Incorporating New Words Detection with Chinese Word Segmentation

Hua-Ping ZHANG[1] **Jian GAO**[1] **Qian MO**[2] **He-Yan HUANG**[1]
[1] Beijing Institute of Technology, Beijing, P.R.C 100081
[2] Beijing Technology and Business University, Beijing, P.R.C 100048
Email: kevinzhang@bit.edu.cn

Abstract

With development in Chinese words segmentation, in-vocabulary word segmentation and named entity recognition achieves state-of-art performance. However, new words become bottleneck to Chinese word segmentation. This paper presents the result from Beijing Institute of Technology (BIT) in the Sixth International Chinese Word Segmentation Bakeoff in 2010. Firstly, the author reviewed the problem caused by the new words in Chinese texts, then introduced the algorithm of new words detection. The final section provided the official evaluation result in this bakeoff and gave conclusions.

1 Introduction

With the rapid development of Internet with Chinese language, word segmentation received extensive attention. In-vocabulary word segmentation and named entity recognition have achieved state-of-art performance. Chinese words are actually not well defined, and there is not a commonly accepted segmentation lexicon. It is hard to collect all possible new words, or predict new words occurred in the future. New words is the bottleneck to Chinese word segmentation. The problem became more severe with word segmentation on special domain texts, such as computer, medicine and finance. There are much specialized words which are difficult to be exported to the lexicon. So new words detection is very important, which would have more substantial impact on the performance of word segmentation than ambiguous segmentation.

In this paper，we presented a method of new words detection, and then detailed the process of Chinese word segmentation incorporating new words detection. The last section provided the evaluation and gave our conclusions.

2 Problem with new words

In the process of Chinese word Segmentation, there are many mistakes because of new words. These new words are Out of vocabulary (OOV), so the system couldn't distinguish them from original texts, and then impacted the results of word segmentation.

We gave an example from Text C in medicine domain to explain and detect the new words.

"我们以阿司匹林作为对照药物，证实盐酸沙格雷酯治疗 12 周后，糖尿病合并 PAD 患者的无痛行走距离和能够耐受疼痛的最大行走距离都明显改善，ABI 明显改善，明显优于阿司匹林的疗效。"

The sentence should be segmented as follows：

"我们 以 阿司匹林 作为 对照 药物 ， 证实 盐酸沙格雷酯 治疗 12 周后 ， 糖尿病 合并 PAD 患者 的 无痛 行走 距离 和 能够 耐受 疼痛 的 最大 行走 距离 都 明显 改善 ， ABI 明显 改善 ， 明显 优于 阿司匹林 的 疗效 。 "

Here, both "阿司匹林" and "盐酸沙格雷酯" are domain words, or new words beyond general segmentation lexicon. Therefore, new words from domain should be detected and added to segmentation lexicon before word segmentation.

3 Word segmentation with new words detection

3.1 Framework

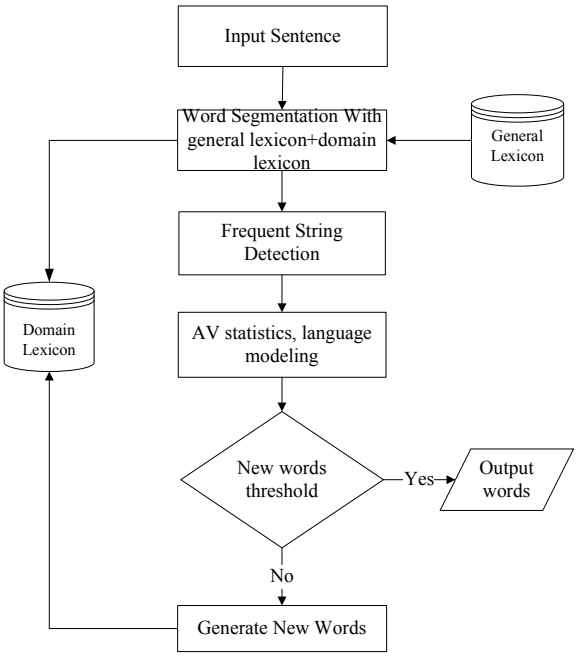

Figure 1: The framework of Chinese word segmentation incorporating with new words detection

As illustrated in Figure 1, Chinese word segmentation with new words detection is a recursive process. The process is given as follows:

1. Making Chinese word segmentation with domain lexicon beyond general lexicon.
2. Frequent string (over twice) finding with postfix tree algorithm, and taking them as new words candidate.
3. Access Variety statistics [Haodi Feng etc. 2004], and language modeling on word formation. [Hemin, 2006]
4. Exporting new words to domain lexicon.
5. Recursively, until no more new word detected.
6. Output final word sequence.

3.2 The process of new words detection

Simple word segmentation is the first step of processing of Chinese language when we deal with a very long Chinese article. The method of word segmentation is based on HHMM, and Zhang and Liu (2003) have given detailed explanation about this.

During the process of word segmentation in the first, the system records the words which occur frequently. We can set a threshold value of words' occurrence frequency. As long as the word occurrence frequency reaches this value, this word could be recorded in the system as frequent string.

