CRFs-Based Chinese Word Segmentation for Micro-Blog with Small-Scale Data

Longyue Wang
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau S.A.R., China.
vincentwang0229@hotmail.com

Derek F. Wong
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau S.A.R., China.
derekfw@umac.mo

Lidia S. Chao
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau S.A.R., China.
lidiasc@umac.mo

Junwen Xing
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau S.A.R., China.
nlp2ct.anson@gmail.com

Abstract

In this paper, we proposed a Chinese word segmentation model for micro-blog text. Although Conditional Random Fields (CRFs) models have been presented to deal with word segmentation, this is still the first time to apply it for the segmentation in the domain of Chinese micro-blog. Different from the genres of common articles, micro-blog has gradually become a new literary with the development of Internet. However, the unavailable of micro-blog training data has been the obstacle to develop a good segmenter based on trainable models. Considering the linguistic characteristics of the text, we proposed some methods to make the CRFs models suitable for segmentation in the domain of micro-blog. Several experiments have been conducted with different settings and then an optimal tagging method and feature templates have been designed. The proposed model has been implemented for the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing Bakeoff (Bakeoff-2012) and achieves a very high F-measure of 93.38% within the test set of 5,000 micro-blog sentences. One of our main contributions is the online version of toolkit[1], which provides segmentation service for Chinese micro-blog text.

1 Introduction

Unlike Roman alphabetic languages such as English, Portuguese, etc., Chinese has no explicit word delimiters within a sentence. Therefore, word segmentation is the very first step in Chinese information processing. After years of intensive researches, Chinese word segmentation has achieved a very good performance (Huang, 2007). However, the performance of segmentation is not so satisfied for tokenizing micro-blog corpora. The main reason is that traditional segmentation models are often trained from the corpora of news, literatures, etc. due to the availability of the corpora in these domains. When using the trained models to the text which is out of the trained domains (e.g. Internet, vernacular records), the precision and recall rates will decline sharply. Among all the proposed methods, character-based tagging with CRFs models have attracted more and more attention since it is firstly introduced into language processing (Lafferty et al., 2001). Reviewing the recent Bakeoffs, we found that Low et al. (2005) and Tseng et al.

[1] It can be accessed at http://nlp2ct.sftw.umac.mo/views/utility.html.

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 51–57,
Tianjin, China, 20-21 DEC. 2012

(2005) in Bakeoff-2005 have obtained the best results based on CRFs. Besides, the model of Zhao and Kit (2008) has been ranked at the top in the closed track of Bakeoff-2008, who integrated unsupervised segmentation and CRFs model. The results fully proved that CRFs can do well for the segmentation task.

In order to solve the segmentation problems with cross-domain data, Qin et al. (2010) proposed novel steps for the first CIPS-SIGHAN segmentation task and achieved 0.9278 of F-measure based on CRFs approach. The result shows that the out-of-domain resources could improve the segmentation performance, especially for the task with small-scale training data.

In our system, we continue to improve the CRFs-based tagging method. Not only the best feature templates are designed, but also that the use of a new 6-tag set, external 1-gram dictionaries and out-of-domain corpus are proposed to further improve the performance of Chinese segmentation for micro-blog. This will be helpful to the research on the tasks of information retrieval, Internet slang analysis and construction of corpus for domain of Chinese micro-blog.

The paper is organized as follows. The linguistics phenomena of micro-blog are analyzed in Section 2. Various tag sets used for segmentation are reviewed and discussed in Section 3. The feature template of the proposed approach is described in Section 4. Results, discussion and comparison between different strategies are given in Section 5 followed by a conclusion to end the paper.

