011

ABSTRACT. *The detection of new words, non-proper nouns in particular is still a bottle-neck in Chinese word segmentation. This thesis analyzes the patterns and rules of new word formation patterns and rules and applies the rules in the detection of new words. Experiments show that the approach is able to achieve a precision of 96.8%.*

Keywords: new word, word formation, automatic Recognition

1. **Introduction.** With the rapid development in society and the increasingly wide use of Internet, new words are constantly formed. Therefore it has become a problem to identify and sort out these new words. This problem is not only the focus of lexicography, but also an important task of Chinese information processing. In the field of lexicography, the work of compiling dictionaries of Chinese new words is basically done artificially at present time. The selection of the new words has been become the "bottleneck" problem, which is difficult to solve. “To select new words, editors usually have to read a large number of texts, and have to rely on their own linguistic knowledge, language intuition to identify a new word. Consequently, the problems of labour, time, inconsistency in knowledge and so on are unavoidably. These will definitely lead to the situation that the work of dictionary compilation of Chinese new words will be lagging behind conspicuously the change of society.”[1] From this point, there’s urgent need to apply computers in the field to find new worlds automatically. In the field of Chinese language processing, because automatic word segmentation is the basic step in natural language processing, the improvement of precision and recall ratio of dividing words has important influence on natural language processing. Although studies in the Chinese words segmentation have achieved great success after more than ten years’ study, there are still dissatisfactory places in recognizing the unknown words. Therefore, how to solve the problem of unknown words is important for Chinese language processing.

1. The Recognition methods. Unknown words are difficult in the field of Chinese information processing. From the angle of computational linguistics, unknown words refer

to the words that don't appear in word list and are therefore unknown by computers. They can be divided into two kinds: proper nouns and non-proper nouns. The former includes names of persons, names of places, names of organizations and so on, while the latter includes new words, abbreviations, dialectal words, learned words, industry words, expressions in Hong Kong and Taiwan etc. At present the research about unknown words identification is mainly concentrated on proper nouns. However, because of the variety in word-formation rules, the studies on the recognition of non-proper nouns are still not in-depth. But in the real texts, the unknown words of non-proper nouns account for a considerable proportion. For automatic word segmentation, it is impossible to improve word segmentation accuracy through infinite expansion of word list because unknown words are actually infinite in essence. With the emergence of new things, new words are constantly emerging. Therefore, the key to improve word segmentation accuracy is enable the word segmentation system to recognize unknown words by rules and patterns.

Currently, there are many methods proposed for identifying unknown words: Sun(1995) identify Chinese names of persons automatically through the construction of database; Chen(1999) proposes a whole solution is to recognize the unknown words among fragments by calculating the in-to-word rate of a single word; Song(1993) identifies personal names using elicitation information in contexts. Generally speaking, the existing methods are often focused on recognition of proper nouns, paying not much attention to new words which have emerged in large quantity recently.

In order to solve the problem of unknown word identification, a database of 15719 new words is built, with information of structural patterns, from which the detailed information of word building can be obtained. On this basis, we count and generalize various types of the new words in detail, then we retrieved the rules of new words' building. We hope that the computer could make full use of these rules to identify the new words effectively. These rules are summed up on the basis of the large-scale corpus, so they have credibility and representativeness. We believe that such achievements will not only provide a foundation for the computer to identify unknown words, but also play an important role in advancing the development of Chinese information processing.

2. Word formation patterns and rules of two-syllable new words. In modern Chinese, the compound pattern of "root + root" plays a major part in the whole vocabulary, mainly because of the fact that Chinese has no morphological changes and relies heavily on combining nouns, verbs, adjective morphemes to make different types of new words. The most common forms of the type are "N+ V", "V + N", "N+A", "A+N", "N+ N", "V+V", "V + A", and "A + A" .From the above description, we can see that no matter in which way the morphemes constitute verbs, nouns or adjectives, these nine "class order" takes up a large proportion.

The above statistics also show that not all part of speeches take part in the new word formation, Nouns, Verbs, Adjectives takes more than 99% of the new words.

