Rules-based Chinese Word Segmentation on MicroBlog for CIPS-SIGHAN on CLP2012

Jing Zhang
Dalian University of Technology,
DaLian, P. R. China.
zhangjingqf@mail.dlut.edu.cn

Degen Huang
Dalian University of Technology,
DaLian, P. R. China.
huangdg@dlut.edu.cn

Xia Han
Dalian University of Technology,
DaLian, P. R. China.
hanxia@mail.dlut.edu.cn

Wei Wang
Dalian University of Technology,
DaLian, P. R. China.
wangwei.dl@263.net

Abstract

In this evaluation, we have taken part in the task of the Word Segmentation on Chinese MicroBlog. In this task, after analysing the feature of the MicroBlog and the result of our original Chinese word segmentation system, four Optimization Rules are proposed to optimize the segmentation algorithm for Chinese word segmentation on MicroBlog corpora. The optimized segmentation system is based on character-based and word-based Conditional Random Fields (CRFs). Experiments show that the optimized segmentation system can obviously improve the performance of CWS on MicroBlog corpora.

1 Introduction

Chinese word segmentation is a crucial fundamental task in Chinese language processing. After years of intensive researches, Chinese word segmentation has achieved a quite high performance. However, it is not so satisfying when the Chinese word segmentation works on MicroBlog corpora. This CIPS-SIGHAN-2012 bake-off task of Chinese word segmentation focuses on the performance of Chinese word segmentation algorithms on MicroBlog corpora. This evaluation is an opened evaluation on simplified Chinese word segmentation task. The task provides no training set, and we are free to use data learned or model trained from any resources.

In this evaluation task, we propose some useful optimization rules for Chinese Word Segmentation (CWS) on MicroBlog corpora, after analysing the results of segmentation on MicroBlog corpora by our original CWS system, which combines character-based and word-based Conditional Random Fields (CRFs).

The rest of this paper is organized as follows. Section II outlines the new Chinese word segmentation algorithm on MicroBlog corpora. Section III reports the results of experiments and some discussions. Finally, some conclusions are presented in Section IV.

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 74–78,
Tianjin, China, 20-21 DEC. 2012

2 Word Segmentation Algorithm

2.1 Machine Learning Models

Conditional random fields (CRFs), a statistical model for sequence labeling, was first introduced by Lafferty, McCallum and Pereira (2001). It is the undirected graph theory that CRFs mainly use to achieve global optimum sequence labeling. It is good enough to avoid label bias problem by using a global normalization.

In previous labeling task of character-based CRFs, the number of the characters in the observed sequence is as same as the one in the annotation sequence. However, for CWS task, the input of n-character will generate the output of m-word sequence on such a condition that m is not larger than n. But this problem can be well solved by word-lattice based CRFs, because the conditional probability of the output sequence depends no longer on the number of the observed sequence, but the words in the output path. For a given input sentence, its possible paths may be various and the word-lattice can well represent this phenomenon. A word-lattice can not only express all possible segmentation paths, but also reflect the different attributes of all possible words in the path. Zhang, Chen and Hu (2012) and Nakagawa (2004) have successfully used the word lattice in Japanese lexical analysis.

Our paper adopt the word-lattice based CRFs that combines the character-based CRFs and the word-based CRFs, and specifically, we put the candidate words selected by the character-based CRFs into a word-lattice, and then label all the candidate words in the word-lattice using word-based CRFs model. When training the word-lattice based CRFs model, the maximum likelihood estimation is used in order to avoid overloading. And Viterbi algorithm is utilized in the decoding process which is similar with (Huang and Tong, 2012).

2.2 Feature Templates

The character-based CRFs in our method adopt a 6-tag set in (Kudo, Yamamoto and Matsumoto, 2004), and its feature template comes from (Huang and Tong, 2010), including C_{-1}, C_0, C_1, $C_{-1}C_0$, C_0C_1, $C_{-1}C_1$ and $T_{-1}T_0T_1$, in which C stands for a character and T stands for the type of characters, such as Number, String. Character and so on, and the subscripts -1, 0 and 1 stand for the previous, current and next character, respectively. Four categories of character sets are predefined as: Numbers, Letters, Punctuation and Chinese characters. The feature templates of the character-based CRFs are described in detail in Table 1.