2 Micro-Blogs

From the perspective of linguistics, micro-blog text is a new domain comparing with the common articles. In order to design a good segmentation system targeted for micro-blog text, several found phenomena are summarized in the followings:

2.1 Unknown Words

Similar to the Internet slang, many new words are used to emerge frequently and disseminate rapidly over the Internet. This will result in a lower recall rate of the segmentation system, because these out-of-vocabulary (OOV) words are not easy to be recognized. Here given some new words which occur on the Internet in recent years. "驴友 (*tour pal*)", "正太 (*cute boy*)" and "木有 (*have nothing*)" are all combined with two common Chinese characters and mostly used in the blogs. In order to improve the ability to identify these words, external word list of popular Internet slang are essential and used in our segmentation model.

2.2 Colloquial

Unlike written language which tends to be formal, users often express their moods and viewpoints with spoken language in their blogs. To simplify or personalize the descriptions, it is very common to see some sentences, which are colloquial, incomplete, or ungrammatical. For instance, the sentence "所有的一切，都在乎，真的 (*everything, treasure, really*)" was not only left out the subject "我 (*I*)", but also disrupted the word order (the formal sentence should be "我真的在乎所有的一切" / *I really treasure everything*). So syntax analysis such as part-of-speech etc. is not helpful to the segmentation in the domain of micro-blog and would seriously interfere the segmentation performance. Different from traditional methods for Chinese word segmentation, syntactic information was not used as features in our segmenter.

2.3 Brief

Micro-blogs are famous for its "micro". In another words, every micro-blog has a length limitation for all the users. For example, Sina Micro-Blog requires each blog has no more than 140 characters. Under this restriction, users get used to texting with shorter sentences. Several strategies to deliver more information with fewer words are adopted. For example, contractions (e.g. "女排" is short for "女子排球队" / *women's volleyball team*), idioms (e.g. "一言难尽" / *it is a long story*), classical Chinese texts (e.g. "但愿人长久，千里共婵娟" / *we wish each other a long life so as to share the beauty of this graceful moonlight, even though miles apart*) and foreign words are often used.

2.4 Non-Chinese Characters

The blog texts are nonstandard, because they are usually composed with a mixture of non-Chinese characters for some special purposes. Punctuations, foreign words, numbers and symbols are commonly used in blogs. For example, URLs often occur after reprints to cite them. Furthermore, several common symbols and numbers can be combined as emoticons (e.g. "^0^ (*smiling face*)"). And young people would like to use some foreign words (mostly English) to make

their expression outstanding. These make the micro-blog more complex compared to the formal text. Therefore, all of the cases should be well considered during the design of useful features for the proposed segmentation model.

According to the discussed phenomena, we analyzed the training data of the 500 micro-blog texts that are provided by the Bakeoff-2012. The detailed distributions are shown in Figure 1. The average length of blogs is around 64.62, which includes both Chinese and non-Chinese characters. In average, more than 60% of tokens are single character words. The length of most tokens (around 98.54%) is no more than 4. We consider the URLs as a single token, and hence URLs usually consists of multiple characters (the length is usually more than 6). So, there are more tokens of which length are more than six than the ones with less length (the lengths are 4, 5, and 6).

