The Multi-layer Language Knowledge Base of Chinese NLP

Hu Junfeng, Yu Shiwen

The Institute of Computational Linguistics,
Dept. of Computer Science and Technology, Peking University,
Beijing, 100871, P.R. China
hujf@pku.edu.cn , yusw@pku.edu.cn

Abstract

This paper introduced the effort to build a multi-layer knowledge base of Chinese NLP which combined with list-based, rule-based and corpus-based language information. Different kinds of information are designed to solve different kind of problems that encountered in the Chinese NLP. The whole knowledge base is designed with theoretical consistency and can easily be put into practice in the application systems.

1. Introduction

For Chinese, there are no remarkable boundary and also lack of inflection of the words. A lot of efforts have been made to give the automatic word segmentation and POS-tagging and most of the systems claim accuracy of 98% or even above. But until now, the result of word segmentation and POS-tagging failed to meet the requirement of some farther applications such as MT.

One of the reason is that there exits a gray region between Chinese compound words and phrases. For example, '树梢'/treetop (树 / tree, 梢/tip) will probably be taken as a word because the character '梢' in this sense is seldom be used as a word separately. But when you come across '树皮'/bark (树/tree, 皮/skin), there will be some disputes about it because 皮/skin can be used a little bit more separately than '梢'. Until now, there don't have a very strict standard to make difference of the compound words and the phrases in Chinese. In most of the systems, a dictionary is used to make the decision and the standard of the word collection of the dictionary is arbitrary. The strings which have the exactly the same grammatical and semantic behavior will probably be treated differently as word or phrase. This will surely lead to the inconsistency of the system and brings a lot of trouble in grammar and semantic analysis.

Other reason that we cannot count on the result of the POS-tagging system is due to the multiple standards of Chinese POS-tagging. In Chinese, the string '发展' , which means develop, have exactly the same form when it behaves like 'development' in a sentence. If you just tag all the existence of '发展' as 'v'(verb), it will not help too much to the farther analysis of the sentence. If you want to tag the string '发展' differently as 'v' or 'n' (noun), according to its grammatical function in each specific sentence, that will be a grammatical problem rather than a lexical one. It will not only lead to the sharp lower of the accuracy rate but also will lead to a recursive problem of POS-tagging and parsing.

In dealing these problems, we think, a multi-layer language knowledge base which combining with different kind of language information and with theoretical consistency, should be needed.

2. The Grammatical Dictionary of Contemporary Chinese.

As the foundation of this knowledge base, ICL started to build "The grammatical dictionary of contemporary Chinese" in year of 1986. With 13 years efforts, this dictionary has now concluded more than 73 thousand Chinese words described with hundreds of grammatical attributes (120 attributes for verb only).

Table 1 illustrated some attributes tagged in the verb table. The column 'DOUBLE OBJECT' shows the verb can have double object or not. If a verb marked "双" in the 'PLURAL SUBJECT' column, it means the verb can only have plural form subject. The 'REAR NOUN PHRASE' means the verb can combine with a rear noun to form a noun phrase. The detailed information of this dictionary can be found in the book "*Grammatical Knowledge base of contemporary Chinese—a Complete Specification* ", Tshinghua university press, 1998.

3. The Inner Structure of the Word

Most of the Chinese characters embedded certain kind of meaning. Most of the Chinese compound words have inner structure just like phrase. The inner structure of a word influences the grammatical performance and even the meaning of this word. The ICL marked up the structure information of all the compound word in the 'grammatical dictionary'.

Table 2 shows some attributes of the structure information that tagged in the dictionary. The column 'front char', 'rear char' indicates the grammatical attributes of the two characters in the two-character words. The structure column shows the forming style of this word.

By doing this, people can easily find some hyponymy of 树/tree by finding all the word in the dictionary that started with character 树 and with the inner structure n-n-定-n (the first character is 'n', the second is 'n', the structure is '定' and the POS is 'n') (table 3a). With the same structure, if the character 树 is in the rear part of the word, different kind of trees will appear in the list (table 3b).

词语 WORD	双宾 DOUBLE OBJECT	复数主 PLURAL SUBJECT	后名 REAR NOUN PHRASE	很 'VERY'	着了过 TENSE MARK	重叠 OVER-LAP	离合 SEPARA-BLE	兼类 ANOTHER POS
保存 save			可		着了过	ABAB		
生存 survive			可		着了			
发展 develop			可		着了			
告诉 tell	双				了过			
协商 discuss		复	可		了			
支持 support				很	着了过	ABAB		
冒险 risk			可		过	VVO	离	a
去 get rid of					了过	vv		

Table 1: some attributes tagged in the verb table

字词 word	前字 front char	后字 rear char	构词 structure	词类 pos	层次 layer
海洋/ocean	n	n	联	n	
摆脱/get rid of	v	v	补	v	
人人/every people	n	n	重	n	
暗室/secret room	a	n	定	n	
海带/kelp	n	n	定	n	
剪刀/scissor	v	n	述	n	
黄山/huang mountain	a	n	地	n	
槟榔/betel nut	x	x	单	n	
调味品/flavoring			定	n	2

Table 2 the structure information of some words

字词	前字	后字	构词	词类
树梢	n	n	定	n
树墩	n	n	定	n
树枝	n	n	定	n
树干	n	n	定	n
树根	n	n	定	n
树冠	n	n	定	n
树胶	n	n	定	n
树敌	v	n	述	v
树立	v	v	联	v

Table 3a some word started with '树'.