(1) As to the rules of word formation of nouns, we can explore them from two aspects. From a general view, the most important way to form a new double-syllable noun is to

combine a modifier and noun morpheme. In the newly formed words, nouns of head-modifier structure take the major part, which is followed by nouns of coordination. In terms of internal structure, Ng+Ng, Ag+Ng, Vg+Ng are the major types, which covers 56.7%,24.8%, and 14.6% respectively, with a sum of 96.1%. We can tell the main characteristics of word formation of double-syllable noun from the rules of word formation: noun morphemes are involved in word formation. In most cases, one noun is formed by two noun morphemes, in the pattern Ng+Ng, which has 3378 examples and account for 56.7%. Another characteristic is that one noun is formed by two morphemes. In most cases, the second morpheme is a noun morpheme, with very few exceptions. One such exception is of the pattern Ng+Vg, such as "果冻", which reflects the changes of attributes of morphemes in word formation.

（2）As to the word-formation rules of verbs, we can also investigate from two aspects: Adv-O verbs, VO verbs, and combined verbs form the main body of the verbs. Combined verb is the most used, which accounted for 27.9%; Adv-O takes the second place with a percentage of 21.5%; the third one is VO verbs which accounted for 21.1%. These three types together cover 70.5% of the verbs in the whole. As far as the internal components is concerned, Vg + Vg type class sequence and Vg + Ng type class sequence are in the majority, the mainly formation characteristics of two-syllable compound verbs is the participation of verb morphemes in word formation. At least one verb morpheme participate in the two morphemes that constitute a verb accounted for 59%, most of them are especially verbs comes first, accounting for 81.1%. that is the combination structure of Vg + Vg type, Adv-O structure in Vg + Ng type and Vg+Ng in VO structure.

（3）From a general view, the head-modifier adjectives and adjectives of coordination are the main types of adjectives. The most frequent is coordination adjectives, followed by VO adjectives; In regard of the internal structures, the pattern Ng+Ng has 531 instances in total, which takes 92.8%. The main characteristics of word formation of the two-syllable words is that adjectives morphemes participate in the word formation, the most frequent one is the one that adjectives morphemes comes first, the class order that without the adjectives morphemes mainly is Vg + Ng type, and Ng + Vg type.

TABLE 1. Statistical Results 1

Component 1	**Component 2**	**Lexical Structure**	**Lexical Category**	**Statistical Results**
Vg	Vg	Z（45.7%）	V（1838）	V 99.5% N 0.5%
		L（31.9%）	V（1285）	
			N（21）	
		B（16.2%）	V（667）	
		C（5.2%）	V（209）	

3. Rules for recognition of new bi-syllabic words. The constructional types above, the ordered morphemes of single construction are decided by descriptive assertion. We statistically sampled and classified those intercrossed morphemes under the consideration of relative difficulty of computational identifying on intercrossed morphemes, the results are given Table 1 to Table 8:

According to the statistical results above, the rule can be given below: Vg+Vg→V, the accuracy reaches up to 99.5%.

TABLE 2. Statistical Results 2

Component 1	**Component 2**	**Lexical Structure**	**Lexical Category**	**Statistical Results**
Vg	Ng	B（73.7%）	V（2428）	V 73.2%
			A（18）	N 26.3%
		D（26.3%）	N（869）	A 0.5%

According to the statistical results above, the derived rule can be: Vg+Ng. If the priority is given to verb-object type, the accuracy is 73.2%; Otherwise the attribute-head type is considered, the accuracy is 26.3%.

TABLE 3. Statistical Results 3

Component 1	**Component 2**	**Lexical Structure**	**Lexical Category**	**Statistical Results**
Ag	Vg	Z（99.6%）	V（743）	V 95.5%
			N（20）	N 2.6%
			A（11）	A 1.9%
		C（0.4%）	A（4）	

According to the statistical results above, we concluded the rule: Ag+Vg→V，the accuracy is 95.5%; Ag+Vg→N, the accuracy is 2.6%, Ag+Vg→A，the accuracy is 1.9%.

TABLE 4. Statistical Results 4

Component 1	**Component 2**	**Lexical Structure**	**Lexical Category**	**Statistical Results**
Vg	Ag	B（63.6%）	V（211）	V 98.8%
			A（2）	A 1.2%
		C（35.8%）	V（120）	
		W（0.6%）	A（2）	

According to the statistical results above, we concluded the rule goes: Vg+Ag→V, the accuracy is 98.8% ; Vg+Ag→A, the accuracy is 1.2%.

