POS Estimation of Undefined Chinese Words

Yu Jiangsheng Duan Huiming
yujs@icl.pku.edu.cn duenhm@pku.edu.cn
Institute of Computational Linguistics, Peking University
Beijing, P. R. China, 100871

Abstract *We have tested the POS formation rules in our lexicon --- Grammatical Knowledge-base of Contemporary Chinese (GKB), which contains at least 70,000 the most popular Chinese words and have found some stable rules for POS formation. Finally, we will give an explanation to those rules.* [1]

Keywords *Part-Of-Speech (POS), POS formation, pant model, semantic constraint*

1 Introduction

Different from indo-european languages, Chinese words are short of morphological transformations. There seems no way to guess the POS of undefined Chinese words only by the weak morphological rules. Because the POS tagging influences the syntactic-semantic analysis, the POS estimation of those undefined Chinese words becomes more important in Chinese Processing. We have tested the POS formation rules in the lexicon which contains at least 70,000 the most popular Chinese words and have found some stable rules as follows:

1) Adj + Adj = Adj;
2) Noun + Noun = Noun;
3) Adj + Noun = Noun;
4) Noun morpheme + suffix = Noun;
5) Noun + Adj = Noun;
6) Adj +Adj morpheme = Noun;
7) Noun + suffix = Noun

The first three have strongly proved the intuition of linguists with high statistical data. We also tested the lexicon of undefined words from the well done segmented and POS tagged corpus of a half-year newspaper (People's Daily). The large corpus contains at least 14,000,000 Chinese characters. The number of POS taggers we used is 40.

In Section 1, we gave a formal definition of undefined words and show the original idea of dynamic lexicon. In order to reach the statistically best segmentation result, we have to delete a set of words from the lexicon, which are obviously rebuilt by some word formation rules. We showed the requirements of POS estimation in Chinese Processing. In Section 2, some data from the two experiments were mentioned and then the POS formation rules with several examples. In Section 3, we gave a linguistic explanation for those rules, which proved the linguists' strong intuitions on the POS (word) formation of two-character words. We proposed the Pants Model and emphasized the importance of qualitative analysis of linguistic objects. Finally in Section 4, we proposed the fuzzy POS formation for the undefined Chinese two-character words.

[1] This work was supported by National 973 high-tech Project, National Science Foundation, and Peking University 985 Project.

2 Undefined Words

Definition 2.1 Let Σ be a set of Chinese Characters, the set of all the Chinese words, W, is a subset of Σ^*. Given a lexicon $L \subset W$, w is called undefined word if and only if $w \in W$, but $w \notin L$.

Definition 2.2 Characteristic function from L to $\{1,0\}$ is defined by:

$$\chi_L(u) = \begin{cases} 0 & u \notin L \\ 1 & u \in L \end{cases}$$

The most popular Chinese words are formed by two characters, so in this paper we just focus on the POS estimation of two-character undefined words.

Example 2.1 POS estimation problem is a main part of word formation. Modifier + (head) Noun = Noun in Chinese:

1. Noun + Noun = Noun
   铁路 and 信纸
2. Adj + Noun = Noun
   温泉 and 红旗
3. Verb + Noun = Noun
   燃料 and 刊物
4. Num + Noun = Noun
   八股 and 千金
5. Pronoun + Noun =Noun
   他人 and 何处
6. Etc.

[1] analyzed all the possible syntactic relations between the POS's, such as,斗争, 挨打, 截获, 迁就, 捐助, 推翻,etc.. There exist many complicated word formation rules among the syntactic relations.

In the system of Segmentation and POS Tagging, we have several guessing rules for undefined words. For example, 数学$\in L$, but 数学家$\notin L$, we have rules:

1) 数学 / 物理学 / ⋮ / 化学 } + 家 = { 数学家 / 物理学家 / ⋮ / 化学家

2) 制药 / 化工 / ⋮ / 轴承 } + 厂 = { 制药厂 / 化工厂 / ⋮ / 轴承厂

3) etc.

