Report to BMM-based Chinese Word Segmentor with Context-based Unknown Word Identifier for the Second International Chinese Word Segmentation Bakeoff

Jia-Lin Tsai
Tung Nan Institute of Technology, Department of Information Management
Taipei 222, Taiwan, R.O.C.
tsaijl@mail.tnit.edu.tw

Abstract

This paper describes a Chinese word segmentor (CWS) based on backward maximum matching (BMM) technique for the 2nd Chinese Word Segmentation Bakeoff in the Microsoft Research (MSR) closed testing track. Our CWS comprises of a context-based Chinese unknown word identifier (UWI). All the context-based knowledge for the UWI is fully automatically generated by the MSR training corpus. According to the scored results of the MSR closed testing track and our analysis, it shows that our BMM-based CWS with the context-based UWI is a simple and effective system to achieve high Chinese word segmentation performance of more than 95.5% F-measure.

1 Introduction

In the research fields of Chinese natural language processing (NLP), a high-performance Chinese word segmentor (CWS) is a useful preprocessing stage to produce an intermediate result for later processes, such as search engines, text mining and speech recognition, etc. The bottleneck of developing a high-performance CWS is to comprise of a high-performance Chinese UWI (Lin et al. 1993; Tsai et al. 2003). It is because Chinese is written without any separation between words and meanwhile more than 50% words of the Chinese texts in web corpus are out-of-vocabulary (Tsai et al. 2003).

Conventionally, there are four approaches to develop a CWS: (1) **Dictionary-based** approach (Cheng et al. 1999), especial forward and backward maximum matching (Wong and Chan, 1996); (2) **Linguistic** approach based on syntax-semantic knowledge (Chen et al. 2002); (3) **Statistical** approach based on statistical language model (SLM) (Sproat and Shih, 1990; Teahan et al. 2000; Gao et al. 2003); and (4) **Hybrid** approach trying to combine the benefits of dictionary-based, linguistic and statistical approaches (Tsai et al. 2003; Ma and Chen, 2003). In practice, statistical approaches are most widely used because their effective and reasonable performance. For a CWS, there are two types of word segmentation ambiguities while there are no unknown words in them: (1) **Overlap ambiguity** (OA), take a character string ABC as an example. If its segmentation can be either AB/C or A/BC depending on different context, the ABC is called an overlap ambiguity string (OAS), such as "將軍(a general)/用(use)" and "將(to get)/軍用(for military use)" (the symbol "/" indicates a word boundary); (2) **Combination ambiguity** (CA), take a character string AB as an example. If its segmentation can be either A/B or AB depending on different context, the AB is called a combination ambiguity string (CAS), such as "才(just)/能(can)" and "才能(ability)." Meantime, there are two types of error segmentation caused by unknown word problem: (1) **Lack of unknown word** (LUW), it means the error segmentation occurred by lack of an unknown word in the system dictionary, such as "瓦/西/里斯"; (2) **Error identified word** (EIW), it means the error segmentation occurred by an error identified unknown words, such as "切合

点.” To sum up, for a CWS in most case the UWI is a pre-processing stage to detect unknown words for the optimization of LUW-EIW tradeoff, and then to disambiguate those auto-detected OAS and CAS problems from the segmentation results.

The goal of this paper is to illustrate and report the effectiveness and the scored results of our BMM-based CWS for the second International Chinese Word Segmentation Bakeoff in the MSR closed (MSR_C) track. For this Bakeoff, our CWS is mainly addressed on optimizing the LUW-EIW tradeoff.

The remainder of this paper is arranged as follows. In Section 2, we present the details of our BMM-based CWS comprised of a context-based UWI. In Section 3, we present the scored results of the CWS in the MSR_C track and give our analysis. Finally, in Section 4, we give our conclusions and suggest some future research directions.

2 Development of BMM-based CWS

As per (Tsai et al. 2004), the Chinese word segmentation performance of BMM technique is about 1% greater than that of FMM technique. Thus, we adopt BMM technique as base to develop our CWS. The descriptions of symbols used in our CWS are given as below:

<BOS>: begin of sentence;
<EOS>: end of sentence;
<BOW>: begin of word;
<EOW>: end of word;
/: word boundary;
+: inner word boundaries of the segmentation of a system word segmented by BMM technique with the system dictionary exclusive of this system word;

SWS (stop word string): for a system word (such as “的(of)”), if the ratio (non-SWS probability) of total frequency of the other system words including it (such as “美的(beautiful)”) and its character string frequency is less than or equal to 1%, it is a SWS;

SWBS (stop word bigram string): for a word bigram (such as “才(just)/能(can)”), if the ratio (non-SWBS probability) of its character string (such as “才能(ability)” frequency and its character string frequency is less than or equal to 1%, it is a SWBS;

