Cascaded Chinese Weibo Segmentation Based on CRFs

Keli Zhong, Xue Zhou, Hangyu Li and Caixia Yuan
School of Computer,
Beijing University of Posts and Telecommunications,
Beijing, 100876 China
zhongkeli@139.com yuancx@bupt.edu.cn
{bupt.zhouxue, hangyuli1209}@gmail.com

Abstract

With the developments of Web2.0, the process for the data on Internet becomes necessary. This Paper reports our work for Chinese weibo segmentation in the 2012 CIPS-SIGHAN bakeoff. In order to improve the recognition accuracy of out-of-vocabulary words, we propose a cascaded model which first segments and disambiguates in-vocabulary words, then recovers out-of-vocabulary words from the fragments. Both the two process are trained by a character-based CRFs model with user-edited external vocabulary. The final performance on the test data shows that our system achieves a promising result.

1 Introduction

Since there are no spaces in Chinese sentences, Chinese word segmentation becomes a vital and fundamental task in Chinese language processing. Many approaches have been implemented in Chinese segmentation, including simple Forward Maximum Match (FMM), statistic based methods like Hidden Markov model, conditional random fields model, along with other learning models(Sproat et al., 1996; Xue and Shen, 2003; Tseng et al., 2005; Song et al., 2006). The main problems of segmentation are word boundary ambiguities and out-of-vocabulary (OOV) word recognition while many researchers have been working on them (Wang et al., 2008; Xu et al., 2010; Koichi et al., 2002).

Recent developments in Web 2.0 have heightened the need for Web text processing (Downey et al., 2007), which makes the problems above more prominent. Being different from traditional texts like news reports and literary works, Web texts like microblogs, tweets tend to be more oral, casual, and have plenty of catchwords, typos and OOVs in them, which bring much challenge to language understanding. For example, “Gelivable” is a Chinglish word coined by Chinese people stands for the word “给力” (awesome), which is a popular Chinese catchword in Web texts. Some users leave the typos deliberately to unique and individual. For instance, “碎叫” (shleep) stands for “睡觉” (sleep). Although human people would understand the meaning of this piece of Chinese tweets, segmenter based on dictionary may never understand how it went wrong (Bian, 2006). In the next place, thousands of new words emerge from current event, social phenomena or even actors’ lines. For instance, “喵星人” and “基友” are the new words that emerged from Internet not long ago, which stands for “cat” and “gay friend” respectively. And the sentence patterns like “神马都是浮云” (Everything is nothing.) a prevalent slogan of many people on the Internet. These phenomena exemplified above exacerbate the OOV problem (Xu et al., 2008). Take weibo, a popular Chinese MicroBlog, for example, within a piece of text restricted to 140 Chinese characters, there are 21.7(15.5%) OOV words on average. Finally, the structure of MicroBlog sentences prone to be simple, elliptical, non-predicate and incompleteness. Some of the sentences are mixed with words in foreign languages and emoticons (like :), ToT). Hence the segmenter based on linguistic knowledge would not be efficient enough (Li et al., 1998).

In order to better solve the Web text problems, we propose an efficient Chinese Web text segmentation model based on CRF model with a user-edited dictionary. Specifically, we first conduct a coarse-grained segment for input Web text, then refine the results through models learned from new word vocabulary provided by users.

Following sections describe in detail the proposed method and its results on the SIGHAN 2012 Chinese MicroBlog segmentation task. In sec-

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 69–73,
Tianjin, China, 20-21 DEC. 2012

tion 2 to 4, we introduce the main idea of our method. Section 5 gives experiment results and related analysis, which proves the effectiveness of our model. Section 6 addresses the future work.

2 Our Method

We use a CRF model[1] based on character to implement Chinese MicroBlog segmentation. Following the work of (Qin et al., 2008), we use a BIO style to formulate the word segmentation into a sequence learning task. We define 6 tags in order to distinguish different roles of characters more accurately. The 6 tags and their descriptions are denoted in Table 1.

label	meaning
B	the start of word
E	the end of word
M1	the 1st character of a word
M2	the 2nd character of a word
M	other characters of a word
S	single-character word

Table 1: Labels and their descriptions.

2.1 Basic procedure

The processing of word segmentation is shown in Fig.1.