FMM (Forward Most Matching) and BMM (Backward Most Matching) are two main methods of segmentation. Besides these, we have some statistical ones, such as HMM (Hidden Markov Model) and its simplified version. Because segmentation and POS tagging are parallel in our system, the identification of undefined words is constrained by word formation and the probability of POS string. So, our work provides the pre-processing of POS estimation. It's not true that the larger the lexicon is, the better the segmentation is. We have tested to expand the lexicon by the software of Finite State Calculus (FSC) in XRCE (Xerox Research Center Europe), the relation between the size of the lexicon

and chaos degree of segmentation shows that:

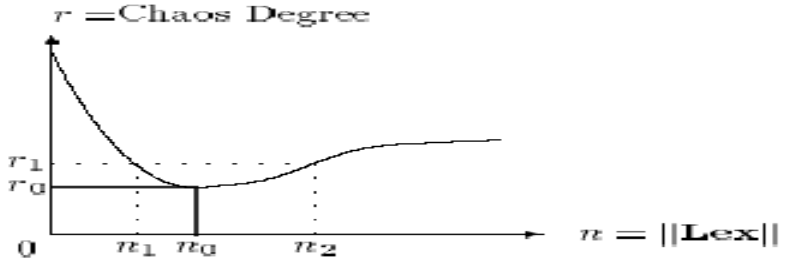

Figure-1

With the increase of the size of lexicon, chaos degree r (more details in [8] and [9]) decreases till some point n_0 where r reaches the least value and then r increases. We are seeking n_0 or $[n_1, n_2]$ where $n_0 \in [n_1, n_2]$ such that $| r_1 - r_2 | < \varepsilon$. From the viewpoint of Dynamic Lexicon, we have to adjust the structure of our lexicon to the segmentation. Thus, some words will become undefined, but which are easily identified by some word formation rules.

3 Test of POS Formation

Definition 3.1 Let $x, y, z \in \mathbf{POS}$, where **POS** is the set of all POS's. Let $\| x + y = z \|$ be the number of samples in the lexicon. Define

1) $$P_1 = \frac{\| x + y = z \|}{\| x \| \times \| y \|}$$

2) $$P_2 = \frac{\| x + y = z \|}{\| z \|}$$

The test is divided into two steps. Firstly, we have tested the POS formation on a lexicon of 10,000 the most popular Chinese words, and then the lexicon of GKB (70,000 words). The threshold level is:

1) Num>10
2) $P_1 > 0.10$
3) $P_2 > 1.00$

There are 18 rules for the POS estimation in the first experiment and 17 rules in the second. The results of the two experiments are as follows:

Experiment-1						Experiment-2					
Rules			Num	P_1	P_2	Rules			Num	P_1	P_2
f	+f	=f	13	3.6	15.1	f	+f	=f	13	2.08	8.07
c	+f	=f	10	2.19	11.6	c	+f	=f	11	0.96	6.83
a	+a	=a	96	0.35	7.36	a	+a	=a	294	0.36	11.6
n	+n	=n	99	0.24	3.16	n	+n	=n	131	0.39	5.89
d	+d	=d	38	0.28	6.57	d	+d	=d	49	0.15	5.48
a	+n	=n	67	0.2	2.14	a	+n	=n	104	0.64	4.72
n	+f	=s	18	0.47	24	n	+f	=s	30	0.21	18.2
r	+k	=r	6	2.31	7.14	c	+c	=c	14	0.66	10.7
r	+f	=f	8	1.62	9.3	N	+k	=n	391	0.25	1.76
P	+u	=p	8	1.57	23.5	n	+a	=n	292	0.18	1.31
						d	+c	=d	13	0.16	1.45
						d	+k	=d	10	0.13	1.12

R	+f	=r	5	1.01	5.95
P	+p	=p	5	0.19	14.7
D	+p	=d	11	0.19	1.9
Q	+f	=t	6	0.18	3.21
N	+f	=t	6	0.16	3.21
F	+n	=f	11	0.29	12.7
F	+q	=b	5	0.15	2.66
R	+q	=r	5	0.11	5.95

