Word Boundary Token Model for the SIGHAN Bakeoff 2007

Tsai Jia-Lin
Department of Information Management
Tungnan University
Taipei 222, Taiwan
tsaijl@mail.tnu.edu.tw

Abstract

This paper describes a Chinese word segmentation system based on word boundary token model and triple template matching model for extracting unknown words; and word support model for resolving segmentation ambiguity.

1 Introduction

In the SIGHAN bakeoff 2007, we participated in the CKIP and the CityU closed tasks. Our Chinese word segmentation system is based on three models: (a) word boundary token (WBT) model and (b) triple context matching model for unknown word extraction, and (c) word support model for segmentation disambiguation. Since the word support model and triple context matching model have been proposed in our previous work (Tsai, 2005, 2006a and 2006b) at the SIGHAN bakeoff 2005 (Thomas, 2005) and 2006 (Levow, 2006), the major descriptions of this paper is on the WBT model.

The remainder of this paper is arranged as follows. In Section 2, we present the WBT model for extracting words from each Chinese sentence. Scored results and analyses of our CWS system are presented in Section 3. Finally, in Section 4, we present our conclusion and discuss the direction of future research.

2 Word Boundary Token Model

To develop the WBT model, first, we define word boundary token. Second, we give definition and computation of the WBT probability and the WBT frequency for a given corpus. Finally, algorithm of our WBT model for word extraction is given.

2.1 Types of Word Boundary Token

We classify WBT into three types: left, right and bi-direction. The left and right word boundary (WB) tokens are the immediately preceding word and the following word of a word in a Chinese sentence, respectively. Suppose $W_1W_2W_3$ is a Chinese sentence comprised of three Chinese words W_1, W_2 and W_3. To this case, W_1 and W_3 are the left and the right WB tokens of W_2, respectively. On the other hand, those words that can simultaneously be left and right WB tokens of a word in corpus are defined as bi-direction WB tokens. Suppose $W_4W_2W_1$ is a Chinese sentence comprised of three Chinese words W_4, W_2 and W_1. Following the above cases, W_1 can be a bi-direction WB token for W_2. Table 1 is the Top 5 left, right and bi-direction WB tokens derived by the Academia Sinica (AS) corpus (CKIP, 1995 and 1996). From Table 1, the Top 1 left , right and bi-direction WB tokens is "的(of)."

	Left	Right	Bi-Direction
Top1	的(of)	的(of)	的(of)
Top2	是(is)	是(is)	是(is)
Top3	在(at)	了(already)	在(at)
Top4	個(a)	在(at)	了(already)
Top5	有(has)	一(one)	與(and)

Table 1. Top 5 left, right and bi-direction WB tokens derived from the AS corpus

2.2 WBT Frequency and WBT Probability

We first give the computation of WBT frequency, then, the computation of WBT probability.

(1) **WBT frequency**: we use WBT_F(*string*, *WBT*, *L*/*R*) as the function of WBT frequency, where *string* is a n-char string containing n Chinese characters, *WBT* is a word boundary token, and *L*/*R* indicates to compute left or right WBT frequency. Now, take WBT_F("我們(we)", "的(of)", *L*) as example. First, we submit the query "的我們" to system corpus. Second, set the number of sentences including this query is the WBT_F("我們(we)", "的(of), L).

(2) **WBT Probability**: we use WBT_P(*string1*, *string2*, *WBT, L/R*) as the function of WBT probability, where *string1* and *string2* are two n-char strings, *WBT* is a word boundary token, and *L*/*R* indicates to compute left or right WBT probability. The equations of left and the right WBT probability are:

$$\text{WBT\_P}(string1, string2, WBT, L) = \frac{\text{WBT\_F}(string1, WBT, L)}{(\text{WBT\_F}(string1, WBT, L)+\text{WBT\_F}(string2, WBT, L))} \quad (1)$$

$$\text{WBT\_P}(string1, string2, WBT, R) = \frac{\text{WBT\_F}(string1, WBT, R)}{(\text{WBT\_F}(string1, WBT, R)+\text{WBT\_F}(string2, WBT, R))} \quad (2)$$

2.3 Algorithm of WBT Model

We use WBTM(*n*, *WBT*, *threshold_p*, *threshold_f*) as the function of the WBT model, where *n* is the window size, *threshold_p* is the threshold value of WBT probability and *threshold_f* is the threshold value of WBT frequency. The algorithm of our WBT model applied to extract words from a given Chinese sentence is as follows:

Step 1. INPUT:
n, *WBT*, *threshold_p* and *threshold_f*;
Step 2. IF sentence length is less or equal to *n* THEN GOTO Step 4;
Step 3.
SET loopCount to one
REPEAT
COMBINE the characters of sentence between loopCount_{th} and $(\text{loopCount} + n - 1)_{th}$ to be a *string_a*
COMBINE the characters of sentence between $(\text{loopCount}+1)_{th}$ and $(\text{loopCount} + n)_{th}$ to be a *string_b*
IF WBT_P(*string_a*, *string_b*, *WBT*, *L*) ≥ *threshold_p* AND
WBT_P(*string_a*, *string_b*, *WBT*, *R*) ≥ *threshold_p* AND
WBT_F(*string_a*, *WBT*, *L*) ≥ *threshold_f* AND
WBT_F(*string_a*, *WBT*, *R*) ≥ *threshold_f* THEN SET *string_a* is as word
ENDIF
IF WBT_P(*string_b*, *string_a*, *WBT*, *L*) ≥ *threshold_p* AND
WBT_P(*string_b*, *string_a*, *WBT*, *R*) ≥ *threshold_p* AND
WBT_F(*string_b*, *WBT*, *L*) ≥ *threshold_f* AND
WBT_F(*string_b*, *WBT*, *R*) ≥ *threshold_f* THEN SET *string_b* to a word
ENDIF
INCREMENT loopCount
UNTIL loopCount > sentence length $- n$
Step 4. END.

loopCount is 1
string_a = 廣義; *string_b* = 義地
WBT_F(*string_a*, "的", L) = 0
WBT_F(*string_a*, "的", R) = 7
WBT_F(*string_b*, "的", L) = 0
WBT_F(*string_b*, "的", R) = 0
WBT_P(*string_a*, *string_b*, "的", *L*) = 0
WBT_P(*string_a*, *string_b*, "的", *R*) = 1
WBT_P(*string_b*, *string_a*, "的", *L*) = 0
WBT_P(*string_b*, *string_a*, "的", *R*) = 0
SET 廣義 to a word
loopCount is 2
string_a = 義地; *string_b* = 地說
WBT_F(*string_a*, "的", L) = 0
WBT_F(*string_a*, "的", R) = 0
WBT_F(*string_b*, "的", L) = 0
WBT_F(*string_b*, "的", R) = 0
WBT_P(*string_a*, *string_b*, "的", *L*) = 0
WBT_P(*string_a*, *string_b*, "的", *R*) = 0
WBT_P(*string_b*, *string_a*, "的", *L*) = 0
WBT_P(*string_b*, *string_a*, "的", *R*) = 0

Table 2. An example of applying WBTM(2, "的", 0.95, 1) to extract word "廣義" from the Chinese sentence "廣義地說"

Table 2 is an example of applying WBTM(2, "的", 0.95, 1) to extract words from the Chinese sentence "廣義地說" by the AS corpus

3 Evaluation

In the SIGHAN Bakeoff 2007, there are five training corpus for word segmentation (WS) task: AS (Academia Sinica), CityU (City University of Hong Kong) are traditional Chinese corpus; CTB (University of Colorado, United States), NCC (State Language Commission of P.R.C., Beijing) and SXU (Shanxi University, Taiyuan) are simplified Chinese corpus. For each corpus, there are closed and open tasks. In this Bakeoff, we attend the AS (Academia Sinica) and CityU (City University of Hong Kong) closed WS tasks. Tables 3 and 4 show the details of CKIP and CityU tasks. From Table 3, it indicates that the CKIP should be a 10-folds design. From Table 4, it indicates that the CityU should be a 5-folds design.

	Training	Testing
Sentence	95,303	10,834
Wordlist	48,114	14,662

Table 3. The details of CKIP WS task

	Training	Testing
Sentence	36,227	8,093
Wordlist	43,639	23,303

Table 4. The details of CityU WS task

3.1 Our CWS System

The major steps of our CWS system with word boundary token model, triple context matching model and word support model are as below:

Step 0. Combine training corpus and testing corpus as system corpus;

Step 1. Generate the BMM segmentation for the given Chinese sentence by system dictionary;

Step 2. Use WBT model with system corpus to extract 2-char, 3-char and 4-char words from the given Chinese sentence, where *WBT* is set to "的," "是," "在," "了," "與," *threshold_p* is set to 0.95 and *threshold_f* is set to 1;

Step 3. Use TCT (triple context template) matching model to extract 2-char, 3-char and 4-char words from the segmented Chinese sentence of Step 1. The details of TCT matching model can be found in (Tsai, 2005);

Step 4. Add the found words of Steps 2 and 3 into system dictionary;

Step 5. Generate the BMM segmentation for the given Chinese sentence by system dictionary;

Step 6. Use word support model to resolve **Overlap Ambiguity (**OA) and **Combination Ambiguity (**CA) problems for the BMM segmentation of Step 5.