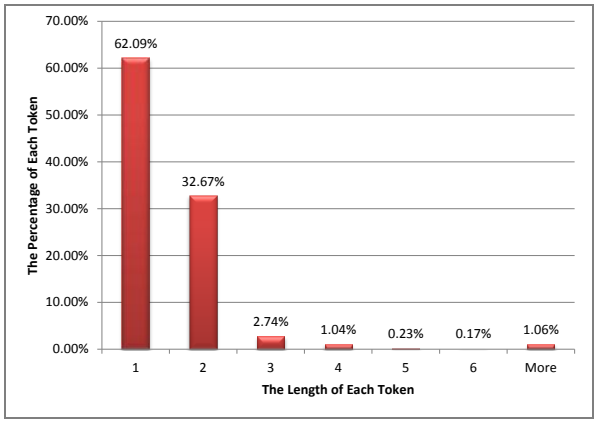

Figure 1. Distribution of token length in micro-blog

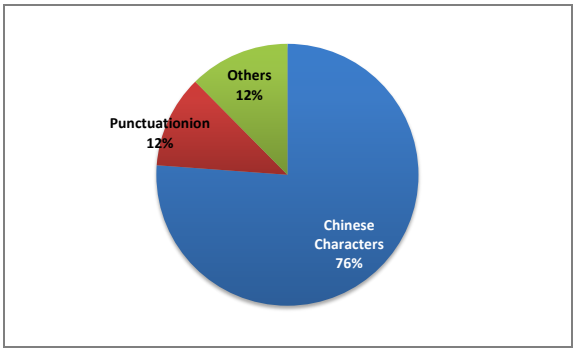

Figure 2. Distribution of different types of characters in micro-blog

Regarding the characters, we found that there are 24% of non-Chinese characters, as shown in Figure 2, which is unusual in comparing with general texts. This fully illustrates its nonstandard phenomena. Among all the non-Chinese characters, half of them are punctuations due to the redundant punctuations used in the blogs for Special expression. So we paid much more attention on those characters during the segmentation.

3 Tag Set

Character based tagging method for Chinese word segmentation, either based on maximum entropy or conditional random fields, views the Chinese word segmentation as a typical sequence labeling problem (Ratnaparkhi, 1996).

There are three kinds of schemes that are commonly used to distinguish the character position in a word, i.e., 6-tag set (Zhao, 2006), 4-tag set (Xue, 2003) and 2-tag set (Tseng, 2005). The details of those schemes are presented in Table 1. A 4-tag set is used for maximum entropy model in (Xue, 2003; Xue and Shen, 2003) and (Low et al., 2005), while a 2-tag set is used for CRFs model in (Peng et al., 2004) and (Tseng et al., 2005). Zhao (2006) extends it into 6-tag set by adding "B2" and "B3" and get a better result in SIGHAN-2006.

6-tag set	
Tag	**Description**
B	First Character (Start of Token)
*B*2	Second Character
*B*3	Third Character
M	The n_{th} Character (n = 4…*len*-1)
E	Last Character (End of Token)
S	Unit Character
4-tag set	
Tag	**Description**
B	First Character (Start of Token)
M	The n_{th} Character (n = 4…*len*-1)
E	Last Character (End of Token)
S	Unit Character
2-tag set	
Tag	**Description**
S	First Character (Start of Token)
N	Continuation

Table 1: Various tag sets used for segmentation

Based on Zhao's 6-tag set, we proposed a different tag set which is more suitable for micro-blog text segmentation. The details of the proposed 6-tag set are shown in Table 2. Our system

pays more attention to the second last character ("*E*2") of a token, instead of the second one ("*B*2"). In order to evaluate that the proposed 6-tag set is more suitable for micro-blog text, several experiments are conducted to compare between the various schemes used in Chinese segmentation, as described in Section 5.

Proposed 6-tag set	
Tag	**Description**
B	First Character (Start of Token)
*B*2	Second Character
M	The n_{th} Character (n = 3…*len*-1)
*E*2	Second Last Character
E	Last Character (End of Token)
S	Unit Character

Table 2: Proposed tag set

4 Feature Template

The selection of feature template is also an important factor. Eight feature templates are selected for this special task.

4.1 Basic features

The basic features of our word segmenter are based on the work of Zhao (2006) and Qin (2010), who achieved very good results in segmentation respectively for common texts and cross domain texts. However, some features are modified to adapt to micro-blog.

The basic feature templates we adopted are given in Table 3. *C* refers to a Chinese character. Templates (a) – (c) refer to a context of three characters (the current character and the proceeding and following characters). C_0 denotes the current character while C_{-1} and C_1 denotes its previous and next character. $C_{n\text{-i}}$ ($C_{n+\text{i}}$) denotes the character *i* positions to the right (left) of the current *n*th character. For example, given the character sequence "我成为微博达人 (I become a micro-Bardon)", when considering the character C_0 "微", C_{-1} denotes "为" and C_0C_1 denotes "微博", etc. Different from the previous work (Low, 2005), we reduced the scope of context from 5 to 3. As stated in Section 2, most tokens are 1-character words or 2-character words.

