Soochow University Word Segmenter for SIGHAN 2012 Bakeoff

Yan Fang Zhongqing Wang Shoushan Li Zhongguo Li Richen Xu Leixin Cai
Natural Language Processing Lab
Soochow University, Suzhou, China, 215006
{fangyan0108, wangzq.antony, shoushan.li, eemath}@gmail.com,
xurichen@yeah.net, leixincai@gmail.com

Abstract

This paper presents a Chinese Word Segmentation system on MicroBlog corpora for the CIPS-SIGHAN Word Segmentation Bakeoff 2012. Our system employs Conditional Random Fields (CRF) as the segmentation model. To make our model more adaptive to MicroBlog, we manually analyze and annotate many MicroBlog messages. After manually checking and analyzing the MicroBlog text, we propose several pre-processing and post-processing rules to improve the performance. As a result, our system obtains a competitive F-score in comparison with other participating systems.

1 Introduction

Because Chinese context is written without natural delimiters, word segmentation becomes an essential initial step in many tasks on Chinese language processing. Though recognizing words seems easy for human beings, automatic Chinese Word Segmentation by computers is not a trivial problem (Xue, 2003; Li et al., 2012). The state-of-the-art Chinese Word Segmentation systems have achieved a quite high precision on traditional media text. However, the performance of segmentation is not so satisfying for MicroBlog corpora. MicroBlog messages are often short, and they make heavy use of colloquial language. Furthermore, they require situational context for interpretation. Thus, we first analyze and annotate some MicroBlog messages, and then propose a novel pre-processing and post-processing approach on the CRF-based segmentation system for the MicroBlog corpora. The experimental results show that our system performs well on MicroBlog corpora and could yield comparable segmentation results with other participants.

2 Our System

2.1 Overview

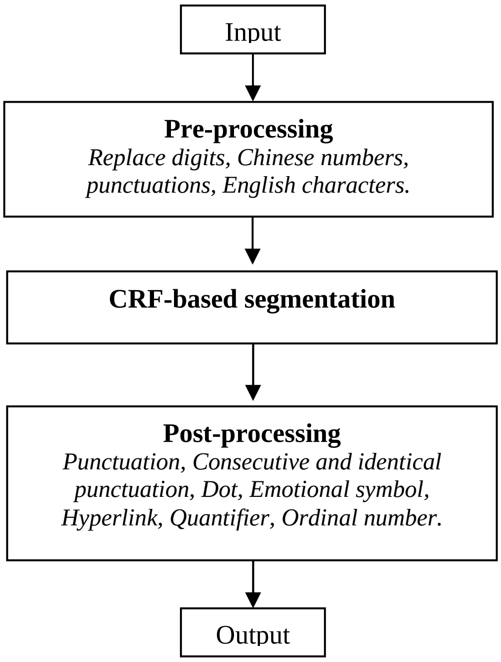
Input
Pre-processing
Replace digits, Chinese numbers, punctuations, English characters.
CRF-based segmentation
Post-processing
Punctuation, Consecutive and identical punctuation, Dot, Emotional symbol, Hyperlink, Quantifier, Ordinal number.
Output

Figure 1: The architecture of our Chinese word segmentation system

Figure 1 illustrates the framework of our Chinese word segmentation system. The whole system contains three main components: preprocessing, CRF-based segmentation, and post-processing. We will introduce them in the following subsections in detail.

2.2 Resources

Note that the 2012 SIGHAN bakeoff task of Chinese Word Segmentation on MicroBlog

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 47–50,
Tianjin, China, 20-21 DEC. 2012

corpora provides no training data. To make our system more adaptive to the new domain, we get the training data by ourselves. The training data we used consists of two parts. The first one is the Peking University Corpora (PKU) from January to June. Secondly, we collect a certain amount of raw sentences from Sina MicroBlog (The size is 90M) for further manual annotation. Due to the big size of the data, we conduct an active learning approach to actively select the informative boundaries for manual annotating and the size of the selected data is reduced to about 3% annotation size (Li et al., 2012).

