Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff

Wei-Yun Ma
Institute of Information science,
Academia Sinica
ma@iis.sinica.edu.tw

Keh-Jiann Chen
Institute of Information science,
Academia Sinica
kchen@iis.sinica.edu.tw

Abstract

In this paper, we roughly described the procedures of our segmentation system, including the methods for resolving segmentation ambiguities and identifying unknown words. The CKIP group of Academia Sinica participated in testing on open and closed tracks of Beijing University (PK) and Hong Kong Cityu (HK). The evaluation results show our system performs very well in either HK open track or HK closed track and just acceptable in PK tracks. Some explanations and analysis are presented in this paper.

1 Introduction

At the first international Chinese Word Segmentation Bakeoff, Academia Sinica participated in testing on open and closed tracks of Beijing University (PK) and Hong Kong Cityu (HK). The same segmentation algorithm was applied to process these two corpora, except that character code conversion from GB to BIG5 for PK corpus and few modifications due to different segmentation standards had been made. The difference between open and closed tracks is that while processing the open track, besides of the lexicon trained from the specific corpus, we also consulted the Academia Sinica lexicon to enhance the word collection.

It is well known that there are two major difficulties in Chinese word segmentation. One is resolving the ambiguous segmentation, and the other is identifying unknown words.

Our earlier work mainly focused on the resolving of segmentation ambiguities and using regular expressions to handle the determinant-measure and reduplication compounds (Chen & Liu 1992, Chen 1999). We adopt a variation of the longest matching algorithm with several heuristic rules to resolve the ambiguities and achieve 99.77% of the success rate without counting the mistakes occurred due to the existence of unknown words. After that, we were paying more attention on the problems of extracting and identifying unknown words (Chen et.al 1997, Chen & Bai 1998, Chen & Ma 2002, Tseng & Chen 2002, Ma & Chen 2003). The process of unknown word extraction could be roughly divided into two steps, i.e. detection process and extraction process. The detection process detects possible occurrences of unknown words (Chen & Bai 1998), so that deeper morphological analysis is carried out only at the places where unknown word morphemes were detected (Chen & Ma 2002). A bottom-up merging algorithm was proposed in (Ma & Chen 2003), which utilizes hybrid statistical and linguistic information to extract unknown words effectively.

In addition to the bakeoff results evaluated by SIGHAN, we also present some other relevant experiment results and provide analysis on the system performance in the following sections.

2 System Overview

Figure 1 illustrates the block diagram of our segmentation system used in this contest. The first two steps of word segmentation algorithm are word matching and resolution for ambiguous matches. These two processes were performed in

parallel. The algorithm reads the input sentences from left to right and matches the input character string with lexemes. In (Chen & Liu 1992), if an ambiguous segmentation does occur, the matching algorithm looks ahead two more words, and the disambiguation rules for those three word chunks is applied afterward. For instance, in (1), the first matched word could be '完' or '完成'. Then the algorithm will look ahead to take all of the possible combinations of three word chunks, as shown in (2), into consideration.

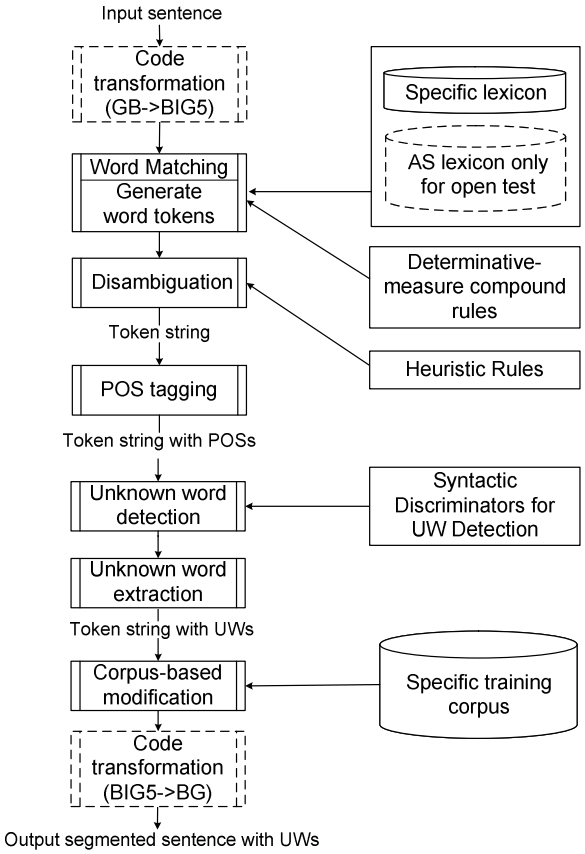

Figure 1. Flowchart of the system

(1) 完成 鑑定 報告
complete authenticate report
"complete the report about authenticating"