For feature (d), it checks whether C_n is a punctuation symbol (such as "?", "–", ",") or not. In our system, we did not take any special symbols like "#", "@", etc. as punctuations. Because of their specific meanings in micro-blog, for example, "#" is a start or end symbol of a topic and they are often appeared in pairs. This is the main difference in this feature.

No.	Type	Feature
a	Unigram	C_n ,n = 0,-1,1
b	Bigram	$C_n\ C_{n+1}$, n = -1,0
c	Skip	$C_{-1}\ C_1$
d	Punctuation	P_n, n = 0,-1
e	Date, Digit and Letter	$T_{-1}T_0T_1$ T_n, n = -1,-2

Table 3: Basic features (a) to (e)

Besides, we should give an explanation to feature template (e). Based on the 4-classification in (Zhao, 2006), we divided the characters into seven classes. The numbers are represented as Class 1, which both include Arabic numbers and Chinese numbers; alphabetic characters belong Class 2; dates ("日 (*day*)", "月 (*month*)", "年 (*year*)") are Class 3; pound sign (#) and at sign (@) are represented as Class 4 and 5; measure word (e.g. "个 (*ge*)" is a quantifier, which is frequently used in modern Chinese) belongs Class 6, while other characters are Class 7. For example, when considering the character "年" in the sequence "1988 年 Born", the feature $T_{-2}T_{-1}T_0T_1T_2$=11322.

Finally, we did not use the feature of "tone" (Zhao, 2006), because there is no improvement when adopting it in the domain of micro-blog.

4.2 External Dictionary

The use of external dictionary in CRFs models was firstly introduced in (Low et al., 2005). In this approach, each possible subsequence of neighboring characters around C_0 in the sentence is firstly looked up from a dictionary based on maximum match strategy. The longest one *W* in the dictionary will be chosen. Finally, the matched words will be represented in the feature templates. However, there is still a fault in the maximum matching method. For example, given the character sequence "金山石化 (*Golden Hill Petrochemical*)", taking "山 (*hill*) " as the current character C_0, the following candidates of "山 (*hill*)", "金山 (*golden hill*) ", "山石 (*hillstone*) ", "金山石 (*jin shan shi*) ", "山石化 (*shan shi hua*) " and "金山石化 (*Golden Hill Petrochemical*)" can be found. Supposed both "金山" and "山石" are the possible lexicons in the dictionary, it is hard to determine which one is better. The

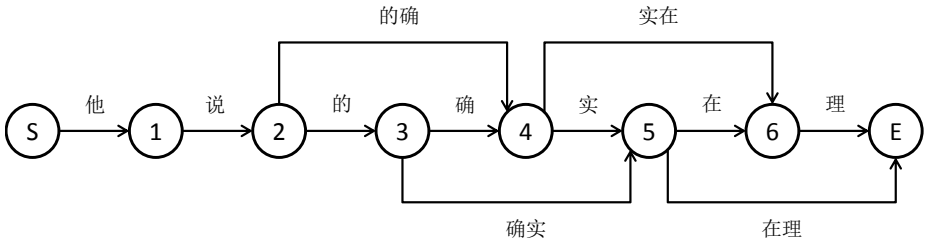

Figure 3. Graph representation of possible segmentations

problem of ambiguity often makes the method fail to determine the correct segmentation, because it does not consider the information of the whole sentence. To solve the conflict, it is used to take the candidate which gives the highest bigram probability in considering of neighboring context.

Therefore, we used the N-Shortest-Paths to fix the ambiguous problems. The details of the approach applied in Chinese word segmentation are discussed in (Leong et al. 2006).