BMM-ASM (BMM ambiguity string mapping table: the BMM-ASM table lists all the pairs of correct SS (given in training corpus) and the error BMM SS (generated by BMM with the training system dictionary). Take the Chinese sentence “效果真好” as an example. As per its MSR-standard segmentation “效果(effect)/真(really)/好(good)” and its BMM segmentation “效(follow)/果真(indeed)/好(good),” the pair “效果/真”-“效/果真” is a BMM-ASM;

TCT (triple context template): a TCT comprised of three items from left to right are: the left word, the segmented system word and the right word, where the system word is not a mono-syllabic Chinese word. Take the Chinese sentence “效果/真/好” as an example. The two generated TCT are:
“<BOS>/效+1-char-word/真”
“<BOS>/1-char-word+果/真”; and

WCT (word context template): a WCT comprised of three items from left to right are: “<BOW>”, the segmented system word and “<EOW>”, where the system word is not a mono-syllabic word. Take the system word “喇嘛寺(lamasery)” as an example. Its two WCT are:
“<BOW>/喇嘛+1-char-word/<EOW>”
“<BOW>/2-char-word+寺/<EOW>.”

The algorithm of our BMM-based CWS comprised of a context-based UWI is as below:

Step 1. Generate BMM segmentation for the input Chinese sentence with system dictionary, firstly. The system dictionary comprised of all word types found in the training corpus. Then, use BMM-ASM table to revise the matched BMM ambiguity string.

Step 2. Use UWI to identify unknown words from the segmentation of Step 1 by the TCT knowledge, firstly. For the matched TCT, the characters between the left word and the right word will be combined as an UWI-identified word. If the UWI-identified word includes a SWS or a SWBS, it will be not an UWI-identified word. Then, use the system dictionary of Step 1 inclusive of the UWI-identified words of this step to repeat Step 1 process.

Step 3. Add tags “<BOW>” and “<EOW>” at

the left-side and right-side of the continue 1-char character segmentations of Step 2, firstly. Then, use UWI to identify unknown words by the WCT knowledge. If the number of characters between "<BOW>" and "EOW>" is same with that of the matched WCT, these 1-char characters will be combined as an UWI-identified word. If the UWI-identified word includes a SWS or a SWBS, it will be not an UWI-identified word. Finally, use the system dictionary of Step 2 inclusive of those UWI-identified words of this step to repeat Step 1 process.

Step 4. Use UWI to combine a word bigram into a word by the following two conditions: (1) if the non-SWS probability of the right first character of the left-side word is greater or equal to 99% and (2) if the non-SWS probability of the left first character of the right-side word is greater or equal to 99%. Take the word bigram "２００/２" as an example. Since the non-SWS probability of the right first character "０" of the left-side word "２００" is 99.95%, "２００２" is identified as an UWI-identified word. If the UWI-identified word includes a SWS or a SWBS, it will be not an UWI-identified word. Finally, use the system dictionary of Step 3 inclusive of those UWI-identified words of this step to repeat the Step 1 process.

Step 5. Repeat the Step 2 process.

Step 6. Repeat the Step 3 process.

Step 7. Repeat the Step 4 process.

Step 8. Stop.

In the above algorithm, Steps 2, 3 and 4 repeated at Steps 5, 6 and 7, respectively, are designed to show the recursive effect of our CWS.

3 The Scored Results and Analysis

In the 2nd Chinese Word Segmentation Bakeoff, there are four training corpus: AS (Academia Sinica) and CU (City University of Hong Kong) are traditional Chinese corpus, PU (Peking University) and Microsoft Research (MSR) are simplified Chinese corpus. Meanwhile, there are two testing tracks of this bakeoff: closed and open. We attend MSR_C track. The non-SWS and the non-SWBS probabilities of our CWS for this bakeoff are all set to 1%. And, the segmentation results of each step of our CWS are collected and scored, respectively.

3.1 The Scored Results

Table 1 shows the details of MSR training and testing corpus. Note that, in Table 1, the details of MSR testing corpus were computed by us according to the MSR gold testing corpus. From Table 1, it indicates that the MSR testing track seems to be a 25-folds experiment design.

Table 1. The details of MSR_C corpus

	Training	Testing
Sentences	86,924	3,985
Word types	88,119	12,924
Words	2,368,391	109,002
Character types	5,167	2,839
Characters	4,050,469	184,356

Table 2 shows the scored results of our CWS in MSR_C track. The performance of "Step 1(P)" in Table 2 was computed by us and the others were from the scored results. It shows a very high performance of 99.1% F-measure can be achieved while the BMM-based CWS by using a system dictionary comprised of word types found in the MSR training and testing corpus at Step 1 ("P" means "Perfect").

Table 2. The performance of each step of our CWS in the MSR-C track (OOV is 0.026)

Step	R	P	F	R_{OOV}	R_{IV}
1(P)	0.993	0.989	0.991	-	-
1	0.963	0.924	0.943	0.025	0.989
2	0.964	0.924	0.944	0.025	0.989
3	0.968	0.938	0.953	0.205	0.989
4	0.958	0.949	0.954	0.465	0.972
5	0.958	0.951	0.954	0.493	0.971
6	0.958	0.952	0.955	0.503	0.970
7	0.958	0.952	0.955	0.504	0.970

3.2 The Analysis

Table 3 (see next page) shows the differences of F-measure and R_{OOV} between each near-by step of our CWS. From Table 3, it indicates that the most contribution for increasing the overall performance (F-measure) of our CWS is at Step 3, which uses WCT knowledge.