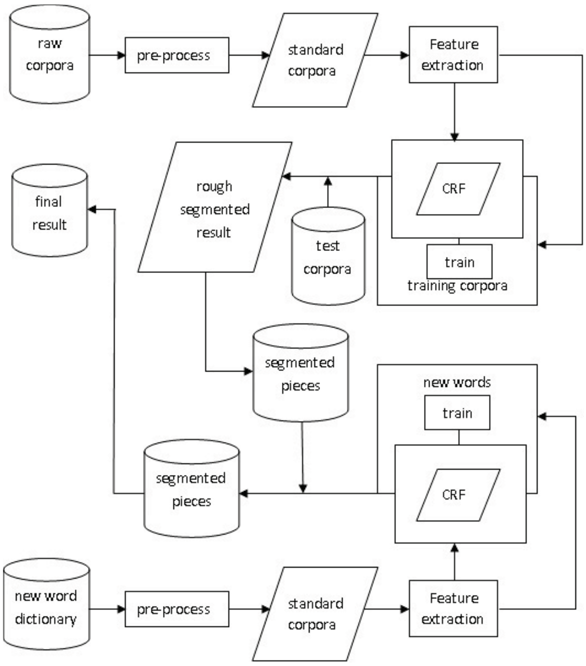

Figure 1: Framework of our segmentation model.

We use 6 months of PKU people's daily data in year 2000 (Yu et al., 2002) as training corpora, in which the sentences in paragraphs have been segmented into words by spaces. In order to construct the character-level based segmenter, we transform the original corpora into the sequential form representing by 6 labels shown in Table 1, and each line only includes one character and its corresponding label.

2.2 Feature selection

As the feature has great influence on the segmentation result, hence what kinds of features should be selected is the key to our task.

We design two classes of feature templates: (1) Unigram feature template, (2) Bigram feature template. Particularly, the Unigram and Bigram that we use here are the count for label that exist in feature, not the count for the character that exist in feature. From this point of view, the meaning of Unigram and Bigram are no longer the same with other existing papers (Jurafsky et al., 2007; Chen et al., 2005).

For character level based Chinese segmentation, character feature is the major concern here. According to the distance from current character, we have features 1-5 respectively as depicted in Table2., and these features belong to Unigram feature templates. The context characters are confined to be two characters around the character at hand. These template features would expand into thousands of features while CRF training, and each feature corresponds to a feature function, which are vital to CRFs model's learning process.Besides the context characters of the current, we also take their bigram sequence into account when designing feature template, which corresponds to features 6-8 in Table 2.

Another critical feature for character tag labeling is the type of the character at hand. We distinguish the character with 4 types including Chinese character, English character, number, punctuation, and add the character type into the feature template as a Unigram feature, which are represented as feature 9 and 10 in Table 2.

The feature templates in Table 2 are basic feature templates designed from character position and their types.

In order to exploit more deliberate properties of how likely a sequence of characters being a word, we investigate the probability of two adja-

[1] CRF++0.54, http://crfpp.sourceforge.net/

No.	feature	feature description
1	C_{-2}	the 2nd lefthand character of C_0
2	C_{-1}	the 1st lefthand character of C_0
3	C_0	current character
4	C_1	the 1st righthand character of C_0
5	C_2	the 2nd righthand character of C_0
6	$C_{-1}C_0$	sequence of C_{-1} and C_0
7	C_0C_1	sequence of C_0 and C_1
8	$C_{-1}C_1$	sequence of C_{-1} and C_1
9	T_0	type of C_0
10	$T_{-1}T_1$	type of C_{-1} and C_0

Table 2: Context features and character type features we used.

cent characters forming a word, that is the cohesion of two characters on word level. Consider the current character C_0, and the probability of being a word with the lefthand character C_{-1} can be computed as:

$$P_{-1,0} = \frac{W(C_{-1}C_0)}{Count(C_{-1}C_0)} \qquad (1)$$

in which $W(C_{-1}C_0)$ represents the amount of $C_{-1}C_0$ as a word that exist in the training corpora, and $Count(C_{-1}C_0)$ represents the amount of $C_{-1}C_0$ that appear in a sentence.

For instance:

1) 中国 的 士兵 (China 's soldier)

2) 中国 的士 (China taxi)

W("的士")=1, while Count("的士")=2.

We used 3 levels to represent the cohesion of two characters, and add them into the feature template as uniform features as is shown in Table 3.

No.	feature	feature description
11	S	$P_{-1,0} < 0.2$ the probability of character C_i and C_j being a word is low
12	NS	$P_{-1,0} > 0.75$ the probability of character C_i and C_j of being a word is high
13	N	$0.2 \leq P_{c_ic_j} \leq 0.75$

Table 3: Character cohesion features.

Finally, 13 features are used for CRF model training, including basic Unigram features in Table 2 and the being-a-word features in Table 3. We train a CRFs model using feature templates listed in Table 2 and 3. This model is then used for the first-round segmentation which yields a word and fragment sequence. Our experiment results depicted later show that this model achieves high performance for in-vocabulary words, while most out-of-vocabulary words are segmented as character fragments. Thus we will investigate the improved model for recognizing such OOV words.