In our system, Google Corpus is used as the external lexicon, which has the 1-gram frequency information of words. As it is the collection of words acquired from online, some popular vocabularies of micro-blog are included. The candidate that gives the highest probability is selected as the final segmentation. In this model, lexicon is represented by a vector of three features and is derived from the used dictionary, as illustrated in Table 4. L_0 is length of current matched word and t_n is position of the nth character in the current matched word. The matching of lexicon is compared against the feature set instead of the lexical item itself.

No.	Type	Feature
f	Length and Position	L_0t_0
g	Character and Position	C_nt_0 (n = -1, 0, 1)
h	Position	t_n (n = -1, 1)

Table 4: Additional features (f) and (g)

Consider the sentence "他说的确实在理 (*what he said is indeed reasonable*)", as shown in Figure 3, it gives several possible segmentation paths, i.e. "他/说/的确/实在/理", "他/说/的确/实/在/理", "他/说/的确/实/在理", "他/说/的/确实/在/理", "他/说/的/确实/在理", and "他/说/的/确/实/在/理". The frequency of each token is treated as the cost of the path. The Dijkstra's algorithm is used for finding optimal path that gives the maximum joint probability. Supposed that the path "他/说/的/确实/在理" is selected and the current character C_0 is "实", the feature templates (f) to (h) are "2E", "实 E" (n = 0) and "S" (n = -1) respectively.

In addition to the Google words, we also include the lists of Chinese idioms, four-word phrases, popular frequently used vocabularies of blogs, and Chinese reduplicating words and emoticons symbols in the proposed system.

4.3 Additional Training Corpus

Corpora in different domains have their own linguistic features and different organizations prepare training corpora in their own standards. These factors mainly limit the amount of training corpora available for micro-blog segmentation. However, the People's Daily Corpus was segmented according to the same segmentation standard (Specification for Corpus Processing at Peking University) (Yu, 2003) as the one adopted by the Bakeoff-2012 for micro-blog. Additionally, Low (2005) presented a method to reduce the OOV problems with additional training corpus. This cross-domain training method is employed in this work to overcome the problem of the micro-blog domain with limited resource.

Therefore, four months of the People's Daily Corpus (1998.01, 1998.09, 2000.03, and 2000.12) were used to extend our limited training data. Several steps are taken for adding additional training corpus:

1. Perform the training step with CRFs models using the original training corpus D_0.
2. Use the trained word segmenter to segment the four-month People's Daily corpus *D*.
3. Suppose a Chinese character *C* in *D* is assigned a boundary tag *t* by the word segmenter with probability *p*. If *t* is identical to the boundary tag of *C* in the gold-standard annotated corpus *D*, and *p* is less than some threshold μ, then the entire correct segment-

ed sentences are added into the original training corpus D_0.

4. Finally, a new word segmenter is trained using the new enlarged dataset.

5 Experiments

In order to obtain the best tag set and best feature templates, we conducted some comparisons with different settings. Due to the limitation of micro-blog corpus, we used a small corpus with 500 sentences, which is released by the CIPS-SIGHAN. 80% of it is used as training data and 20% is for testing set.

	2-Tag Set	4-tag Set	6-Tag Set	Proposed 6-Tag Set
P	0.9199	0.9275	0.9262	**0.9330**
R	0.9275	0.9315	**0.9317**	0.9281
F	0.9237	**0.9295**	0.9289	**0.9305**

Table 5: Evaluation results of various tagging schemes

Firstly, we tested our system with different tag sets. It is found that the model with 4-tag set gives a better result than that of 2-tag set and 6-tag set, while the model with the proposed 6-tag set achieves the best performance among all schemes. The results are shown in Table 4. 6-Tag Set achieves the highest recall value (0.9317), but a little lower than both the proposed tag set and 4-tag set in precision. Although the improvement of the proposed is not very clear, it is only evaluated with 500 sentences. So a good performance of the tag set still can be expected.