2.3 Segmentation Method

The approach of character-based tagging is popular for Chinese word segmentation (Xue, 2003; Xue and Shen, 2003). The backbone of our system is a character-based segmenter with the application of CRF (Zhao and Kit, 2008; Li and Huang, 2009) that provides a framework to use a large number of linguistic features. It can avoid the so-called 'label-bias' problem in some degree and is originally introduced into the language processing tasks in Lafferty et al. (2001).

The probability assigned to a label sequence for a particular sequence of characters by a CRF is given by the following equation:

$$P_{\lambda}(Y \mid X) = \frac{1}{Z(x)} \exp(\sum_{c \in C} \sum_{k} \lambda_k f_k (Y_c, X, c))$$

Y is the label sequence for the sentence; X is the sequence of unsegmented characters; $Z(X)$ is a normalization term; f_k is a feature function, and c indexes into characters in the sequence being labeled.

The character based tagging model for Chinese word segmentation is usually based on either maximum entropy or CRF which regards a segmentation procedure as a tagging process. For detailed information, please refer Adwait (1996). The probability model and corresponding feature function is defined over the set $H \times T$, where H represents the set of possible contexts and T represents the set of possible tags. Generally, a feature function can be found as follows,

$$f(h,t) = \begin{cases} 1, \text{if } h{=}h_i \text{ is satisfied and } t{=}t_j \\ 0, \text{otherwise,} \end{cases}$$

where $h_i \in H$ and $t_i \in T$

The features used in our experiments are straightforward and include the following types:

$$c_0, c_1, c_{-1}c_0, c_0c_1, c_1c_2$$

Where c stands for character (Zhao et al., 2006). The subscripts are position indicators. 0 means the current word; -1,-2, the first or second word to the left; 1, 2, the first or second word to the right.

A forward-backward algorithm is used in training and the Viterbi algorithm is used in decoding.

As for tag set, we apply a four-tag tagging scheme. That is, each Chinese character can be assigned to one of the tags in {B, M, E, S}. The tag B, M, E represent the character being the beginning, middle, and end of a multiple-character word respectively while the tag S represents the character being a single-character word.

3 The Preprocessing and Post-processing Rules

3.1 Preprocessing

Before applying the training data to train CRF, we use some preprocessing rules on training data.

Because English characters and digits are frequently out-of-vocabulary words, we replace all the English character and digits to special characters before segmentation processing, and we will restore all these special characters to the original character after segmentation processing. The following table shows the character type we choose in the pre-processing step.

Type	Example
English characters	Today is Friday
Chinese digital	一百五十九
Digital	2012
Punctuations	“，”，“。”，“！”

Table1 Explaining of preprocessing

3.2 Post-processing

In the segmentation result from the CRF segmenter, we find that some errors could be corrected by some heuristic rules. For this purpose, we propose seven rules as follows.

- **Punctuation**: punctuation tends to be a single-character word. If a punctuation’s previous character and next character are both Chinese characters, i.e. not punctuation, digit, or English character, we always regard the punctuation as a word.
- **Consecutive and identical punctuation**: some consecutive and identical punctuation tend to be joined together as a word. For example, “———” represents a Chinese hyphen which consist of three “—”, and consecutive punctuations of “.” or “。” all presents suspension points. Inspired by this observation, we would like to join some consecutive and identical punctuations as a single word.
- **Dot**: when the character “•” appears in the training data, it is generally used as a connection symbol in a foreign personal name, such as “奥黛丽•赫本”. Taking this observation into consideration, we always join the character “•” and its previous and next segment units into a single word. A similar rule is designed to join consecutive digits on the sides of the symbol “.”, ex. “0.99”.
- **Emotion symbol**: some consecutive punctuations have special meanings. For example, “'^_^” and “:-)” all mean smiling expressions. “'T_T” and “Q_Q” all mean sad expressions. This is a kind of network language features. So when we come across these consecutive punctuations, we applied a rule to join them together as a single word.
- **Hyperlink**: MicroBlog corpora contain so many web sites, and there are always than one hyperlinks appear together. Under these circumstances, the CRF-based segementer always has difficulties to separate them. So we get a rule to correct it.
- **Quantifier**: some quantifiers after numbers were connected as one word in our result. Such as “三个”, “5 斤”, “1cm”. So we proposed a rule to split those words whose previous character is a number and next character is a quantifier or a unit. But the word “一个” would be regarded as an exception.
- **Ordinal number**: in Chinese, ordinal numbers are regard as one word such as the word “第一” . In MicroBlog corpora, there are many cases that a digit after the character “第” like “第 3”, we also regard them as one word. To this end, we join the character “第” with its next segment which consists of digits completely. A similar rule is designed to join integers or decimals with its next character “%”.