With the frequent strings detected, we can do the further analysis. For every frequent string, we check its left and right adjacent one in the original text segmented respectively. Through this step, we find the adjacent words which occur next to some frequent string detected. If the adjacent word also occurs very frequently, or even it occurs at the left or right of the frequent string every time, it's great possibility that the string detected and the adjacent word could merge into one word.

With the detection in above steps, we gain new words from Chinese texts. Then we import these new words into domain lexicon and our lexicon is updated. With the lexicon containing new words, we can do the next cycle recursively and revise continually.

Then, we can see this is a recursive structure. Through the continued process of word segmentation and new words detection, the state of segmentation tends to be steady. The condition of steady state has several kinds such as no more new words detected or the latest result equal to the previous one. At this time, we can break the recursion and output the final result.

This is an example. This sentence is from Text D in finance domain

"雷曼兄弟公司倒闭不到一年，金融市场已经稳定，股市也已回升。" ("The financial market has been stable and the stock has rebounded in less than one year time after Lehman Brother Corporation went bankrupt.")

After word segmentation with original lexicon, this altered sentence is:

"雷/曼/兄弟/公司/倒闭/不/到/一/年/，/金融市场/已经/稳定/，/股市/也/已/回升/。/"

"雷曼兄弟" is a new word as a organization name and it is hard to be collected. Like this kind of word, there are difficulties to add new words to update the lexicon in time. So it is normal to segment this word "雷曼兄弟" into three words.

Through frequent string detection, we gain these three words “雷”, “曼” and “兄弟”. With the adjacent analysis, we find the word “雷” occurs 6 times, “曼” 3 times and “兄弟” 3 times.

The character “雷” occurs 3 times in the detected word “布雷迪” and 3 times at the left of the word “曼”. So we can consider the word “雷曼” as a whole word.

Then we can easily find the words “兄弟” are always at the right of words “雷曼”. So it’s necessary to consider “雷曼兄弟” as a whole word.

4 Evaluation

The performance of word segmentation is measured by test precision (P), test recall (R), F score (which is defined as 2PR/(P+R)) and the OOV recall rate.

In this competition, our test corpus involved literature, computer, medicine and Finance, totally 425KB. We take 6 months data of The People's Daily to be the training corpus. From Table 1, we can see the official evaluation result.

	R	P	F1	OOV R	OOV RR	IV RR
A-Literature	0.965	0.94	0.952	0.069	0.814	0.976
B-Computer	0.951	0.926	0.938	0.152	0.775	0.982
C-Medicine	0.953	0.913	0.933	0.11	0.704	0.984
D-Finance	0.963	0.938	0.95	0.087	0.758	0.982

Table 1. Official evaluation result

Our system got high Precision Rate and Recall Rate after testing the texts in four domains, especially Recall Rate is all over 95%. And we also could see that this system detected most new words through several measures of OOV, especially IV RR is all over 97.5%. This proved that the system could be able to get a nice result through processing professional articles in literature, computer, medicine and finance domains, and we believed it also could do well in other domains. This also proved that the method of new words detection with Chinese word segmentation was competitive.

5Conclusion

Through this competition, we’ve found a lot of problems needed to be solved in Chinese word segmentation and tried our best to improve the system. Finally, we proposed the method of new words detection in Chinese word segmentation. But we still had some shortage during the evaluation and need to improve in the future.

References

Lawrence. R.Rabiner.1989. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. Proceedings of IEEE 77(2): pp.257-286.

Hua-Ping Zhang, Qun Liu. *Model of Chinese Words Rough Segmentation Based on N-Shortest-Paths Method*. Journal of Chinese information processing, 2002,16(5):1-7 (in Chinese)

ZHANG Hua-Ping, LIU Qun, Zhang Hao and Cheng Xue-Qi. 2002. *Automatic Recognition of Chinese Unknown Words Recognition*. Proc. of First SigHan attached on COLING 2002

ZHANG Hua-Ping, LIU Qun, YU Hong-Kui, CHENG Xue-Qi, BAI Shuo. *Chinese Named Entity Recognition Using Role Model*. International Journal of Computational Linguistics and Chinese language processing, 2003,Vol. 8 (2)

Mao-yuan Zhang, Zheng-ding Lu, Chun-yan Zou. *A Chinese word segmentation based on language situation in processing ambiguous words.* Information Sciences 162 (2004) 275–285

Gao, Jianfeng, Andi Wu, Mu Li, Chang-Ning Huang,Hongqiao Li, Xinsong Xia, and Haowei Qin. *Adaptive Chinese word segmentation.* ACL2004. July 21-26.

Haodi Feng, Kang Chen, Xiaotie Deng, Weimin Zheng *Accessor Variety Criteria for Chinese Word Extraction*, Computational Linguistics March 2004, Vol. 30, No. 1: 75–93.

Hemin, **Web-Oriented Chinese Meaningful String Mining, M.Sc** Thesis of Graduate University of Chinese Academy of Scienses. 2006