Based on the basic feature templates and proposed tag set, three strategies were evaluated. Firstly, there were not any additional dictionaries or corpora involved in the segmented models and this evaluation is the baseline of our experiments. And the Strategy A is applied with all the dictionaries listed in Section 4.2. Finally, both additional dictionaries and corpus were used in Strategy B. As shown in Table 5, the presence of both Strategy A and B achieve much better performance than the baseline proves that additional resources could be helpful to the segmentation for micro-blog. The recall value of Strategy A is higher than that in Strategy B, which prove that additional training corpus do well in reducing the OOV problem. However, the precision declines due to the different domain of data used for training models.

After obtaining the best strategy, a CRFs-based model was trained using the corpus with 500 sentences. And then our Chinese word segmenter was evaluated in an open track, on the test set of 5,000 micro-blog sentences which is released by the second CIPS-SIGHAN.

	Baseline	Strategy A	Strategy B
P	0.8349	0.9330	0.9293
R	0.8284	0.9281	**0.9375**
F	0.8316	**0.9305**	**0.9334**

Table 6: Evaluation results with different strategies

Table 6 shows the official bakeoff results. The column of "Proposed System" shows the precision, recall, F-measure and correct sentence (CS) of our system, which are all very closed to the values of Strategy B in Table 4. This is mainly because a suitable ratio (80% training set and 20% test set) was selected to evaluated presented approach. The third column gives the best value in each measure while Δ stands for the difference between our result and the best one. There is only a gap of 1.4% in F-measure between our system and the best one. The result shows a good performance of the segmentation in the domain of Chinese micro-blog using CRFs-based methods.

	Proposed System	Best Value	Δ
P	0.9294	0.9460	0.0166
R	0.9383	0.9496	0.0113
F	**0.9338**	0.9478	0.0140
CS (%)	37.34%	44.88%	**7.54%**

Table 7: The official bakeoff results

6 Conclusion

This article presents a CRFs-based approach for Chinese micro-blog segmentation. We not only consider the linguistic characteristics of micro-blog, but also solve the problem of small-scale training data with technique to enhance the training corpus. This is the first time to deal with Chinese micro-blog segmentation using CRFs methods. Through the comparison experiments, we found the best tag set, feature template and additional resource for this special task and achieve a good result with a very small training corpus. The performances showed that this method can do a good job of Chinese micro-blog segmentation.

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

Huang C. and Zhao H. 2007. Chinese word segmentation: A decade review. *Journal of Chinese Information Processing*. 21:8–20.

Lafferty J.D., McCallum A., and Pereira F.C.N. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*. 282–289.

Leong K.S., Wong F., Tang C.W., and Dong M.C. 2006. CSAT: A Chinese segmentation and tagging module based on the interpolated probabilistic model. *Proceedings of the Tenth International Conference on Enhancement and Promotion of Computational Methods in Engineering and Science* (*EPMESC-X*). 1092–1098.

Low J.K., Ng H.T., and Guo W. 2005. A maximum entropy approach to Chinese word segmentation. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*. 161-164.

Qin X., Zong L., Wu Y., Wan X., and Yang J. 2010. CRF-based Experiments for Cross-Domain Chinese Word Segmentation at CIPS-SIGHAN-2010. *In CLP2010*.

Ratnaparkhi A. and others. 1996. A maximum entropy model for part-of-speech tagging. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1:133–142.

Tseng H., Chang P., Andrew G., Jurafsky D., and Manning C. 2005. A conditional random field word segmenter for sighan bakeoff 2005. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*. vol. 171.

Xue N. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*. 8:29–48.

Yu S., Duan H., Zhu X., Swen B., and Chang B. 2003. Specification for corpus processing at Peking University: Word segmentation, POS tagging and phonetic notation. *Journal of Chinese Language and Computing*. 13:121–158.

Zhao H., Huang C.N., and Li M. 2006. An improved Chinese word segmentation system with conditional random field. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. vol. 1082117.

Zhao H. and Kit C. 2008. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*. 106–111.