Table 4 (see next page) shows the distributions of four segmentation error types (OAS, CAS, LUW and EIW) for each step of our CWS. From Table 4, it shows that our context-based UWI with the knowledge of TCT and WCT can

effectively to optimize the LUW-EIW tradeoff. Moreover, from Table 4, it also shows that the knowledge of SWS, SWBS and BMM-ASM can effectively to resolve the CAS errors.

Table 3. The differences of F-measure and R_{OOV} between near-by steps of our CWS

Step	F	F(d)	R_{OOV}	R_{OOV}(d)
1	0.943	-	0.025	-
2	0.944	0.001	0.025	0
3	0.953	0.011	0.205	0.18
4	0.954	0.001	0.465	0.26
5	0.954	0	0.493	0.028
6	0.955	0.001	0.503	0.01
7	0.955	0	0.504	0.001

Table 4. The number of OAS (types), CAS (types), LUW (types) and EIW (types) for each step of our CWS

	OAS	CAS	LUW	EIW
1	210(194)	233(80)	2702(1930)	157(96)
2	184(173)	233(80)	2698(1927)	157(96)
3	185(174)	232(80)	2169(1473)	187(126)
4	250(226)	226(77)	1373(1090)	946(609)
5	250(226)	226(77)	1283(1018)	991(658)
6	251(227)	224(77)	1255(1001)	1005(669)
7	262(216)	224(76)	1260(1005)	1007(668)

4 Conclusions and Future Directions

In this paper, we have applied a BMM-based CWS comprised of a context-based UWI to the Chinese word segmentation and obtained a high performance of 95.5% F-measure in the MSR closed track. To sum up the results of this study, we have following conclusions and future directions:

(1) Since the F-measure of Step 1 of our CWS is 94.3%, it indicates that the BMM with BMM-ASM knowledge is a simple but probably effective technique as a good base in developing a high performance CWS;

(2) Since 82% of segmentation errors of our CWS caused by LUW problem, this result supports that a high performance CWS is relied on a high performance Chinese UWI.

(3) For a CWS, there are two critical and probably independent tasks: the optimization of LUW-EIW tradeoff and the detection and disambiguation of OAS and CAS error segmentation. We believe the former task is more critical than the later one.

(4) We will continue to expand our CWS with other linguistic knowledge (such as part-of-speech information and morphology) and BTM model (Tsai 2005) to improve our BMM-based CWS for attending the third International Chinese Word Segmentation Bakeoff in both closed and open testing tracks.

References

Chen, Keh-Jiann and Wei-Yun, Ma. 2002. Unknown Word Extraction for Chinese Documents, *Proceedings of 19th COLING 2002*, Taipei, 169-175.

Cheng, Kowk-Shing, Gilbert H. Yong and Kam-Fai Wong.. 1999. A study on word-based and integral-bit Chinese text compression algorithms. *JASIS*, 50(3): 218-228.

Gao, Jianfeng, Mu Li and Chang-Ning uang. 2003. Improved Source-Channel Models for Chinese Word Segmentation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 272-279.

Lin, Ming-Yu, Tung-Hui Chiang and Keh-Yi Su. 1993. A preliminary study on unknown word problem in Chinese word segmentation. ROCLING 6, 119-141.

Ma, Wei-Yun and Keh-Jiann Chen, 2003, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, pp168-171.

Sproat, R. and C., Shih. 1990. A Statistical Method for Finding Word Boundaries in Chinese Text. Computer proceeding of Chinese and Oriental Language, 4(4):336 349.

Teahan, W. J., Yingying Wen, Rodger McNad and Ian Witten. 2000. A compression-based algorithm for Chinese word segmentation. *Computational Linguistics*, 26(3): 375-393.

Tsai, Jia-Lin, C.L., Sung and W.L., Hsu. 2003. Chinese Word Auto-Confirmation Agent, *Proceedings of ROCLING XV*, Taiwan, 175-192.

Tsai, Jia-Lin, G., Hsieh and W.L., Hsu. 2004. Auto-Generation of NVEF knowledge in Chinese, *Computational Linguistics and Chinese Language Processing*, 9(1):41-64.

Tsai, Jia-Lin. 2005. A Study of Applying BTM Model on the Chinese Chunk Bracketing. *Proceedings of IJCNLP, 6th International Workshop on Linguistically Interpreted Corpora,* Jeju Island.

Wong, Pak-Kwong and Chorkin ChanWong. 1996. Chinese Word Segmentation. based on Maximum Matching and Word Binding Force. *Proceedings of the 16th International conference on Computational linguistic,* 1:200-203.