Rules Design in Word Segmentation of Chinese Micro-Blog

Zong Hao **Derek F. Wong** **Lidia S.Chao**
NLP^{2}CT Research Group, Department of Computer and Information Science, University of Macau, Macau SAR, China
{MB15463, DerekFw, Lidiac}@umac.mo

Abstract

This paper proposed a Hidden Markov Model (HMM) based tokenizer for Chinese micro-blog texts. Comparing with normal Chinese texts, micro-blog texts contain more uncertainties. These uncertainties are generally aroused by the irregular use of bloggers (such as network words, dialect words, wrong written characters, mixture of foreign words and symbols, etc.). Besides the lack of the annotated training corpus is also a restriction in solving this task. Hence the segmentation for micro-blogs is much more difficult than that of general text, we present an HMM based segmentation model integrated with a pre and post correction module. The evaluation results show that the proposed approach can achieve an F-measure of 90.98% on test set of 5,000 sentences.

1 Introduction

Word segmentation is a common task in Chinese information processing. This task is to split a character sequence into many small groups by inserting a space between two neighbor groups. Each group, as a Chinese word, represents an independent meaning. For example, given a character sequence "李明是个好人" (*Li Ming is a good man*); the segmentation result will be "李明 /是/个/好人". We select this task as our research target because it is a very common task and many scholars had done a lot of experiments on it. We can easily compare our method with others', more importantly, segmentation is normally the first step to process the Chinese text. The quality of it may seriously affect on the later processing.

Micro-blog has more uncertainties than normal Chinese text. For instance, the micro-blog texts contain a large number of *network words* like "打酱油" and "楼主" which are easily to be mis-segmented due to the arbitrary nature of language. The dialect words and wrong written words are also easily to be mis-segmented according to a limited knowledge of these.

To accomplish this task, many approaches had been proposed. The first adapted and efficient approach is the Maximum Matching (Wong & Chan, 1996). Its segmentation accuracy is depending on the quality of system dictionary. System dictionary is a manual defined lexicon that it contains the majority of standardized words. However, with the development of language, new words are springing up. The system dictionary cannot track of newly born vocabularies. Several years later machine learning approaches had been applied. The Maximum Entropy (Shi, 2005) achieved the highest accuracy among most of the tasks in SIGHAN-2005 Bake-off[1] Segmentation contest. During the same contest Conditional Random Fields (CRFs) (Zhou et al, 2005) has the best performance in solving the out-of-vocabulary (OOV) problem. Hidden Markov Model (HMM) (Zhang et al, 2003) is another efficient approach in Chinese segmentation. It has an efficient approach in handling the word ambiguity[2] issue. Furthermore it achieved the best result in the first Chinese segmentation competition[3].

The most two common issues in Chinese segmentation are OOV and ambiguity. In this word we assume that the some existing segmentation

[1] http://www.sighan.org/swclp4/

[2] Example: "长春市长春饭店" can be segmented as "长春 市长 春 饭店" or "长春市 长春 饭店". So this sentence is ambiguity.

[3] Proceedings of the Second SIGHAN Workshop on Chinese Language Processing task2: Chinese segmentation.

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 90–94,
Tianjin, China, 20-21 DEC. 2012

tool is already very good. Based on this tool we designed several rules to modify the segmentation result to overcome its inadaptation for this domain.

2 Task Specialty

Comparing with normal Chinese text segmentation, micro-blog text segmentation has to overcome more difficulties due to the arbitrary nature of language. We will show this in detail in the following sections.

Network words: thanks to the speed of spread in the internet age, a large amount of irregular words had been widely emerged and accepted. These words such as "屌丝" (*diao si*) usually have the rich connotation and can represent the heartfelt idea. Therefore although the network words are irregularly written and some of them even are not grammatical, they are still widely used.

Accent words: accent words such as "木有" (*mu you*) and "酱紫" (*jiang zi*) are nonstandard pronounced words. These words are widely used because it can show their accent and sounds cute.

Wrong written words: these words such as "戒子" (*jie zi*) (refer to "戒指" (*ring*)) and "针贬" (*zhen bian*) (refer to "针砭" (*zheng bin*)) are very are hard to be recognized by the current segmentation approaches.