TABLE 5. Statistical Results 5

Component 1	**Component 2**	**Lexical Structure**	**Lexical Category**	**Statistical Results**
Ng	Vg	Z（63. 1%）	V（229）	V 95.3% N 4.1% A 0.6%
		W（36.9%）	V（117）	
			N（15）	
			A（2）	

According to the statistical results above, we concluded the rule goes: Ng+Vg→V， the accuracy is 95.3% ; Ng+Vg→N， the accuracy is 4.1% ; Ng+Vg→A， the accuracy is 0.6%.

TABLE 6. Statistical Results 6

Component 1	**Component 2**	**Lexical Structure**	**Lexical Category**	**Statistical Results**
Ng	Ng	D（95%）	N（3378）	N 100%
		L（5%）	N（171）	

According to the statistical results above, we concluded the rule goes: Ng+Ng→N， the accuracy reaches up to 100%.

TABLE 7. Statistical Results 7

Component 1	Component 2	Lexical Structure	Lexical Category	Statistical Results
R	Vg	W（95.7%）	V（43）	V 97.9% A 2.1%
			A（1）	
		Z（4.3%）	V（3）	

According to the statistical results above, we concluded the rule goes: R+Vg→V， the accuracy is 97.9% ; R+Vg→A， the accuracy is 2.1%.

TABLE 8. Statistical Results 8

Component 1	Component 2	Lexical Structure	Lexical Category	Statistical Results
Tg	Vg	Z（81.3%）	V（13）	V 100%
		D（18.7%）	V（3）	

According to the statistical results above, we concluded the rule goes:Tg +Vg→V， the accuracy is 100%.

4. The result of the experiment. We take advantage of the rules of word constitution, which are summarized from the foregoing paragraphs, to detect the specific linguistic materials of People's Daily from April 1st to 7th in 1998. From this detection, we can identify 1627 new words, among them, there are 1038 words composed of the attributive central language structure, accounting for 63.8%; the words that consist of the center adverbial phrase structure are 106, accounting for 6.52%; the words made up of V-O constructions are 295, accounting for 18.13%; the words that composed of the structures of predication are 7, accounting for 0.43%; there are 101 words that are combined structures, accounting for 6.52%; 3 are added structure words, accounting for 0.19%; besides, there are 10 prefix words and 15 suffix words, 0.61% share of the total, respectively, and 0.92%.

From the result of this detection we can see that: (1) the patterns summarized in the previous paragraphs are able to cover all the new words found in the text; (2) the two ways of word-formation are basically identical. The word formation mode summarized above is effective.

Of course, we are mainly doing analysis from the point of view of static new words. In the future, there is a need to combine the static with the dynamic, in order to provide a better basis for automatic word segmentation.

REFERENCES

[1] Chen Xiaohe. A package of Solutions to Unknown Words in Automatic Word Segmentation [J]. Application of Language, 1999, (3).

[2] Guo Wei et al. The Method of Identifying New Words Based on the Delay Decision and Slope [J]. Journal of Sichuan University (natural science edition), 2007, (3).

[3] Liu JianZhou, Tingting He, Changri Luo. The New Words Automatic Recognition Based on the Network and Corpus [J]. Computer Applications, 2004, (7).

[4] Wang Yuanyuan, Zhongshi He. The Detection Algorithm of Chinese Name Recognition Based on the Part-of-speech Detecting [J]. Computer Science, 2005, (4).

[5] Yuan Chunfa, Changning Huang. The Study of Chinese Morpheme and Word-Formation Morpheme Based on Morpheme Database [J]. Application of Language, 1998, (3).

[6] Zhou Lei, Qiaoming Zhu.The Study of the Unknown Words Identification Method Based on the

Statistical and Rules [J]. Computer Engineering, 2007, (8).

[7] Zhou Lei. The Method of Unknown Words identification Based on the Debris Participle [J]. Journal of Changshu Institute of Technology, 2007, (2).

[8] Sun Maosong et al. The Automatic Identification of Chinese Name [J]. Journal of Chinese Information, 1995, (2).

[9] Song Rou et al. The Recognition Method of Names Based on the Corpus and Rules [M]. Computational Linguistics Research and Application. Press of Beijing Language College, 1993.

Annotations:

[1] Chen YuQuan et al. Computer Aided Dictionary Compilation of New Words[J]. Journal of Shanghai Jiaotong University, 2000, (7).