Word Segmenter for Chinese Micro-blogging Text Segmentation — Report for CIPS-SIGHAN’2014 Bakeoff

Lu Xiang **Xiaoqing Li** **Yu Zhou**

Institute of Automation Chinese Academy of Sciences, Beijing, China
{lu.xiang, xqli, yzhou}@nlpr.ia.ac.cn

Abstract

This paper presents our system for the CIPS-SIGHAN-2014 bakeoff task of Chinese word segmentation. This system adopts a character-based joint approach, which combines a character-based generative model and a character-based discriminative model. To further improve the performance in cross-domain, an external dictionary is employed. In addition, pre-processing and post-processing rules are utilized to further improve the performance. The final performance on the test corpus shows that our system achieves comparable results with other state-of-the-art systems.

1 Introduction

Because Chinese text is written without natural delimiters, word segmentation is a prerequisite and fundamental task in Chinese natural language processing. And many approaches have been proposed for this task. Among these methods, the character-based tagging approach (Xue, 2003) has become the prevailing technique for Chinese word segmentation (CWS) due to its good performance. In recent years, within the framework of character-based, much efforts (Tseng et al., 2005; Zhang et al., 2006; Jiang et al., 2008) have been made to further improve word segmentation’s performance.

The character-based joint model (Wang et al., 2010, Wang et al., 2012) achieves a good balance between in-vocabulary (IV) words recognition and out-of-vocabulary (OOV) words identification. So, in this evaluation task, following their work we adopt the character-based joint model as our basic system, which combines a character-based discriminative model and a character-based generative model. The generative module holds a robust performance on IV words, while the discriminative module can handle the extra features easily and enhance the OOV words segmentation.

Because the 2014 SIGHAN bakeoff task of Chinese Word Segmentation is an opened evaluation task and no training set is provided, the OOV problem will be more serious. Although the discriminative module can handle some cases of OOV, the performance is less preferable if no technique is utilized. So to further improve the performance of the basic system and minimize the OOV, we employ an external dictionary containing a large set of unknown words from different domains. Another notable problem is the Microblog text segmentation because Microblog has become a new Internet literary which is different from the genres of common text. To make our system more robust on Microblog text, we propose several simple but novel pre-processing and post-processing approaches in our system.

The final results show that our system performs well on test set and achieves comparable segmentation results with other participants.

2 System Description

2.1 Character-Based Joint Model

The character-based joint model in our system consists of two basic components:

- The character-based discriminative model.
- The character-based generative model.

The character-based discriminative model (Xue, 2003) is based on a Maximum Entropy (ME) framework (Ratnaparkhi, 1998) and can be formulated as follows:

$$P\left(t_1^n \mid c_1^n\right) \approx \prod_{k=1}^{n} P\left(t_k \mid t_{k-1}, c_{k-2}^{k+2}\right) \quad (1)$$

Where t_k is a member of {**B**, **M**, **E**, **S**}, in which **B**, **M** and **E** indicate the *Beginning*, *Middle* and *End* of character c_k in its associated word respectively, and **S** denotes that it’s a *Single-character* word. For example, the word “北京市 (Beijing

City)" will be assigned with the corresponding tags as: "北/B (North) 京/M (Capital) 市/E (City)".

This discriminative model can incorporate extra features easily and the Maximum Entropy Modeling Toolkit[1] given by Zhang Le is used to implement the module. In our experiments, this model is trained with Gaussian prior 1.0 and 600 iterations.

The character-based generative module is a character-tag-pair-based trigram model (Wang et al., 2009) and can be expressed as below:

$$P\left([c,t]_1^n\right) \approx \prod_{i=1}^{n} P\left([c,t]_i \mid [c,t]_{i-2}^{i-1}\right) \qquad (2)$$

SRI Language Modeling Toolkit[2] (Stolcke, 2002) is used to train the generative trigram model with modified Kneser-Ney smoothing (Chen and Goodman, 1998) in our experiments.

The character-based joint model combines the above discriminative module and the generative module with log-linear interpolation as follows:

$$\begin{aligned} Score(t_k) &= \alpha \times \log\left(P\left([c,t]_k \mid [c,t]_{k-2}^{k-1}\right)\right) \\ &\quad + (1-\alpha) \times \log\left(P\left(t_k \mid t_{k-1}, c_{k-2}^{k+2}\right)\right) \end{aligned} \qquad (3)$$

Where the parameter $\alpha\left(0.0 \le \alpha \le 1.0\right)$ is the weight for the generative model and can be obtained from the development set. $Score(t_k)$ will be directly used to search for the best sequence. We set an empirical value 0.4 to α as there is no development-set for various domains.