字词	前字	后字	构词	词类
梨树	n	n	定	n
枫树	n	n	定	n
桦树	n	n	定	n
柏树	n	n	定	n
桉树	n	n	定	n
柳树	n	n	定	n
植树	v	n	述	v
建树	v	v	联	n
枯树	a	n	定	n

Table 3b some word end with '树'

4. Word Formation Rules and the Unlisted Word Discovery

4.1 Word Formation Templates

Since we have marked the inner structure of the words in the dictionary, in the other viewpoint, the inner structure of all these words have made up a word structure template set. Table4 shows some most frequent used structure template of Chinese two-char word. The templates that have been used more than one time in the dictionary are defined as **generative template**. The total number of the generative template that we have got is 382 (only for two-char words).

To go one step farther, we assume that there are two empty slots in each generative template, one is for front char and the other is for rear char. Farther more, if a character appeared in one of these templates in our dictionary, we say this character can form compound word by using this specific template. Then come two formulas to describe the probability of each character to enter these templates.

The probability of character X enter template Si (front position)

$$Pf(X \mid Si) = \frac{\text{Front occur } (X \mid Si)}{\text{Front occur } (X)} \qquad ①$$

The probability of character X enter template Si (rear position)

$$Pr(X \mid Si) = \frac{\text{Rear occur } (X \mid Si)}{\text{Rear occur } (X)} \qquad ②$$

Using these two formulas, we got a character-based list which described the frequency of each character entering different kind of word formation template.

Table5a described the ability of some characters to enter the templates in the front position, table5b is for the rear position.

The frequency of each character entering different generative template reflects the ability of this character in generating words with this template.

In this case, the template is called **morphological template** and the percentage of entering the template is called the **intimate degree** of the character with this template. In this paper, the character, the morphology templates and the intimate degree that reflected the relationship between them are defined as **word formation rules** of this Chinese char.

前字 front	后字 rear	构词 struct	词类 pos	频度 times	例词 example
n	n	定	n	9434	案卷
a	n	定	n	3285	暗坝
v	v	联	v	3067	遨游
v	n	述	v	2258	昂首
n	n	联	n	1537	笆篱
v	n	定	n	1498	案语
a	a	联	a	1163	暗淡
v	v	状	v	614	爱抚
n	k	后	n	590	案子

Table4 some frequent used generative template, the column 'times' shows the number of occurrence of this template in the grammatical dictionary.

前位字 front char	前位总数 front occur times	构词模版 morphological template	出现 occur times	结合强度 intimate degree	例词 example word
树	27	n_n_定_n	22	.8148148148	树权
树	27	n_n_联_n	2	.074074074	树木
树	27	v_n_述_v	1	.0370370370	树敌
树	27	v_v_联_v	1	.0370370370	树立
树	27	n_a_定_n	1	.0370370370	树懒
庶	4	a_n_定_n	4	1	庶民
数	16	n_n_定_n	11	.6875	数词
数	16	n_n_联_n	2	.125	数量
数	16	n_k_后_n	1	.0625	数落

Table5a character based word formation rules for front char, totally 11072 lines.

后位字 rear char	后位总数 rear occur times	构词模版 morphological template	出现 occur times	结合强度 intimate degree	例词 example word
树	32	n_n_定_n	25	.78125	桉树
树	32	a_n_定_n	3	.09375	大树
树	32	x_n_定_n	1	.03125	柞树
树	32	v_v_联_n	1	.03125	建树
树	32	v_n_述_v	1	.03125	植树
庶	1	a_a_联_a	1	1	富庶
数	83	n_n_定_n	42	.50602	辈数
数	83	a_n_定_n	10	.120481	常数
数	83	v_n_定_n	10	.120481	变数
数	83	v_n_述_v	6	.072289	报数

Table5b character based word formation rules for rear char, 11664 lines all-together

4.2 Unlisted Word Discovery in the Poems of Song Dynasty

Word formation rules restricted the way of a certain character to generate compound word. For example, the character '树', in front position, can use five kind of morphological template to generate words. The '庶' can only have one. If we make the hypothesis that the word collection of the 'grammatical dictionary' is adequate and rational, we can edict a more exaggerate conclusion that:

The necessary condition for a two- character string 'XY' to be a compound word is that there exists at least one morphological template that the X can enter this template in the front position and the Y can enter it in rear position.

Can these features been applied into the refinement of the result of the unlisted word discovery?