Table-1

d	+c	=c	10	0.12	7.69
a	+k	=b	13	0.11	1.99
a	+A	=n	226	0.11	1.02
n	+k	=n	248	1	1.12

Table-2

There are three distinct cases:

1) Pseudo Rules --- in Table-1 but not in Table-2:
   R+k=r r+f=f p+u=p R+f=r
   F+n=f p+p=p d+p=d Q+f=t
   N+f=t f+q=b r+q=r
2) New Rules --- in Table-2 but not in Table-1:
   n+k=n c+c=c N+k=n N+a=n
   d+c=d d+k=d d+c=c A+k=b
   a+A=n
3) Stable Rules --- both in Table-1 and Table-2:
   f+f=f c+f=f n+f=s A+a=a
   d+d=d n+n=n a+n=n

The order of the reliability is: Pseudo ≺ New ≺ Stable

4 Linguistic Explanation

The stable rules for POS formation is as follows:

1) Adj + Adj = Adj;
   矮小, 暗淡, 饱满, 专横, 壮实, etc.
2) Noun + Noun = Noun;
   堤坝, 钟摆, 碑帖, 笔画, 草图, etc.
3) Adj + Noun = Noun;
   矮墙, 暗号, 白人, 长剑, 痴情, etc.
4) Adverb + Adverb = Adverb;
   必定, 单独, 过多, 未必, 重新, etc.
5) Noun + Location (f) = place(s);
   岸上, 城外, 地下, 湖底, 幕后, etc.
6) etc.

1) to 5) are the most familiar rules in Chinese. The statistical data from both experiments have proved the intuition of linguists. Some reliable new rules:

1) Noun morpheme + suffix = Noun;
   案头, 鼻炎, 磁场, 敌方, 汉学, etc.
2) Noun + Adj = Noun;
   财富, 汗臭, 权贵, 水旱, 月亮, etc.
3) Adj + Adj morpheme = Noun;

小康, 羞耻, 冤孽, 重孝, 壮锦, etc.

4) Noun + suffix = Noun

癌症, 笔者, 画法, 草体, 孔型, etc.

For the stable and new rules of POS formation, there is only one rule for adjective, that is, Adj + Adj = Adj. And more than 1/3 of the left ones are for nouns:

1) Noun + Noun = Noun;
2) Adj + Noun = Noun;
3) Noun morpheme + suffix = Noun;
4) Noun + Adj = Noun;
5) Adj + Adj morpheme = Noun;
6) Noun + suffix = Noun.

There is no rule for verbs in our experiments, which indicates that the Chinese verb formation is quite complicated at the level of POS. As described in the following Pants Model, more rules about adjectives and verbs need finer POS set and more semantic constraints.

Pants Model

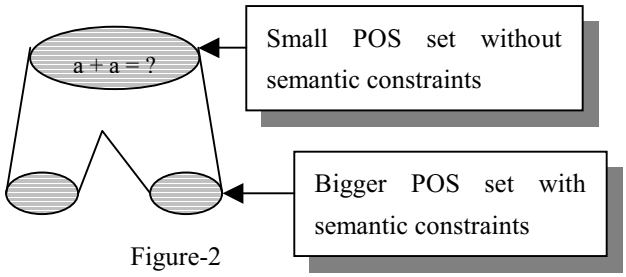

Figure-2

Usually the mathematical methods in Computational Linguistics are quantitative analysis (fuzzy math. or statistics). Some qualitative analysis (such as, the topological structures) of linguistic objects has been done in [4] and [5].