(2) 完 成 鑑定
完成 鑑定 報
完成 鑑定 報告

The disambiguation algorithm will select the first word of the most plausible chunks as the solution according to heuristic rules. The first heuristic rule is:

Longest Matching Rule: The most plausible segmentation is the three word sequence with the longest length.

In the above example, the longest matched three-word chunk is (1). Therefore the first segmented word is '完 成'. This heuristic rules achieves as high as 99.69% accuracy and a high applicability of 93.21%, i.e. the 93.21% of the ambiguities were resolved by this rule. However there are still about 6.79% of ambiguities, i.e. the three word chunks with the same length but with different segmentations, which cannot be resolved by the maximal matching rule. The following heuristic rules were used for further resolution.

Word Length Rule: Pick the three-word chunk that has the smallest standard deviation in length of the three words.

Morphemic Rules:

(a). Pick the chunk with fewer bound morphemes.
(b). Pick the chunk with fewer characters in compound words.

Probability Rule:

(a). Pick the chunk with the high frequency monosyllabic words.
(b). Pick the chunk with the highest probability value.

After disambiguation process, an input sentence is segmented into a word sequence. Then for the needs of the following unknown word extraction, a Pos bi-gram tagging model is applied to tag Pos of words.

It is clear that unknown words in the input text will be segmented into several adjacent tokens (known words or monosyllabic morphemes). Then at unknown word detection stage, every monosyllable is decided whether it is a word or an unknown word morpheme by a set of syntactic discriminators, which are trained from a word segmented corpus.

(3) 若 能 提升 毛利率
if can increase gross profit rate
"if gross profit rate can be increased…"

(4) after first step word segmentation:
若 能 提升 毛 利 率
after unknown word detection:
若 能 提升 毛(?) 利(?) 率(?)
after unknown word extraction:
若 能 提升 毛利率

For example, the correct segmentation of (3) is shown, but the unknown word "毛利率" is segmented into three monosyllabic words after the first step of word segmentation process. In (4), The unknown word detection process will mark the sentence as "若() 能() 提升() 毛(?) 利(?) 率(?)", where (?) denotes the detected monosyllabic unknown word morpheme and () denotes common words. During extracting process, the rule matching process focuses on the morphemes marked with (?) only and tries to combine them with left/right neighbors according to the rules for unknown words. After that the unknown word "毛利率" is extracted.

We adopt a bottom-up merging algorithm (Ma & Chen 2003), which utilizes hybrid statistical and linguistic information, to extract unknown words.

3 Adaptation for Different Tracks

It is known that different segmentation standards could affect the performance of segmentation significantly. In this contest, due to limited preparing time, we mainly focused on adjusting the regular expressions for determinant-measure compounds according to the HK and PK segmentation standards.

While processing the PK track, a shortcut method of converting GB codes to BIG5 codes was adopted to cope with the problem of character coding difference. Instead of re-design or re-implement the GB segmentation system, we convert the codes of training and testing PK corpora into BIG5 versions and perform the segmentation under the BIG5 environment. The segmented results are then translated back to GB code as the final outputs. In contrast, processing of HK corpus is easier for us, because our system was designed for the BIG5 environment.

As for the lexicons, for closed test, both PK and HK lexicons are derived from the word sets of each respective training corpus. For the open test, each lexicon was enhanced by adding the lexical entries in the CKIP lexicon. The sizes of lexicons are shown in table1.

	HK	PK
# of lexical entries (HK/PK)for closed test	22K	50K
# of lexical entries (HK/PK join CKIP) for open test	140K	156 K

Note: # lexicon of (CKIP) is 133K

Table 1. The sizes of lexicons

Syntactic categories of a words were utilized in unknown word detection and extraction processes. We don't have syntactic categories for words which are not in the CKIP lexicon. Therefore, we (Chen et.al 1997, Tseng & Chen 2002) use association strength between morphemes and syntactic categories to predict the category of a new word. The accuracy rate is about 80%.