The mixture of foreign words: many people like to write with foreign words such as words or phrases of "打 (*da*) ball" (*play basketball*) and "很 (*hen*) down" (*very disappointed*). It is very popular and common in some specific topic. Using these words can express the richest meaning with the less characters.

3 Rules design for Chinese micro-blog segmentation

Chinese micro-blog texts are unrestrained. In this task we followed the tagging schema of "Specification for Corpus Processing at Peking University"[4] in the design of our model.

In this word, under the assumption that a segmentation system for general text is already good , for a special domain we only need to do some modification to make the segmentation result better. The main frame of this system is using ICTCLAS[5] as the segmentation tool. Based on it result, we do a group of preprocessing and postprocessing to get a better result.

After analyzed the 500 sentences train corpora, we found that there are some rules in the segmentation that it is very difficult to use other approaches to recognize them. Therefore we designed rules as the pre-processing and postprocessing of this system.

Pre-processing rules

The rules designed for preprocessing are the URL and E-mail address. In ICTCLAS, URL and E-mail address cannot be segmented at all and these mis-segmented URL and E-mail address may encounter more segmentation errors later.

This system used regular expression[6] to define the segmentation rules for URL and E-mail address. Figure 1 showed the improvement after applying the preprocessing rules.

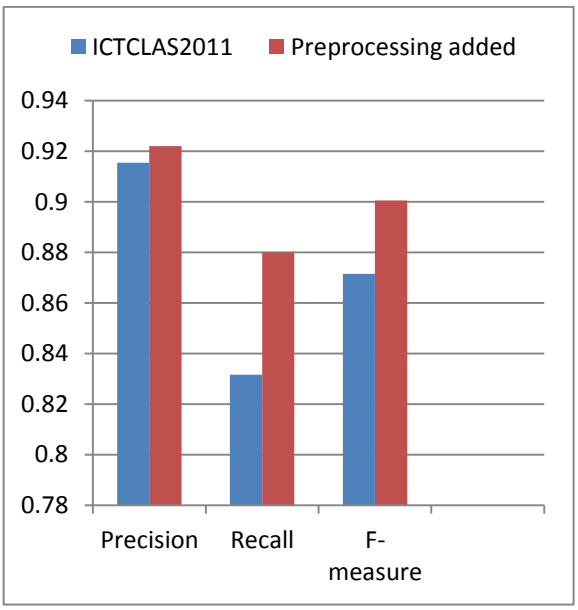

Figure 1. Improvement after preprocessing added.

Postprocessing rules

These rules are generated after analyzing the fragments of ICTCLAS2011's segmentation result. The fragments revealed that the ICTCLAS2011 cannot segment the roll-call system in micro-blog which will frequently occur in micro-blog texts. For instance, "@一移已易-YEE33333" will be segmented as "@/一/移/已/易/-YEE33333" while the right segmentation is "@/一移已易/-/YEE33333". Error about this is complicated and we believed that if the system

[4] http://www.icl.pku.edu.cn/icl_groups/corpus/coprus-annotation.htm

[5] http://www.ictclas.org/index.html

[6] Regular expression referenced from http://en.wikipedia.org/wiki/Regular_expression

use rules as the segmentation constrain, this error will be totally correct.

The detail of this roll-call system rules is followed:

1. If the text starts with a group of meaningful Chinese words, use normal segmentation strategy (ICTCLAS). For example, "@花心女想要去流浪 1989" should be segmented as "@/花/心/女/想/要/去/流浪/1989"
2. If the text starts with a group of meaningless Chinese characters, group these characters together. For example, "@一移已易-YEE33333" should be segmented as "@/一移已易/- /YEE33333" rather than "@/一/移/已/易/ - /YEE33333".
3. If the text starts with an account ID, the ID characters should be grouped together. Example: "@super_lv" should be segmented as "@/super_lv".
4. When the symbol "-" or "_" is between English and Chinese. If the left is Chinese and the right is English the symbol should be segmented alone. Else it should group to the English. Example: "@一移已易-YEE33333" should be segmented as "@/一移已易/-/YEE33333", "@小丁_Vic" segmented as "@/小丁/_/Vic", "@12th_章" segmented as "@/12th_/章", "@BETTY-萍萍" segmented as "@/BETTY-/萍萍"
5. If the roll-call system contains Chinese personal name, the surname should be separated. Example: "@刘彦友 2527" should be segmented as "@/刘/彦友/2527".