In the research work of the ancient poems of Song dynasty, a statistic modal has been used in finding the word in the corpus (1.6milliom characters). Since there does not have a standard word dictionary for ancient Chinese, all the words in the corpus have to be treated as unlisted words. For the first step, a list of 73,218 candidates (two character string) was created via statistic method, and then, 14,067 words were confirmed manually. In the 73,218 candidates, there are 34,592 strings that do not have valid morphological template; among them 1452 words are found. In these 1452 words, 932 words are proper nouns. That means, if we simply neglected all the strings that do not have a valid morphological template, the accuracy rate of the unlisted

word discovery will go 12.8% up with 10% lose of recall. Considering there often have given method to deal with the unlisted proper noun, morphological template can be used as a good standard to eliminate the statistical garbage in the process of unlisted word discovery. Table 6 shows some words that do not have valid morphological template. The value 'rm', 'dm' refer to the people' name and location.

词语 word	成词度 statistic value	属性 POS
槲叶	3.91	n
绸缪	3.37	v
纷纭	2.7	a
蛾眉	2.4	n
未央	2.33	dm
茅茨	2.3	n
芋魁	2.29	n
貂裘	2.26	n
萧瑟	2.16	a
仲尼	2.04	rm
唐虞	2.01	rm
会稽	1.98	dm

Table6 some words which do not have valid morphological template which appears in the poems of the Song dynasty

5. The Processing of the Large Corpora of People's Daily

Unlike some other scholars, we think most of the Chinese words have unique POS. Only quite a few among them have more than one POS (table2 last column). The reason that many words, like '发展', can play different kinds of grammatical roles in a sentence is because of the lack of inflection in Chinese. But it is important to make difference the different kind of the usage of these words in analyzing a sentence. In processing the *'People's Daily'* of 1998 (around 26,000,000 Chinese characters) , we induced four more tags, i.e. vn, an, vd, ad, which have not be adopted in the Grammatical Dictionary, in dealing with the multi-function phenomena of Chinese words. For example:

经济发展很重要。(The development of economic is very important.)

The verb '发展'(develop), which behaves as a nominal in this sentence, is tagged 'vn' instead of its original form 'v'.

The tagging result of this sentence is like:

经济(economic)/n 发展(develop)/vn 很(very)/d 重要(important)/a

This makes it possible to identify the multi-function of Chinese words according to the statistic information rather than the parsing information. By now, more than 12,000,000 chars of the words of the corpus have been processed.

Table 7 shows the four special tags that has been used in processing the corpus.

标记 tags	名称 name	英文解释 English explanation
ad	副形词	adjective used as adverbial modifier
an	名形词	noun used as an adjective
vd	副动词	verb used as adverbial modifier
vn	名动词	verb used as a noun

Table 7 the four special tags for corpus processing

Some example of the processed corpus is showed below:

新华社/nt 北京/ns 3月/t 9日/t 电/n [中共中央/nt 办公厅/n]nt 近日/t 发出/v 通知/n ，/w 要求/v 各级/r 党委/n 组织/v 干部/n 群众/n 认真/ad 学习/v 悼念/v 邓/nr 小平/nr 同志/n 的/u 重要/a 文献/n 。/w 我们/r 要/v 化/v 悲痛/an 为/v 力量/n ，/w 继承/v 邓/nr 小平/nr 同志/n 的/u 遗志/n ，/w 以/p 更加/d 努力/ad 做/v 好/a 各/r 方面/n 工作/vn 的/u 实际/a 行动/vn ，/w 来/v 表达/v 对/p 邓/nr 小平/nr 同志/n 的/u 爱戴/v 和/c 怀念/v 之/u 情/Ng 。/w

6. Conclusion

For Chinese, most of the compound words have inner structure just like phrase. Marking the inner structure of the words make it possible to give a consistent word level interface for the parser. The automatic extraction of the word formation template gives another rule-based standard for Chinese word and is helpful for unlisted word discovery. The processing of the large-scale corpus makes it possible to get some very important statistic information of Chinese NLP. Morpheme dictionary, grammatical dictionary, character based word formation

templates (22,736 in total), large-scale processed corpus make up a multi-layer knowledge base of Chinese NLP.

As we know, for most of the NLP applications, only the grammatical information is not enough. For corpus processing, segmentation and POS tagging is just the first step. There still have a lot of work needs to be done on the phrase binding, the maximal noun phrase discovery and some other fields. A semantic dictionary concerning with valence dependency and semantic classification is under construction.

7. References

[1] Fu Guohong, Wang Xiaolong. (1999). Unsupervised Word Segmentation and Unknown Word Identification. 5th National Language Processing Pacific Rim Symposium

[2] Swen Bing,Yu Shiwen. (1999). A graded Approach for the Efficient Resolution of Chinese Word Segmentation Ambiguities, 5th National Language Processing Pacific Rim Symposium

[3] 俞士汶等, (1998). 《现代汉语语法信息词典详解》, 清华大学出版社

[4] 俞士汶,朱学锋,李峰. (1999) 现代汉语语素库的开发及应用. 《世界汉语教学》