4 Evaluation Results

There are several evaluation indexes provided by SIGHAN, i.e. test recall (R), test precision (P), F score[2], the out-of-vocabulary (OOV) rate for the test corpus, the recall on OOV words (R_{oov}), and the recall on in-vocabulary (R_{iv}) words.

Tables 2 shows the evaluation results of our system in HK closed and open tracks. For both tracks, our system achieved the top ranks on F scores.

	R	P	F	OOV	R_{oov}	R_{iv}
Closed	0.947	0.934	0.940	0.071	0.625	0.972
Open	0.958	0.954	0.956	0.071	0.788	0.971

Note: The word count of testing corpus is 34955

Table 2. Scores for HK

The evaluations of our system in PK closed and open tracks are shown in table 3. For PK closed track, our system ranks 6th among 10 systems. And for PK open track, our system ranks 3rd among 8 systems.

	R	P	F	OOV	R_{oov}	R_{iv}
Closed	0.939	0.934	0.936	0.069	0.642	0.961
Open	0.939	0.938	0.938	0.069	0.675	0.959

Note: The word count of testing corpus is 34955

Table 3. Scores for PK

Because Academia Sinica corpora (AS) are provided by us, we are not allowed to participate any AS track at this contest. Therefore, in this report, we still show the performance of our system evaluating AS closed track in table 4. Our system would have the top rank if the result was compared with the other 6 participants of AS closed track.

R	P	F	OOV	R_{oov}	R_{iv}
0.968	0.966	0.967	0.022	0.657	0.975

Note: The word count of testing corpus is 11985

Table 4. Scores for AS closed

5 Discussions and Conclusions

The evaluation results show that our system performs very well in either HK closed track or HK open track. We think the key to the success is our unknown word extraction performs better than other participants. This could be observed by the results of HK closed track, the 2th and 3th system, which have better performance in R_{iv} but worse R_{oov} than our system, performs worse than our system in f score. Furthermore to have better performance, high precision for unknown word extraction is necessary, since one identification error may cause at least two segmentation errors.

The performance in PK tracks are not as well as HK. An important reason is that coding conversion may cause errors. For instance, in the conversion of the GB code of "里約" (the capital of Brazil) to its BIG5 codes, Since GB code to BIG5 conversion is a one-to-many mapping, the above example is wrongly converted to "裡約". This kind of errors do affect accuracy of the segmentation significantly, especially for the unkown word processes. To solve this problem, we think the best and direct solution is to re-implement the GB segmentation version without any code conversion.

Variation on the word segmentation standards is another reason of causing segmentation errors. Some of the standards were even not available to the public. It is better to propose a uniform word segmentation standard in the future.

Regarding evaluation index, we suggest that an error type of crossing error should be take into consideration, since noncrossing errors are more or less related to segmentation standards and crossing errors are more severe.

6 References

[1] Chen, K.J. & S.H. Liu, 1992,"Word Identification for Mandarin Chinese Sentences," *Proceedings of 14th Coling*, pp. 101-107

[2] Chen, C. J., M.H. Bai, & K.J. Chen, 1997," Category Guessing for Chinese Unknown Words," Proceedings of the Natural Language Processing Pacific Rim Symposium, 35-40, Thailand.

[3] Chen, K.J. & Ming-Hong Bai, 1998, "Unknown Word Detection for Chinese by a Corpus-based Learning Method," international Journal of Computational linguistics and Chinese Language Processing, Vol.3, #1, pp.27-44

[4] Chen, Keh-jiann,1999,"Lexical Analysis for Chinese- Difficulties and Possible Solutions", Journal of Chinese Institute of Engineers, Vol. 22. #5, pp. 561-571.

[5] Chen, K.J. & Wei-Yun Ma, 2002. Unknown Word Extraction for Chinese Documents. In Proceedings of COLING 2002, pages 169-175

[6] Tseng, H.H. & K.J. Chen, 2002. Design of Chinese Morphological Analyzer. In Proceedings of SIGHAN, pages 49-55

[7] Ma Wei-Yun & K.J. Chen, 2003. A bottom-up Merging Algorithm for Chinese Unknown Word Extraction. In Proceedings of SIGHAN