Table 2 summarizes all the rules we utilized in the post-processing step.

Rule type	Example
Punctuation	你好吗？很好。
Consecutive and identical punctuation	思考中。。。。。。
Dot	奥黛丽•赫本
Emotion symbol	今天很开心^_^
Hyperlink	http://www.taobao.com/
Quantifier	买了 5 斤苹果
Ordinal number	开学的第一天

Table 2 Explaining of post-processing

4 Experiments

For this CIPS-SIGHAN bakeoff, we focus on the Chinese Word Segmentation task on MicroBlog corpora. Before the final test, we use the data provided by SIGHAN 2012 which consists of approximately 500 messages from MicroBlog to test our approaches described in the previous sections. The results are shown in Table 3, where P, R, F represents the precision rate, recall rate and harmonic average measure rate respectively. The approaches we used are:

- **Basic** represents the result of our model using only the corpora of PKU.
- **+Pre** represents the result of our model using the preprocessing rules.
- **+Post** represents the result of our model using the post-processing rules.

- **+Ann** represents the result of our model using the annotated data.

As the table shows, after the use of preprocessing rules, the results are somehow decreased. The reason for a worse performance is that when we use preprocessing rules, we treat all the digits, other types alike, as the same, whereas they are always different in some circumstance. For example, we always regard "一个" as one word, but others like "三个", "五个" all regard as two words. These problems are solved in post-processing, and we can see that the designed post-processing rules are effective and thus could greatly improve the results.

	P	R	F
Basic	0.8959	0.8613	0.8782
+Pre	0.8589	0.8585	0.8587
+Pre +Post	0.9225	0.9153	0.9187
+Pre+Post+Ann	0.9336	0.9224	0.9279

Table 3 Performances tested before final test

P	R	F	CS	CSP
0.9383	0.9346	0.9365	1909	38.18

Table 4 Performance of the final test.

The final test data consists of approximately 5,000 texts from MicroBlog. The performances are shown in Table 4, where CS indicates the sum of correct sentences, and CSP indicates the percentage of correct sentences in all the sentences. The F-score we achieved is 0.9365, which is higher than the results when only 500 texts are used.

5 Conclusion

In this paper, we introduce our Chinese word segmentation system for SIGHAN 2012. The nice performance of system are attributed to three main aspects: the CRF learning algorithm, the newly annotated data on Sina MiroBlog, the preprocessing and post-processing rules.

References

Adwait R. 1996. A Maximum Entropy Part-of-speech Tagger. In Proceedings of the Empirical Method in Natural Language Processing Conference, 133-142. University of Pennsylvania.

Lafferty J., A. McCallum and F. Pereira. 2001. Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning, 282-289. June 28-July 01, 2001.

Li S., G. Zhou, and C. Huang. 2012. Active Learning for Chinese Word Segmentation. In Proceeding of COLING-2012, poster. To appear.

Li S., C. Huang. 2009. Word Boundary Decision with CRF for Chinese Word Segmentation. In proceeding of PACLIC-2009, pages 726-732.

Xue N. 2003. Chinese Word Segmentation as Character Tagging. Computational Linguistics and Chinese Language processing, Vol. 8(1): 29-48.

Xue N. and L. Shen. 2003. Chinese Word Segmentation as LMR Tagging. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL'03, 176-179. Sapporo, Japan.

Zhao H. and C. Kit. 2008. Unsupervised Segmentation Helps Superivised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. In Proceedings of SIGHAN-6 2008, pages 106-111.

Zhao H., C. Huang, M. Li and B. Lu. 2006. Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling. In Proceedings of PACLIC-2006, pages 87-94.