BMM-based Chinese Word Segmentor with Word Support Model for the SIGHAN Bakeoff 2006

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) for the third International Chinese Language Processing Bakeoff (SIGHAN Bakeoff 2006). We participate in the word segmentation task at the Microsoft Research (MSR) closed testing track. Our CWS is based on backward maximum matching with word support model (WSM) and contextual-based Chinese unknown word identification. From the scored results and our experimental results, it shows WSM can improve our previous CWS, which was reported at the SIGHAN Bakeoff 2005, about 1% of F-measure.

1 Introduction

A high-performance Chinese word segmentor (CWS) is a critical processing stage to produce an intermediate result for later processes, such as search engines, text mining, word spell checking, text-to-speech and speech recognition, etc. As per (Lin et al. 1993; Tsai et al. 2003; Tsai, 2005), the bottleneck for developing a high-performance CWS is to comprise of high performance Chinese unknown word identification (UWI). It is because Chinese is written without any separation between words and more than 50% words of the Chinese texts in web corpus are out-of-vocabulary (Tsai et al. 2003). In our report for the SIGHAN Bakeoff 2005 (Tsai, 2005), we have shown that a highly performance of 99.1% F-measure can be achieved while a BMM-based CWS using a perfect system dictionary (Tsai, 2005). A perfect system dictionary means all word types of the dictionary are extracted from training and testing gold standard corpus.

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.

To develop UWI, there are three approaches: (1) **Statistical approach**, researchers use common statistical features, such as maximum entropy (Chieu *et al*. 2002), association strength, mutual information, ambiguous matching, and multi-statistical features for unknown word detection and extraction; (2) **Linguistic approach**, three major types of linguistic rules (knowledge): morphology, syntax, and semantics, are used to identify unknown words; and (3) **Hybrid approach**, recently, one important trend of UWI follows a hybrid approach so as to take advantage of both merits of statistical and linguistic approaches. Statistical approaches are simple and efficient whereas linguistic approaches are effective in identifying low frequency unknown words (Chen *et al*. 2002).

To develop WSD, there are two major types of word segmentation ambiguities while there are no unknown word problems with them: (1) **Overlap Ambiguity** (OA). Take string C1C2C3

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 130–133,
Sydney, July 2006.

comprised of three Chinese characters C1, C2 and C3 as an example. If its segmentation can be either C1C2/C3 or C1/C2C3 depending on context meaning, the C1C2C3 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 string C1C2 comprised of two Chinese characters C1 and C2 as an example. If its segmentation can be either C1/C2 or C1C2 depending on context meaning, the C1C2 is called a combination ambiguity string (CAS), such as "才(just)/能(can)" and "才能(ability)." Besides the OA and CA problems, the other two types of word segmentation errors are caused by unknown word problems. They are: (1) **Lack of unknown word** (LUW), it means segmentation error occurred by lack of an unknown word in the system dictionary, and (2) **Error identified word** (EIW), it means segmentation error occurred by an error identified unknown words.

The goal of this paper is to report the approach and experiment results of our backward maximum matching-based (BMM-based) CWS with word support model (WSM) for the SIGHAN Bakeoff 2006. In (Tsai, 2006), WSM has been shown effectively to improve Chinese input system. In the third Bakeoff, our CWS is mainly addressed on improving its performance of OA/CA disambiguation by WSM. We show that WSM is able to improve our BMM-based CWS, which reported at the SIGHAN Bakeoff 2005, about 1% of F-measure.

The remainder of this paper is arranged as follows. In Section 2, we present the details of our BMM-based CWS comprised of WSM. In Section 3, we present the scored results of the CWS at the Microsoft Research closed track and give our experiment results and analysis. Finally, in Section 4, we give our conclusions and future research directions.

2 BMM-based CWS with WSM

From our work (Tsai et al. 2004), the Chinese word segmentation performance of BMM technique is about 1% greater than that of forward maximum matching (FMM) technique. Thus, we adopt BMM technique as base to develop our CWS. In this Bakeoff, we use context-based Chinese unknown word identification (CCUWI) (Tsai, 2005) to resolve unknown word problem. The CCUWI uses template matching technique to extract unknown words from sentences. The context template includes triple context template (TCT) and word context template (WCT). The details of the CCUWI can be found in (Tsai, 2005). In (Tsai, 2006), we propose a new language model named word support model (WSM) and shown it can effectively perform homophone selection and word-syllable segmentation to improve Chinese input system. For this Bakeoff, we use WSM to resolve OA/CA problems.

The two steps of our BMM-based CWS with WSM are as below:

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

Step 2. Use WSM to resolve OA/CA problems for the BMM segmentation of Step 1. Now, we give a brief description of how we use WSM to resolve OA/CA problem. Firstly, we pre-collect OA/CA pattern-pairs (such as "就/是"-"就是") by compare each training gold segmentation and its corresponding BMM segmentation. The pattern of OA/CA pattern-pairs can be a segmentation pattern, such as "就/是," or just a word, such as "就是." Secondly, for a BMM segmentation of Step 1, if one pattern matching (matching pattern) with at least one pattern of those pre-collected OA/CA pattern-pairs (matching OA/CA pattern-pairs), CWS will compute the word support degree for each pattern of the matching OA/CA pattern-pair. Finally, select out the pattern with maximum word support degree as its segmentation for the matching pattern. If the patterns of the matching OA/CA pattern-pair having the same word support degree, randomly select one to be its segmentation. The details of WSM can be found in (Tsai, 2006).

3 Scored Results and Our Experiments

In the SIGHAN Bakeoff 2006, there are four training corpus for word segmentation (WS) task: 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. And, for each corpus, there are closed and open

track. In the Bakeoff 2006, we attend the Microsoft Research closed (MSR_C) track.