No.	Feature	Description of Feature
1	C_0	The current character
2	C_1	The later character
3	C_{-1}	The former character
4	$C_{-1}C_0$	The former and the current characters
5	C_0C_1	The current and the later characters
6	$C_{-1}C_1$	The former and the later characters
7	$C_{-1}C_0C_1$	The former, current and the later characters
8	$T_{-1}T_0T_1$	The type of the former, current and the later characters

Table 1: The feature templates of the character-based CRFs

Two kinds of features are selected for the word-based CRFs, like (Huang and Tong, 2012): unigram features and bigram features. The unigram ones only consider the attributes information of current word, and bigram ones are also called compound features, which utilize contextual information of multiple words. Theoretically, the current word's context sliding window can be infinitely large, but due to efficiency factors, we define the sliding window as 2. The specific features are W_0, T_0, W_0T_0, W_0T_1, T_0T_1, W_0W_1, where W stands for the morphology of the word, T stands for the part-of-speech of the words, and subscript 0 and subscript 1, respectively, stand for the former and the latter of two adjacent words. Furthermore, the Accessor Variety (AV) in (Zhao, Huang and Li, 2006) is applied as global feature. The feature templates of the word-based CRFs are shown in Table 2.

No.	Feature	Description of Feature
1	W_0	The current word
2	T_0	The POS of the current word
3	$T_{-1}T_0$	The POS of the former and the current words
4	T_0T_1	The POS of the current and the later words

Table 2: The feature templates of the word-based CRFs

2.3 Optimization Rules

As we all know, there exist plenty of new words, a great variety of symbols, and a good deal of URLs in MicroBlog corpora. Those features bring a big challenge to Chinese word segmentation. Considering the features of MicroBlog corpora and the segmentation result of our original Chinese word segmentation system, we propose several rules to optimize the segmentation result on MicroBlog corpora.

The features of MicroBlog corpora we summarized is as follows:

I. There are a lot of new words in MicroBlog, such as "团购" tuan-gou (online shopping), "点评网" dian-ping-wang (HankowThames), "有木有" you-mu-you (yes or not) and so on.
II. Many kinds of special symbols are used in MicroBlog, and what we deal with is mainly included in the following three cases:
   A. All kinds of combinations of the punctuation, especially, "！", "。", "-", for example, "其实应该很开心的呀！ ！ ！ ！" (Actually we are suposed to be very happy!!!!), "我要虚脱了。。。" (I am exhausted。。。).
   B. The frequently use of "@", e.g. "@姚晨" @-yao-chen.
   C. There also exist large number of emoticon icons, for instance, "^_^", "→_→" and so on.
III. The expression forms of time or date are quite various.
IV. The vast majority of the MicroBlog have URLs.

Our original segmentation system does not solve those problems mentioned above very well. Therefore, considering these characteristics of the MicroBlog, we propose some optimization rules to optimize the original results, which finally improve the segmentation results.

The rules are described in detail as follows:

Optimization Rule 1: With regards to the first feature, we use the contextual information, which is described in detail in (Huang and Tong, 2012) to calculate the frequency of the new words, and then added the high-frequency words to the dictionary.

Optimization Rule 2: According to the second feature, we have collected some commonly used combinations of punctuations to the dictionary.

Optimization Rule 3: Considering the third feature, the original system can not deal with the string of time very well, for instance, "2012年11月8日" (November 8, 2012), the string of time is segmented as "2012/年/11/月/8/日", while the correct segmentation is "2012年/11月/8日". Under this circumstance, we have built a set of Time Templates. If the string matches any of the Time Templates, it will be segment as Time.

Optimization Rule 4: As to the last point, first, we search for the key word "http", and then we look for the right boundary of the URLs. At last, we merge all the string between the "http" and the right boundary together.

