An Improved Chinese Word Segmentation System with Conditional Random Field

Hai Zhao, Chang-Ning Huang and Mu Li
Microsoft Research Asia,
49, Zhichun Road, Haidian District,
Beijing, P. R. China, 100080
Email: {f-hzhao,cnhuang}@msrchina.research.microsoft.com,
muli@microsoft.com

Abstract

In this paper, we describe a Chinese word segmentation system that we developed for the Third SIGHAN Chinese Language Processing Bakeoff (Bakeoff-2006). We took part in six tracks, namely the closed and open track on three corpora, Academia Sinica (CKIP), City University of Hong Kong (CityU), and University of Pennsylvania/University of Colorado (UPUC). Based on a conditional random field based approach, our word segmenter achieved the highest F measures in four tracks, and the third highest in the other two tracks. We found that the use of a 6-tag set, tone feature of Chinese character and assistant segmenters trained on other corpora further improve Chinese word segmentation performance.

1 Introduction

Conditional random field (CRF) is a statistical sequence modeling framework first introduced into language processing in (Lafferty et al., 2001). Work by Peng et al. first used this framework for Chinese word segmentation by treating it as a binary decision task, such that each Chinese character is labeled either as the beginning of a word or not (Peng et al., 2004).

Since two participants, Ng and Tseng in Bakeoff-2005, gave the best results in almost all test corpora (Low et al., 2005), (Tseng et al., 2005), we continue to improve CRF-based tagging method of Chinese word segmentation on their track. Our implementation used CRF++ package Version 0.41[1] by Taku Kudo.

In our system, a Chinese character is labeled by a tag which stands for its position in the Chinese word that the character belongs to. We handle closed test and open test in the same way. The difference is that those features concerned with additional linguistic resources are added in the feature set of closed test to produce the feature set used in open test.

2 Tag Set Selection

Character based tagging method for Chinese word segmentation, either based on maximum entropy or CRF, views Chinese word segmentation as a label tagging problem, which is described in detail in (Ratnaparkhi, 1996).

The probability model and corresponding feature function is defined over the set $H \times T$, where H is the set of possible contexts (or any predefined condition) and T is the set of possible tags. Generally, a feature function can be defined as follows,

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

where $h_i \in H$ and $t_j \in T$.

For convenience, features are generally organized into some groups, which used to be called feature templates. For example, a bigram feature template C_1 stands for the next character occurring in the corpus after each character.

[1] http://chasen.org/ taku/software/CRF++/

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 162–165,
Sydney, July 2006. ©2006 Association for Computational Linguistics

As for tag set, there are two kinds of schemes that are used to distinguish the character position in a word in the previous work, i.e., 4-tag set and 2-tag set. The details are listed in Table 1. Notice Xue and Ng use a 4-tag set in maximum entropy model. While Peng and Tseng used a 2-tag set in CRF model.

Table 1: Tag sets in Chinese word segmentation in the previous work

4-tag set Ng/(Xue)		2-tag set Peng/Tseng	
Function	Tag	Function	Tag
begin	B(LL)	start	Start
middle	M(MM)	continuation	NoStart
end	E(RR)		
single	S(LR)		

Generally speaking, activated feature functions in practice like (1) are determined by both feature template and tag set. In the existing work, tag set is specified aforehand. To effectively perform tagging for those long words, we extend the 4-tag set of Ng/Xue into a 6-tag set. Two tags, 'B2' and 'B3', are added into a 4-tag set to form a 6-tag set, each additional tag stands for the second and the third character position in a Chinese word, respectively.

3 Feature Templates for Closed Test

The feature template set we selected for closed test is shown in Table 2. We give an explanation to feature template (e) and (f).

Feature template (e) is improved from the corresponding one in (Low et al., 2005). $T_n, n = -1, 0, 1$ stands for predefined class. There are four classes defined: numbers represent class 1, those characters whose meanings are dates represent class 2, English letters represent class 3, and other characters represent class 4.

As for feature template (f), $To(C_0)$ stands for the tone of current character. There are five possible types of tones for Chinese characters in mandarin, we just assign 0, 1, 2, 3 and 4 as feature values. For example, consider some characters, '中','国','很','大' and '吗', $To(C_0)$ is 1, 2, 3, 4 and 0, respectively.

4 Feature Templates for Open Test

In open test, we use two kinds of additional feature templates to improve the performance upon closed test.

4.1 External Dictionary

This method was firstly introduced in (Low et al., 2005). We continue to use the online dictionary from Peking University downloadable from the Internet [2], consisting of about 108,000 words of length one to four characters. If there is some sequence of neighboring characters around C_0 in the sentence that matches a word in this dictionary, then we greedily choose the longest such matching word W in the dictionary. The following features derived from the dictionary are added:

(g) Lt_0

(h) $C_n t_0 (n = -1, 0, 1)$

where t_0 is the boundary tag of C_0 in W, and L is the number of characters in W.

4.2 Assistant Segmenter

We observed that although there exists different segmentation standards, most words are still segmented in the same way according to different segmentation standards. Thus, though those segmenters trained on different corpora will give some different segmentation results, they agree on most cases. In fact, we find that it is feasible to customize a pre-defined standard into any other standards with TBL method in (Gao, 2005). And it is also valuable to incorporate different segmenters into one segmenter based on the current standard. For convenience, we call the segmenter subjected to the current standard main segmenter, and the other assistant segmenters.