2.2 Features

The feature templates used in the character-based discriminative model are listed below:

(a) $C_n\left(n=-2,-1,0,1,2\right)$;

(b) $C_nC_{n+1}\left(n=-2,-1,0,1\right)$;

(c) $C_{-1}C_1$;

(d) $T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$

In the above templates, C_n represents a Chinese character and the index n indicates the position. For example, when we consider the third character "奥" in the sequence "北京奥运会", template (a) results in the features as following: C_{-2}=北, C_{-1}=京, C_0=奥, C_1=运, C_2=会, and template (b) generates the features as: $C_{-2}C_{-1}$=北京, $C_{-1}C_0$=京奥, C_0C_1=奥运, C_1C_2=运会, and template (c) gives the feature $C_{-1}C_1$=京运.

Template (d) is the feature of character type and five type classes are defined: dates ("年", "月", "日", the Chinese character for "year", "month" and "day" respectively) represents class 0; foreign alphabets represent class 1; Arabic and Chinese numbers represent class 2; punctuation represents class 3 and other characters represent class 4. For example, when considering the character "，" in the sequence "八月，阿Q", the feature $T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$ will be set to "20341".

2.3 External Dictionary

OOV words is a main problem faced by a Chinese word segmenter and it will lead to lower accuracy if the sentence to be segmented contains many OOV words. To address the problem of OOV words, we use an external dictionary containing a large set of predefined words. We following the method presented in Low et al. (2005) to use the dictionary. In this method, some sequence of neighboring characters around C_0 will be looked up in a dictionary using maximum match strategy. And the longest matching word *W* will be chosen. Let t_0 be the boundary tag of C_0 in *W*, *L* the number of characters in *W*, and $C_1\left(C_{-1}\right)$ be the character immediately following (preceding) C_0 in the sentence. We then add the following features derived from the dictionary:

(e) Lt_0

(f) $C_nt_0\left(n=-1,0,1\right)$

For example, consider the sentence "北京奥运会...". When processing the current character C_0 "京", we will try to match the following candidates "京", "北京", "京奥", "北京奥", "京奥运", "北京奥运" and "京奥运会" against existing word in the external dictionary. Assuming that both "京奥" and "京奥运" are found in the dictionary, then the longest matching word "京奥运" will be chosen. And the value of *W*, t_0, *L*, C_{-1} and C_1 are "京奥运", **B**, 3, "北" and "奥" respectively.

In this work, we collect dictionaries from the Internet, including the title of Wikipedia[3], the title of Hudong Baike[4], Sogou word bank[5] and

[1] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

[2] http://www.speech.sri.com/projects/srilm/

[3] http://zh.wikipedia.org

[4] http://www.baike.com/

some other internet dictionaries. Finally, we obtain a dictionary containing 5,893,038 words in our system.

2.4 Restrictions in Constructing Lattice

When considering a character in the sequence, we take the type information of both the previous and the next character into consideration and use some restrictions to obtain a better tag lattice (Wang et al., 2010). The restrictions are listed as follows:

- If the previous, the current and the next characters are all English or numbers, we would fix the current tag to be "M";
- If the previous and the next characters are both English or numbers, while the current character is a connective symbol such as "-", "/", "_", "\" etc., we would also fix the current tag to be "M";
- Otherwise, all four tags {B, E, M, S} would be given to the current character.

3 Rule-based Adaptation

The state-of-the-art Chinese word segmentation systems can achieve a quite high performance on well-formed text, while the performance of Microblog text segmentation is not satisfying due to the specificity of Microblog text. For example, there are lots of emotion symbols, URLs, abbreviations, consecutive and identical punctuations and special characters in Microblog text. In order to make our system more robust on segmenting Microblog data, we propose some heuristic pre-processing and post-processing rules to avoid some segmentation errors.

3.1 Pre-processing

As mentioned above, the Microblog texts contain much noise like special format words and characters. And such kind of noise will affect the segmentation performance. In order to remove these noise, we will pre-process the text before segmentation.

Since URL, email and consecutive punctuations should be treated as one word and these content types can be easily recognized using the regex expressions, we first replace all these content to special characters before segmentation, and then restore all the special characters to the original characters after the segmentation. Table 1 shows the content type we will process in the pre-processing stage.

Table 1: Content type of pre-processing

Type	Example
URL	http://t.cn/RPdBAPV
Email	hanhuahr@126.com
Consecutive punctuations	。。。 ！！！！

3.2 Post-processing

We use some heuristic rules to further post-process the results generated by the segmenter and the rules are described below:

1) **Numeral and Quantifier**: In our results, some numerals and quantifiers such as "两个" and "三张" are segmented as one unit. But in fact, the numeral and quantifier should be segmented into two words except some few words like "一个". So we use a simple rule to split these cases in which the previous word is a numeral and the next word is a quantifier.
2) **Continuous mimetic words**: There are many continuous mimetic words in Microblog, such as "哈哈哈哈哈", "呵呵呵" . This kind of words should be treated as one unit. But our system splits each character into one word. Hence, we apply a rule to group the continuous mimetic words together.
3) **Emoticons**: some consecutive punctuations like ":-)" represent an emoticon and have some certain meanings. These emoticons should be grouped together. We have collected a list of emoticons from the web. For any consecutive punctuations, we join them together as a single word if they appear in the emoticon list.