3 User Editable Dicitionary

In order to make model exploit external knowledge about OOV words and easily adapt to different user demand, we design a plug-in user dictionary, which is used to refine the segmentation model trained in Section 2. For SIGHAN MicroBlog segmentation task, we collect 278,060 words from Sogou word bank[2]. Due to MicroBlogs are the epitome of people's life, so the new words we collected from Sogou word bank are close to the type that used in MicroBlogs, which consists of newly invented words on the Internet, dishes' name, celebrities' name, online shopping words (product names, brands, etc.) and others that is related with people's daily life.

4 Refined OOV Word Recognition Model

Quite amount of OOV would emerge during the MicroBlog segmentation. Based on the vocabulary collected in Section 3, we refine the segmentation results yielded in the first-round segmentation depicted in Section 2. The refined model is trained on the user-edited vocabulary and is to used for a second-round segmentation. Each word is viewed as a training sample. Besides feature templates listed in Table 2, we design several new features for the refined model which is described in Table 4.

No.	feature	feature description
14	$C_{-2}C_{-1}$	sequence of C_{-2} and C_{-1}
15	C_1C_2	sequence of C_1 and C_2
16	$C_{-1}C_0P_{-1,0}$	sequence of C_{-1}, C_0 and $P_{-1,0}$

Table 4: New context features and character type features in Model 1, while other features are already shown in Table 2.

The function of Model 1 is to segment test corpora for the first time. And the features it uses is shown in Table 4.

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

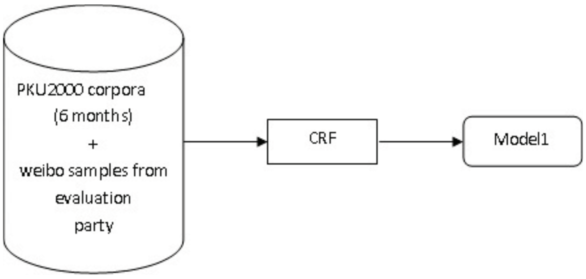

Figure 2: Training process.

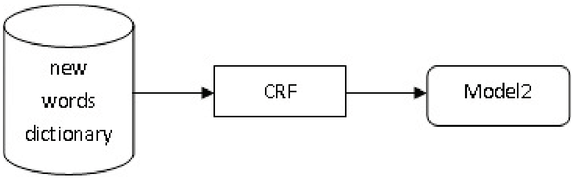

Figure 3: New words training process.

Model2 is trained using new words from user-editable dictionary. Each word is viewed as a training sample and features are extracted according to feature templates shown in Table 2.

The whole structure of the Model is shown in Fig.4.

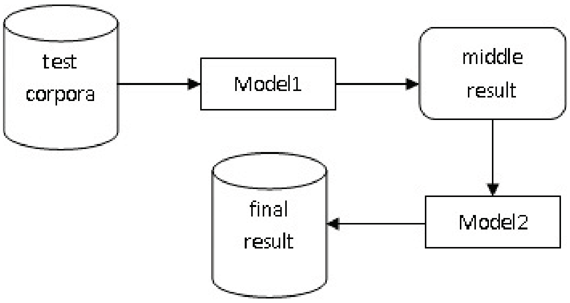

Figure 4: Model predicting process.

5 Experiment

We design 4 experiments to test contributions of different features, and the effectiveness of our proposed model. The comparison of the result is made and shown in Table 5. The base training data we used is 6 months of People Daily in year 2000 built by Peking University (Yu et al., 2002). Experiment 1 uses features listed in Table 2, and experiment 2 adds features listed in Table 3. The test data of experiment 1 and 2 are MicroBlog training samples. In experiment 3, we add half of training samples of SIGHAN, while the rest half is used for test data. Experiment 4 uses base training data and all the MicroBlog training samples provided by SIGHAN, and is evaluated on the test data provided by SIGHAN. From the results of experiment 1 and 2,we can observe that adding cohesion ratio of two characters listed in Table 3 achieves a higher accuracy. The cohesion ratio of characters is a strong sign for them being a word or not. From the result of experiment 2 and 3, we learn that to achieve a better performance in mirco-blog segmentation, more corpora or features that embody the characteristics of MicroBlog is vitally needed.

No.	**1**	**2**
Training data	PKU	PKU
Features	Feature1-10	Feature1-13
test data	Weibo	Weibo
Recall	0.897	0.925
Precision	0.915	0.927
F1 measure	0.906	0.926
No.	**3**	**4**
Training data	PKU+1/2 Weibo	PKU+Weibo
Features	Feature1-16	Feature1-16
test data	1/2 Weibo	test data
Recall	0.928	0.932
Precision	0.935	0.935
F1 measure	0.932	0.933

Table 5: Experiment results comparison in different data settings, in which Weibo stands for Weibo samples and test data is the given Weibo test data.