A feature template will be added for a assistant segmenter:

(i) $t(C_0)$

[2] http://ccl.pku.edu.cn/doubtfire/Course/Chinese %20Information%20Processing/Source Code/ Chapter 8/Lexicon full 2000.zip

Table 2: Feature templates

Code	Type	Feature	Function
a	Unigram	$C_n, n = -1, 0, 1$	The previous (current, next) character
b	Bigram	$C_nC_{n+1}, n = -1, 0$	The previous (next) character and current character
c	Jump	$C_{-1}C_1$	The previous character and next character
d	Punctuation	$Pu(C_0)$	Current character is a punctuation or not
e	Date, Digital and Letter	$T_{-1}T_0T_1$	Types of previous, current and next character
f	Tone	$To(C_0)$	Tone of current character

where $t(C_0)$ is the output tag of the assistant segmenter for the current character C_0. For example, consider character sequence, '我们都是中国人', an assistant segmenter gives the tag sequence 'BESSBES' according to its output segmentation, then $t(C_0)$ by this assistant segmenter is 'B', 'E', 'S', 'S', 'B', 'E', and 'S' for each current character, respectively.

In our system, we integrate all other segmenters that are trained on all corpora from Bakeoff-2003, 2005 and 2006 with the feature set used in closed test. The segmenter, MSRSeg, described in (Gao, 2003) is also integrated, too.

Our assistant segmenter method is more convenient compared to the additional training corpus method in (Low et al., 2005). Firstly, the performance of additional corpus method depends on the performance of the trained segmenter that carries out the corpus extraction task. If the segmenter is not well-trained, then it cannot effectively extract the most wanted additional corpus to some extent. Secondly, additional corpus method is only able to integrate useful corpus, but it cannot integrate a well-trained segmenter while the corpus cannot be accessed. Finally, additional corpus method is very difficult to use in CRF model, the reason is that the increase of corpus can lead to a dramatic increase of memory and time consuming in this case, while assistant segmenters just lead to little increase of memory and time consuming in training.

It is more interesting that we may also regard the external dictionary method as another assistant segmenter in some degree, that is, a maximal matching segmenter with the specified external dictionary. Thus, all of our additional methods in open test can be viewed as assistant segmenter ones.

5 Evaluation Results

We took part in six segmentation tasks in Bakeoff-2006, namely the closed and open track on three corpora, Academia Sinica (CKIP), City University of Hong Kong (CityU), and University of Pennsylvania/University of Colorado (UPUC).

The comparison between our official results and best results in Bakeoff-2006 are shown in Table 3.

Our system achieved the highest F measures in four tracks, and the third highest in the other two tracks. However, a format error unfortunately occurred in the open test of UPUC corpus as we submitted our final results. Thus an abnormal result in this task is obtained, the official F measure in open test is the same as that in closed test. We get the actual F measure of 0.953 after the bug is fixed.

The results in MSRA corpus from our evaluation are listed in Table 4.

Table 4: Comparison between our results and best results of Bakeoff-2006 on MSRA corpora

Type	F Measures	
	Bakeoff-2006	Ours
Closed Test	0.963	0.970
Open Test	0.979	0.982

The sizes of training corpora (in number of characters) and difference of our results between open

Table 3: Comparison between our official results and best results of Bakeoff-2006

Type	Participant	F measures on Different Corpora		
		CKIP	CityU	UPUC
Closed Test	Best results of Bakeoff-2006	0.958	0.972	0.933
	Our results	0.958	0.971	0.933
Open Test	Best results of Bakeoff-2006	0.959	0.977	0.944
	Our results	0.959	0.977	0.933

test and closed test are shown in Table 5. This illustrates how much assistant segmenters improve the performance of segmentation in different sizes of training corpora, also, this shows how the size of training corpus affects the improvement contributed by assistant segmenters.

Table 5: The sizes of training corpora and difference of our results between open test and closed test

	F Measures			
	CKIP	CityU	MSRA	UPUC
$F_{open} - F_{closed}$	0.001	0.006	0.012	0.020
Size of training corpus	9M	2.9M	2.3M	0.88M

References

Jianfeng Gao, Mu Li, Andi Wu and Chang-Ning Huang. 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computational Linguistics*, Vol. 31(4): 531-574.

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

Nianwen Xue and Libin Shen. 2003. Chinese Word Segmentation as LMR Tagging. In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL'03*, 176-179. Sapporo, Japan

Richard Sproat and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. *The Second SIGHAN Workshop on Chinese Language Processing*, 133-143. Sapporo, Japan.

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 123-133. Jeju Island, Korea.

Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo 2005. A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 161-164. Jeju Island, Korea.

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

Fuchun Peng, Fangfang Feng and Andrew McCallum. 2004. Chinese Segmentation and New Word Detection Using Conditional Random Fields. In *COLING 2004*, 562-568. August 23-27, 2004, Geneva, Switzerland

John Lafferty, 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

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

Jianfeng Gao, Mu Li, and Chang-Ning Huang. 2003. Improved Source-Channel Models for Chinese Word Segmentation. In *Proceedings 41nd Annual Meeting of the Association for Computational Linguistics*, 272-279. Sapporo, Japan, July 7-12, 2003.