3.1 Scored Results and our Experiments

Tables 1a and 1b show the details of MSR training and testing corpus for 2nd (2005) and 3rd (2006) bakeoff. From Table 1a and 1b, it indicates that MSR track of 3rd bakeoff seems to be a more difficult WS task than that of 2nd bakeoff, since (1) the training size of 2nd bakeoff is two times as great as that of 3rd bakeoff; (2) in training data, the word type number of 3rd bakeoff is less than that of 2nd bakeoff, and (3) in testing data, the word type number of 3rd bakeoff is greater than that of 2nd bakeoff.

	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 1a. Details of MSR_C corpus of 2nd bakeoff.

	Training	Testing
Sentences	46,364	4356
Word types	63,494	13,461
Words	1,266,169	100,361
Character types	4,767	3,103
Characters	2,169879	172,601

Table 1b. Details of MSR_C corpus of 3rd bakeoff.

Table 2 shows the scored results of our CWS at the MSR_C track of this bakeoff. In Table 2, the symbols a, b and c stand for the CWS with a, b and c system dictionary. The system dictionary "a" is the dictionary comprised of all word types found in the MSR training corpus. The system dictionary "b" is the dictionary comprised of "a" system dictionary and the word types found in the testing corpus by CCUWI with TCT knowledge. The system dictionary "c" is the dictionary comprised of "a" system dictionary and the word types found in the testing corpus by CCUWI with TCT and WCT knowledge. Table 3 is F-measure differences between the BMM-based CWS system and it with WSM and CCUWI using "a", "b" and "c" system dictionary in the MSR_C track.

From Tables 2 and 3, we conclude that our CWS of 3rd bakeoff improve the CWS of 2nd bakeoff about 1.8% of F-measure. Among the 1.8% F-measure improvement, 1% is contributed by WSM for resolving OA/CA problems and the other 0.8% is contributed by CCUWI for resolving UWI problem.

System	R	P	F	R_{OOV}	R_{IV}
a	0.949	0.897	0.922	0.022	0.982
b	0.954	0.921	0.937	0.163	0.981
c	0.950	0.930	0.940	0.272	0.974

Table 2. The scored results of our CWS in the MSR_C track (OOV is 0.034) for 3rd bakeoff.

System	R	P	F	Improve
a1.BMM	0.949	0.897	0.922	
a2.BMM+WSM	0.958	0.907	0.932	0.010
b1.BMM	0.946	0.911	0.928	
b2.BMM+WSM	0.954	0.921	0.937	0.009
c1.BMM	0.938	0.920	0.929	
c2.BMM+WSM	0.950	0.930	0.940	0.011

Table 3. The F-measure improvement between the BMM-based CWS and it with WSM in the MSR_C track (OOV is 0.034) using a, b, and c system dictionary.

3.2 Error Analysis

Table 4 shows the F-measure and R_{OOV} differences between each result of our CWS with a, b and c system dictionaries. From Table 4, it indicates that the most contribution for increasing the overall performance (F-measure) of our CWS is occurred while our CWS comprised of WSM and CCUWI with TCT knowledge.

System	F	F(d)	R_{OOV}	R_{OOV}(d)
a	0.922	-	0.022	-
b	0.937	0.015	0.163	0.141
c	0.940	0.003	0.272	0.109

Table 4. The differences of F-measure and ROOV between near-by steps of our CWS.

	OA	CA	LUW	EIW
a	667(389)	403(194)	3268(2545)	0(0)
c	160(147)	231(150)	2310(1887)	805(605)

Table 5. The number of OAS (types), CAS (types), LUW (types) and EIW (types) for our CWS.

Table 5 shows the distributions of four segmentation error types (OA, CA, LUW and EIW) for each result of our CWS with a and c system dictionaries. From Table 5, it shows CCUWI with the knowledge of TCT and WCT can be used to optimize the LUW-EIW tradeoff. Moreover, it shows that WSM can effectively to reduce the number of OA/CA segmentation errors from 1,070 to 391.

4 Conclusions and Future Directions

In this paper, we have applied a BMM-based CWS comprised of a context-based UWI and word support model to the Chinese word segmentation. While we repeat the CWS with the MSR_C track data of 2nd bakeoff, we obtained 96.3% F-measure, which is 0.8% greater than that (95.5%) of our CWS at 2nd bakeoff. To sum up the results of this study, we have following conclusions and future directions:

(1) **UWI and OA/CA problems could be independent tasks for developing a CWS**. The experiment results of this study support this observation. It is because we found 1% improvement is stable contributed by WSM and the other 0.8% improvement is stable contributed by the CCUWI while the BMM-based CWS with difference a, b and c system dictionaries and different MSR_C training and testing data of 2nd and 3rd bakeoff.

(2) About 89% of segmentation errors of our CWS caused by unknown word problem. In the 89%, we found 66% is LUW problem and 23% is EIW problem. This result indicates that the major target to improve our CWS is CCUWI. The result also supports that a high performance CWS is relied on a high performance Chinese UWI (Tsai, 2005).

(3) We will continue to expand our CWS with other unknown word identification techniques, especially applying n-gram extractor with the TCT and WCT template matching technique to improve our CCUWI for attending the fourth SIGHAN Bakeoff.

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.

Chieu, H.L. and H.T. Ng. 2002. Named Entity Recognition: A Maximum Entropy Approach Using Global Information. *Proceedings of 19th COLING 2002*, Taipei, 190-196.

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.

Tsai, Jia-Lin. 2006. Using Word Support Model to Improve Chinese Input System. *Proceedings of ACL/COLING 2006*, Sydney.

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.