No.	**5**	**6**
Training data	PKU	PKU
Features	Feature1-10	Feature1-13
Test data	1 month of PKU	
Recall	0.951	0.962
Precision	0.967	0.973
F1 measure	0.959	0.967
OOV Recall	0.847	0.860
IIV	0.957	0.968

Table 6: Feature used here is the cohesion ratio feature.

Table 6 demonstrates test result on the text from a month of People Daily. We can observe that F score is improved to 0.973 after adding cohesion features of characters, which is consistent with the observation on MicroBlog data in Experiment 2.

6 Future Work

In this paper, we try to implement micro blog segmentation, finding out the cohesion ratio of characters is a crucial feature for them being a word or not. Meanwhile, the user-editable vocabulary can not only provide flexibility for domain adaptation, but also be used as external knowledge to improve OOV recognition rate.

The current system is far from our goal, and there still has a lot of work to do:

(1)We use PKU corpora mainly for training, with a little corpora from micro blogs. Sufficient corpora is needed to extract the cohesion ratio features in MicroBlog. So active-learning (Baldridge et al., 2004; KimS et al., 2006) can be implemented here to achieve better performance through iterative training on relative small scale of manually labeled data.

(2)A method that can express the cohesion ratio feature between characters more efficiently is required. In this paper, we just calculated the probability of being a word between characters in a simple statistical way. Therefore another direction of future work is to explore the relationship between words to reflect the relationship between characters.

Acknowledgement

This research has been partially supported by the National Science Foundation of China (NO. NSFC61202248). We also thank Xiaojie Wang and Huixing Jiang for useful discussion of this work.

References

J. Wang, J. Liu, P. Zhang. 2008. *Chinese Word Sense Disambiguation with PageRank and HowNet.* In Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing.

X. Xu, M. Zhu, X. Fet, J. Zhu. 2010. *High OOV-Recall Chinese Word Segmenter.* In CIPS-SIGHAN Joint Conference on Chinese Language Processing.

G. Bian. 2006. *Chinese Word Segmentation using Various Dictionaries.* In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing.

Z. Xu, X. Qian, Y. Zhang, Y. Zhou. 2008. *CRF-based Hybrid Model for Word Segmentation, NER and even POS Tagging.* In Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing.

H. Li, B. Yuan. 1998. *Chinese Word Segmentation.* In Language, Information and Computation(PACLIC12), 19-20 Feb, 1998, 212-217.

Y. Qin, X. Wang, Y. Zhong. 2008. *Cascade Identification of Chinese Chunks.* In the Journal of Beijing University of Posts and Telecommunications.

R. Sproat, C. Shin, W. A. Gale, and N. Chang. 1996. *A stochastic finite-state word-segmentation algorithm for Chinese.* In Computational Linguistic, 22(3):337-404.

N. Xue, L. Shen. 2003. *Chinese word segmentation as lmr tagging.* In Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan.

H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. D. Manning. 2005. *Conditional random field word segmenter.* In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.

D. Jurafsky, James H. Martin 2007. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Prentice Hall.

S. Yu, H. Duan, X. Zhu, B. Sun. 2002. *The Basic Processing of Contemporary Chinese Corpus at Peking University SPECIFICATION.* In Journal of Chinese Information Processing. Vol.15 No.5.

K. Tangigaki, H. Yamamoto, Y. Sagisaka. 2000. *A Hierarchical Language Model Incorporating Class-Dependent Word Models For OOV Words Recognition.* In the Proceedings of the 6th International Conference on Spoken Language Processing.

D. Downey, M. Broadhead, O. Etzioni. 2007. *Locating Complex Named Entities in Web Text.* In IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence.

D. Song, Anoop Sarkar. 2006. *Voting between Dictionary-Based and Subword Tagging Models for Chinese Word Segmentation.* In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing.

A. Chen, Y. Zhou, A. Zhang, G. Sun. 2005. *Unigram language model for Chinese word segmentation.* In the Fourth SIGHAN Workshop on Chinese Language Processing (Second International Chinese Segmentation Bakeoff)

J. Baldridge, M. Osborne. 2004. *Active learning and the total cost of annotation.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP):9 – 16. ACL Press. 2004.

S. Kim, Y. Song, K. Kim, J. Cha, G. G. Lee. 2006. *MMR-based active machine learning for bio named entity recognition.* In Proceedings of Human Language Technology and the North American Association for Computational Linguistics (HLT-NAACL):69 – 72.ACL Press. 2006.

L. Zhou 2007. *The Recognition Method of Unknown Chinese Words Based on Fragments Segmentation.* In the Journal of Changshu Insititue of Technology(Natural Sciences). Vol 2.