4 Experiments

4.1 Data sets

Since the Chinese word segmentation task focuses on the performance of multi-domain, we use five datasets as our test data. Four of the five datasets are the test data of SIGHAN10 closed track and the rest one is the 500 Microblog messages released by SIGHAN12. Hence, our test data covers 5 domains: Literature (Testing-A, containing 671 sentences), Computer (Testing-B, containing 1,330 sentences), Medicine (Testing-C, containing 1,309 sentences), Finance (Testing-D, containing 561 sentences) and Microblog (Testing-E, containing 500 sentences). The training data of our segmenter consists of two parts: one is the Peking University Corpora (PKU)

[5] http://pinyin.sogou.com/dict/

from January to June and the other is manually annotated Microblog data which contains nearly 7000 sentences.

4.2 Experimental Results

We first evaluate our approach on the five test datasets using different strategies. The results are shown in Table 2 and the evaluation criterion is F-score. The strategies we used are:

- **Joint**: represents the result of our model without dictionary.
- **+Dic**: represents the result of our model using the external dictionary.
- **+Rule**: represents the result of our model using the external dictionary and the pre-processing and post-processing rules.

Table 2: Evaluation results with different strategies

	Joint	+Dic	+Rule
Testing-A	0.9590	0.9622	0.9628
Testing-B	0.9589	0.9630	0.9634
Testing-C	0.9522	0.9557	0.9557
Testing-D	0.9670	0.9686	0.9696
Testing-E	0.9338	0.9381	0.9412

As Table 2 shows, our joint model performs well on all the five datasets even though the domain of the training data which is mainly composed of news data is different from the test sets. This shows that our character-based joint model is very robust and can achieve a good balance between in-vocabulary (IV) words recognition and OOV words identification

After the external dictionary added, the performance increased a lot, which shows the external dictionary is very useful and can help alleviate the OOV problem efficiently. Finally, we adopt the pre-processing and post-processing rules in our system, the performance can be further improved on all testing set except Testing-C.

Table 3: Final Result of the Test Set

	P	R	F
Final Test	0.9592	0.9566	0.9578

Since the final test data will be multi-domain, we add all the five datasets to the training data and retrain the segmentation model. Then we apply the retrained model to the final test data (containing 1,665 sentences) and the performance is shown in Table 3. Table 3 shows that our system can achieve an F-score of 0.9578.

5 Conclusion

Our system is based on a character-based joint model, which combines a generative module and a discriminative module. In addition, we employ an external dictionary and propose several pre-processing and post-processing rules to further improve the performance. Our system achieves comparable performance with other participants.

Acknowledgements

We would like to thank Mr. Zhilin Zhang for his great help in implementation and experiments of our system. This work is supported by the Hi-Tech Research and Development Program (863) of China under grant No. 2012AA011101 and High New Technology Research and Development Program of Xinjiang Uyghur Autonomous Region under grant No. 201312103.

References

Stanley F. Chen and Joshua Goodman, 1998. An empirical study of smoothing techniques for language modeling. *Technical Report TR-10-98, Harvard University Center for Research in Computing Technology*.

Wenbin Jiang, Liang Huang, Qun Liu and Yajuan Lu, 2008. A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In *Proceedings of ACL*, pages 897-904.

Adwait Ratnaparkhi, 1998. Maximum entropy models for natural language ambiguity resolution. University of Pennsylvania.

Andreas Stolcke, 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 311-318.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning, 2005. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168-171.

Kun Wang, Chengqing Zong and Keh-Yih Su, 2009. Which is more suitable for Chinese word segmentation, the generative model or the discriminative one? In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC23)*, pages 827-834.

Kun Wang, Chengqing Zong and Keh-Yih Su, 2010. A Character-Based Joint Model for CIPS-SIGHAN Word Segmentation Bakeoff 2010. *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP2010)*, pages 245-248.

Kun Wang, Chengqing Zong and Keh-Yih Su, 2010. A Character-Based Joint Model for Chinese Word Segmentation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, China, August 23-27, 2010. Pages 1173-1181.

Kun Wang, Chengqing Zong and Keh-Yih Su, 2012. Integrating generative and discriminative character based models for Chinese word segmentation. *ACM Transactions on Asian Language Information Processing (TALIP).*

Low, Jin Kiat et al., 2005. A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 161-164.

Nianwen Xue, 2003. Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing*, 8 (1). pages 29-48.

Ruiqiang Zhang, Genichiro Kikui and Eiichiro Sumita, 2006. Subword-based Tagging for Confidence dependent Chinese Word Segmentation. In *Proceedings of the COLING/ACL*, pages 961-968.

Xiaojin Zhu, 2006. Semi-supervised learning literature survey. *Technical Report 1530*, Computer Sciences, University of Wisconsin-Madison.