3.2 Bakeoff Scored Results

Table 5 is the comparison of scored results between our CWS and the SIGHAN Bakeoff 2007 baseline system for the CKIP closed WS task by the SIGHAN Bakeoff 2007. Table 6 is the comparison between our CWS and the SIGHAN Bakeoff 2007 baseline system for the CityU closed WS task by the SIGHAN Bakeoff 2007.

	Baseline	Our CWS	Increase
R	0.8978	0.915	0.0172
P	0.8232	0.9001	0.0769
F	0.8589	0.9075	0.0486

Table 5. The comparison of scored results between our CWS system and the SIGHAN Bakeoff 2007 baseline system for the CKIP closed WS task

	Baseline	Our CWS	Increase
R	0.9006	0.9191	0.0185
P	0.8225	0.9014	0.0789
F	0.8598	0.9102	0.0504

Table 6. The comparison of scored results between our CWS system and the SIGHAN Bakeoff 2007 baseline system for the CityU closed WS task

From Tables 5 and 6, it shows the major improvement of our CWS for the baseline system is on the precision of word segmentation. That is to say, the major target system for improving our CWS system is the unknown word extraction system, i.e. the word boundary model and the triple context template matching model.

3.3 Analysis

Table 7 is the coverage of 2-char, 3-char, 4-char and great than 4-char error words extracting by our CWS for the CKIP and the CityU closed WS tasks.

	Coverage (%)			
	2-char	3-char	4-char	> 4-char
CKIP	68%	24%	4%	4%
CityU	78%	19%	2%	1%
Total	75%	21%	3%	1%

Table 7. The coverage of 2-char, 3-char, 4-char and great than 4-char error words extracting by our CWS for the CKIP and the CityU closed WS tasks

From Table 7, it shows the major n-char unknown word extraction for improving our CWS system is on 2-char unknown word extraction. It is because that the total coverage of 2-char word errors extraction of our CWS system for the CKIP and the CityU WS tasks is 75%.

4 Conclusions

In this paper, we describes a Chinese word segmentation system based on word boundary token model and triple context matching model (Tsai, 2005) for extracting unknown words; and word support model (Tsai, 2006a and 2006b) for resolving segmentation ambiguity. To develop the word boundary model, we define WBT and classify WBT into three types of left, right and bi-direction. As per three types of WBT, we define WBT probability and WBT frequency.

In the SIGHAN Bakeoff 2007, we take part in the CKIP and the CityU closed word segmentation tasks. The scored results show that our CWS can increase the Bakeoff baseline system with 4.86% and 5.04% F-measures for the CKIP and the CityU word segmentation tasks, respectively. On the other hand, we show that the major room for improving our CWS system is the 2-char unknown word extraction of the word boundary model and triple context matching model. The performance of word support model is great and supports our previous work (Tsai, 2006a and 2006b).

We believe one major advantage of the WBT model is to use it with web as live corpus to minimum the corpus sparseness effect. Therefore, in the future, we shall investigate the WBT model with the web corpus, such as the searching results of GOOGLE and Yahoo!, etc.

References

CKIP (Chinese Knowledge Information Processing Group). 1995. *Technical Report no. 95-02, the content and illustration of Sinica corpus of Academia Sinica*. Institute of Information Science, Academia Sinica.

CKIP (Chinese Knowledge Information Processing Group). 1996. *A study of Chinese Word Boundaries and Segmentation Standard for Information processing* (in Chinese). Technical Report, Taiwan, Taipei, Academia Sinica.

Levow, Gina-Anne. 2006. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition, In *Proceedings of SIGHAN5 the 3rd International Chinese Language Processing Bakeoff at Coling/ACL 2006*, July, Sydney, Australia, 108-117.

Thomas, Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff, In *Proceedings of The 2nd International Chinese Word Segmentation Bakeoff at SIGHAN-4*, October, Jeju Island, Korea, 123-133.

Tsai, Jia-Lin. 2005. Report to BMM-based Chinese Word Segmentor with Context-based Unknown Word Identifier for the Second International Chinese Word Segmentation Bakeoff, In *Proceedings of The 2nd International Chinese Word Segmentation Bakeoff at SIGHAN-4*, October, Jeju Island, Korea, 142-145.

Tsai, Jia-Lin. 2006. Using Word Support Model to Improve Chinese Input System, In *Proceedings of Coling/ACL 2006*, July, Sydney, Australia, 842-849.

Tsai, Jia-Lin. 2006. BMM-based Chinese Word Segmentor with Word Support Model for the SIGHAN Bakeoff 2006, In *Proceedings of SIGHAN5 the 3rd International Chinese Language Processing Bakeoff at Coling/ACL 2006*, July, Sydney, Australia, 130-133.