2.4 Word Segmentation Process

The Process of the optimized segmentation system is as follows:

Step1. Collect the commonly used combinations of punctuations to the dictionary which is mentioned in Rule 2.

Step2. Put all the candidate words in 3-Best paths selected by the character-based CRFs model into the word-lattice.

Step3. To build the word-lattice, in other word, give properties and costs to each node, the candidate words selected by character-based CRFs in Step2, in the word-lattice, which is divided into four cases to deal with:

①If the candidate words are in the system dictionary, then assign the properties and cost of the words in the system dictionary directly to the candidate words in the word-lattice.

②If the candidate words are not in the system dictionary, then we use Optimization Rule 1, search the dictionary of contextual information, if it is in there, then the properties of the words in the contextual information dictionary will be assiged to the candidate words, and a weight value, calculated by Eq. (1), will be added to the cost of the candidate words.

$$cost'(w)=\begin{cases}\dfrac{1.0}{rNum+1}\times cost_0(w) & rNum>0\\ \left(\dfrac{0.2}{\log(frequency+2)}+0.8\right)\times cost_0(w) & rNum=0\end{cases} \quad (1)$$

Where *w* stands for the word, and *t* on behalf of the Part of Speech (POS), and *Cost* represents the difficulty of the emerging of a candidate

word, and *Frequency* delegates the frequency of being a candidate word, and *rNum* is in the name of the frequency of being the node in the final segmentation path. Besides, $cost_0(w)$ stands for the original cost of the words.

③If the candidate words is not in the system dictionary, neither in the contextual information dictionary, then we will search the synonyms forest to find a synonym of the candidate words. If the synonym exits in the system dictionary, we'd like to replace the candidate word with it.

④If the above cases are not suitable for the candidate words, then the candidate words will be classified according to the classification mentioned above.

Step4. To find the optimal path, the least costly path of word segmentation, in the word-lattice using the Viterbi algorithm according to Eq. (4), and the values of $TransCost(t_i,t_{i+1})$ and $Cost(w_i)$ can be calculated by Eq. (2) and Eq. (3), respectively. Since all feature functions are binary ones, the cost of the word is equal to the sum of all the weight of the unigram features about the word, and the transition cost is equal to the sum of all bigram features about the two parts of speech.

$$Cost(w) = -factor * \sum_{f_k \in U(w)} \lambda_{f_k} \quad (2)$$

$$TransCost(t_1,t_2) = -factor * \sum_{f_k \in B(t_1,t_2)} \lambda_{f_k} \quad (3)$$

Where $U(w)$ is the unigram feature set of the current word, $B(t_1, t_2)$ is the bigram feature set of the adjacent words t_1 and t_2. λ_{fk} is the weight of the corresponding feature f_k and factor is the amplification coefficient.

$$Score(Y) = \sum_{i=0}^{Y^{\#}} (TransCost(t_i,t_{i+1}) + Cost(w_i)) \quad (4)$$

It can be seen from the above process that the factors of recognizing the territorial words are considered in Step3. Contextual information as well as synonym information is used to adjust the cost and the properties of the candidate words in the path, which can contribute to the follow-up Step4 to select the best path.

Step5. To optimize the original segmentation results. Optimization Rule 1 and Optimization Rule 2 have been used in the previous steps, while Optimization Rule 3 and Optimization Rule 4 are utilized in the end. The purpose of these two rules is to revise the segmentation results. In another word, some errors in the segmentation results can be corrected by Rule 3 and Rule 4.

3 Experiment Results

3.1 Data Sets

Our method is tested on the simplified Chinese MicroBlog testing data and the training data from the CIPS-SIGHAN-2012 bake-off task. The test corpus consists of approximately 5,000 texts from MicroBlog, and the training data includes 500 texts from MicroBlog with the gold standard result. The experiment results are evaluated by P (Precision), R (Recall) and F-measure. The system dictionary we used is extracted from the People's Daily from January to June, in 2000, containing 85000 words, with the POS. The word-based CRFs model is trained by the corpus with POS tag which is from the People's Daily of January, in 1998).