Beside the roll-call system, two other rules are also applied in this system.

1. For continuous symbols "." And "。", every three of them should be a group. For instance, "….." should be segmented as "…/.." and "。。。。。。" should be segmented as "。。。/。。。".
2. For continuous mimetic words, they should grouped together. For example "哈哈哈哈哈哈哈" should be segmented as "哈哈哈哈哈哈哈" and "呵呵呵呵呵呵" should be segmented as "呵呵呵呵呵".

Figure 2 showed the improve after postprocessing added.

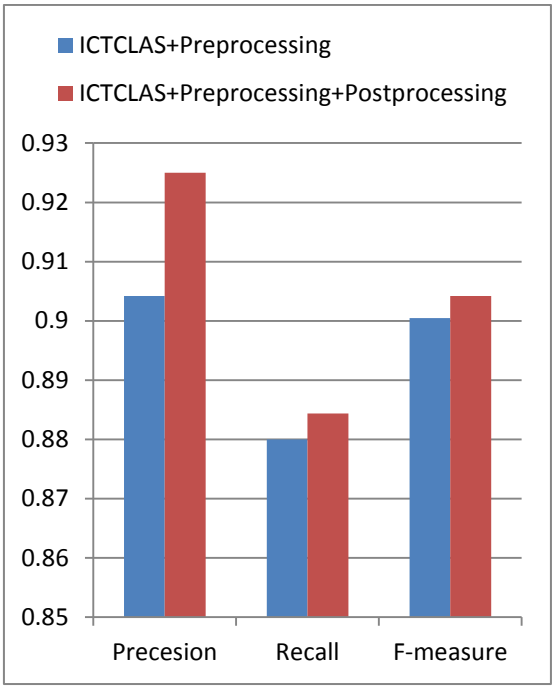

Figure 2. Improvement after postprocessing added

Although the improvement is not very obvious, we can ensure that all the special case covered by these rules will be completely correct.

4 External Dictionary

In order to overcome the sparsity of training data, an external dictionary is necessary.

To get a micro-blog related dictionary we referenced a famous Chinese Input Method: Sougou[7] Input. We got the network dictionary (9850 words) and applied in this system. But mechanically added this network dictionary did not improve the result a lot. Therefore we analyzed the detail terms in this dictionary. We found that many terms like "祝妈妈身体健康"(*wish mom healthy*) did not be segmented. Then we use ICTCLAS again to segment the terms in this dictionary and group all singer character together. For example: "醉驾" (*drunk driving*) will be segmented as "醉/驾", then we group these two singer characters together as "醉驾".

Figure 3 showed the improvement after applied the external dictionary.

[7] http://pinyin.sogou.com/

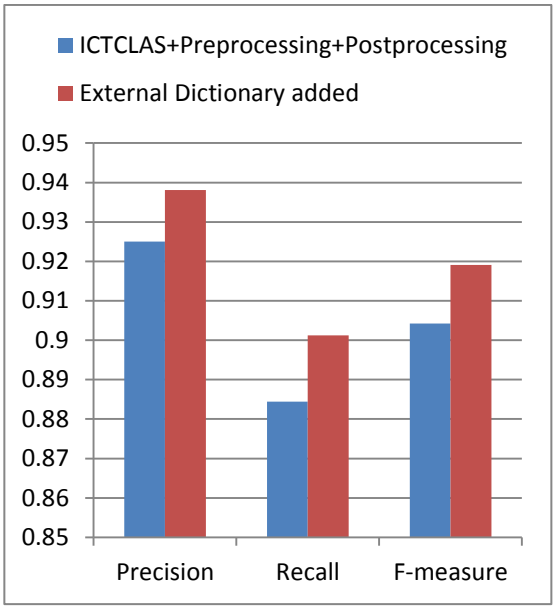

Figure 3. Improvement after external dictionary added

5 Named Entity Recognition

After applied all approaches above the evaluation result still can not reach the state-of-the-art, the segmentation error showed that the named entities encountered much error. Then named entity recognition procedure imported into this system.