5 Fuzzy POS Formation

Definition 5.1 Let $\chi_z : \mathbf{POS} \times \mathbf{POS} \to [0,1]$ be a fuzzy function describing the degree of $x + y = z, \forall x, y \in \mathbf{POS}$, where

$$\sum_{x,y \in \mathbf{POS}} \chi_z(x,y) = 1$$

Example 5.1 POS estimation for Chinese nouns and adjectives are defined by the following functions:

POS estimation function for Chinese nouns

(x,y)		$\chi_n(x,y)$
(N,k)	$\to$	1/7
(n,k)	$\to$	1/7
(a,n)	$\to$	1/7
(n,a)	$\to$	1/7
(n,n)	$\to$	1/7
(a,A)	$\to$	1/7
Otherwise	$\to$	1/11158

Table-3

POS estimation function for Chinese adjectives

(x,y)		$\chi_a(x,y)$
(a,a)	$\to$	1/2
Otherwise	$\to$	1/3198

Table-4

We have tested the POS estimation rules on the set of undefined words from the well done segmented and POS tagged corpus (a half-year newspaper of *People's Dail*y) with more than 14,000,000 characters, the qualitative results are the same with our conclusion. One part of the further work is to apply the result in the POS estimation parallel with the calculation of the probability of POS sequence in HMM.

6 Conclusion

The automatic identification of undefined words and their POS estimations are the most important steps in Chinese Processing. In this paper, we discussed the POS estimations for undefined Chinese words. By two experiments, we got some POS formation rules, mainly for nouns and adjectives. Then, we mentioned some examples and the analysis of the linguistic foundations. From the Pants Model, we could simulate the process of POS (or word) formation at different level. [1] has tried in a deeper syntactic way. [2], [3], [6] and [10] have realized the importance of POS (word) formation in Chinese Processing. More details will be found in our further work.

7 Acknowledgements

We appreciate all the colleagues in our institute who discussed the POS estimation problems with us, especially the members of the seminar of Natural Language Processing. Also we thank Ms. Feng who helped us on the test of POS formation rules. And we thank our friends in XRCE, especially J-P. Chanod, L. Karttunen, A. Schiller and J. Quint who helped us on FSC. Finally, we show our respect to Prof. Yu and Prof. Zhu for their hard work on the construction of syntactic lexicon and Computa- tional Lexicology.

References

[1] Liu Yun et al, *Construction of the Contemporary Chinese Compound Words Database and its Application*, in Proceedings of the Second International Conference on New Technologies in Teaching and Learning Chinese, pp273- 228, 2000

[2] Lu Zhiwei, *Chinese Word Formation*, Zhonghua Publishing Co., 1975

[3] Lü Wenhua, The Design of Morpheme in Chinese Teaching, in Study of the Grammatical System of Chinese Teaching, Beijing Language and Culture University Press, 1999

[4] R. Thom, *Stabilité Structurelle et Morphogénèse*, W. A. Benjamin Reading, Mass. 1972

[5] R. Thom, *Mathematical Models of Morphogenesis*, Ellis Horwood Limited, 1983

[6] Yu Shiwen and Zhu Xuefeng, *The Development and Application of Morpheme-base of Contemporary Chinese*, in Chinese Teaching in the World, Vol.2, 1999

[7] Yu Shiwen and Zhu Xuefeng et al, *The Grammatical Knowledge - base of Contemporary Chinese --- Complete Specification*, Tsinghua Press, 1998

[8] Yu Jiangsheng, *Machine Segmentation Ambiguities and Dynamic Lexicon*, in the Association of Artificial Intelligence in China, 2000

[9] Yu Jiangsheng and Yu Shiwen, *Some Problems of Chinese Segmentation*, in The First International Workshop on MultiMedia Annotation (MMA --- 2001)

[10] Yuan Chunfa, Huang Changning, *The study of Chinese Morphemes and Word Formation based on Morpheme Database*, in Chinese Teaching in the World, Vol. 2, 1998

[11] Zhang Wangxi and Cui Yonghua, *The Basic Situations of the Study of Grammatical Problems in Chinese Teaching*, in The Study of Applied Linguistics in China Next Century (Chen Zhangtai et

al ed.), Chinese Teaching Press, 1999

[12] L.A. Zadeh, *fuzzy Sets*, Inf. Control. 8, p338-353, 1965

[13] L.A. Zadeh, *The Concept of a Linguistic Variable and its Application to Approximate Reasoning*, 1975