3.2 Evaluation Metrics

The metrics we used in this bake-off task is as follows:

$$\text{Precision} = \frac{\text{Num1}}{\text{Num2}} * 100\%$$

$$\text{Recall} = \frac{\text{Num1}}{\text{Num3}} * 100\%$$

$$F\text{-}measure = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} * 100\%$$

Num1 means the number of words correctly segmented.

Num2 stands for the number of words segmented.

Num3 means the number of words in the reference.

3.3 Experimental Results

Test Track	P	R	F
$Base_{500}$	78.76	88.59	83.39
$Final_{500}$	83.50	89.21	86.26
$Final_{5000}$	83.35	89.43	86.28

Table 3: The result of the experiments

In our experiments, at first, we use our original Chinese word segmentation system as the Baseline, and the 500 MicroBlog corpora provided by the organization are used as the test corpora. The segmentation result is shown in the first row of Table 3.

After that, in order to compare with the Baseline, we use the segmentation system added the optimization rules segments the 500 MicroBlog corpora, and the second row of Table 3 shows the result of this experiment. From the result we can see that our optimization works very well, and the F-measure is promoted obviously.

At last, we use the 5000 MicroBlog corpora to test our Final system, the segmentation system added the optimization rules, and we can see the result from the last row of Table 3, having the similar promotion with the second row.

From the above, we can clearly get that our optimized segmentation system can promote the segmentation performance significantly.

3.4 Error Analysis

Although the optimization rules improve the segmentation performance significantly, several typical errors are observed in the results of the experiment.

First, those problems we mentioned above are not solved thoroughly, especially the variety of punctuation problems. Because the combination is so flexible to sum up, we just summarize some frequently used combinations of punctuations.

Second, there still exist many new words which occur just a few times in the corpora, so they have not been added into the system dictionary eventually.

4 Conclusions

In this evaluation task, according to the features of MicroBlog, we propose several optimization rules of Chinese word segmentation on MicroBlog corpora. In the processing, experiments show that those optimization rules works very well on this task. While there still exit amount of problems need to be solved when Chinese word segmentation works on MicroBlog, and we have a lot of works to do in the future.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (No.61173100, No.61173101, No.61272375), Fundamental Research Funds for the Central Universities (DUT10RW202). The authors wish to thank Wu Qiong, Wang Dandan and for their useful suggestions, comments and help during the design and editing of the manuscript.

References

Huang Degen and Tong Deqin. 2012. Context Information and Fragments Based Cross-Domain Word Segmentation. J. China Communications, 9 (3): 49-57

Huang Degen, Tong Deqin, and Luo Yanyan. 2010. HMM Revises Low Marginal Probability by CRF for Chinese Word Segmentation. Proc of CIPS-SIGHAN Joint Conference on Chinese Processing. 216-220. ACL, Beijing

Kudo T, Yamamoto K, and Matsumoto Y. 2004. Applying conditional random fields to Japanese morphological analysis. Proc of EMNLP2004. 230-237. ACL, Barcelona

Lafferty J, McCallum A, and Pereira F. 2001. Conditional Random Fields: probabilistic models for segmenting and labeling sequence data. Proceedings of ICML2001. 282-289. Morgan Kaufmann, San Francisco

Nakagawa T. 2004. Chinese and Japanese word segmentation using word-level and character-level information. Proc of COLING 2004. 466-472. ACL, Geneva

Zhang Chongyang, Chen Zhigang, and Hu Guoping. 2012. A Chinese Word Segmentation System Based on Structured Support Vector Machine Utilization of Unlabeled Text Corpus. Proc of CIPS-SIGHAN Joint Conference on Chinese Processing. 221-227. ACL, Beijing

Zhao Hai, Huang Changning, and Li Mu, et al. 2006. Effective tag set selection in Chinese word segmentation via Conditional Random Field modeling. In PACLIC-20. 87-94. ACL, Wuhan