Chinese Named Entity Recognition (NER) is more complex than English Named Entity Recognition because it contains a segmentation step before. In this system NER is playing a very important role. For those unlabeled data, it will do NER first. If this system recognizes that the name in this text is not a Named Entity (NE), it will directly assert that this text belongs to the OTHER class. If the name in the text is a NE, we will then mark all the NE in this text to help the later work.

Before we do NER we have to do the Chinese segmentation and Part-of-Speech (POS) tagging. Here this system used ICTCLAS 2011 with additional user dictionary to improve the segmentation and POS tagging accuracy.

Conditional Random Fields (CRFs) is the most popular approach to do NER task. This approach is easy to implement and usually achieve a very high accuracy. A Study on Features of the CRFs-based Chinese Named Entity Recognition (Duan & Zheng, 2011) did a lot of work on this task and gave a conclusion of the feature selection. This system also used CRFs to do the NER. The CRFs toolkit adopted in this system is CRF++[8] toolkit and used feature is three single characters (before, current, after), three POS tags (before, current, after), some suffix and prefix (s/f) information and three segmentation label sets (before, current, after). The training data set is January-June People's Daily 1998. We get F-measure 91.4% from our test set.

Figure 4 showed the improvement of NER added into this system.

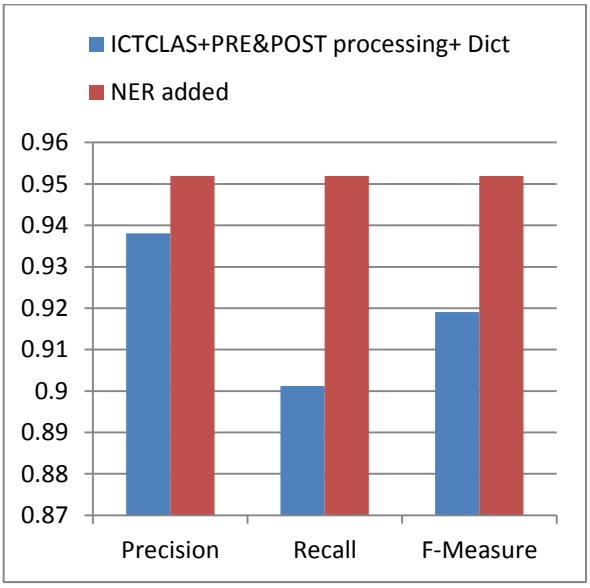

Figure 4. Improvement after NER added

6 Conclusion

This paper proposed a modification of ICTCLAS a basic segmentation tool for the Chinese microblog segmentation. These modifications contain preprocessing, postprocessing in rule level, an external network dictionary with a little amelioration and a named entity recognition. All these modifications improved the original segmentation result in 8.4 percent which is a very obvious improvement in Chinese segmentation. However due to the time limit, there are still some other issues we had not considered such as wrong written error and the mixture of foreign words.

Table 1 showed our final evaluation result in SIGHAN-2012 Bake-off Task 1.

Precision	0.9000
Recall	0.9199
F-measure	0.9098
All right sentences	1,388
All right sentence rate	27.76%

Table 1. Final evaluation result

[8] CRF++: Yet Another Toolkit [CP/OL]. http://crfpp.googlecode.com/svn/trunk/doc/index.html

The reason why we get a large decrease may be that the train corpora is so small that we have not anticipated any other error in the test set.

Acknowledgments

This work was partially supported by the Research Committee of University of Macau under grant UL019B/09-Y3/EEE/LYP01/FST, and also supported by Science and Technology Development Fund of Macau under grant 057/2009/A2.

References

Wong, P. & Chan, C. 1996. *Chinese word segmentation based on maximum matching and word binding force*, Proceedings of the 16th conference on Computational linguistics-Volume 1, 200–203.

Shi, W. 2005. *Chinese Word Segmentation Based On Direct Maximum Entropy Model*, Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.

Zhang, H.P. and Liu, Q and Cheng, X.Q and Zhang, H and Yu, H.K 2003. *Chinese lexical analysis using hierarchical hidden markov model*, Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17, 63-70.

Duan, H. and Zheng, Y. 2011. *A study on features of the CRFs-based Chinese Named Entity Recognition*, International Journal of Advanced Intelligence-Volume 3, 287-294.