Maximal Match Chinese Segmentation Augmented by Resources Generated from a Very Large Dictionary for Post-Processing

Ka-Po Chow **Andy C. Chin** **Wing Fu Tsoi**
Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
{kapo.chow,cochin,rlwftsoi}@cityu.edu.hk

Abstract

We used a production segmentation system, which draws heavily on a large dictionary derived from processing a large amount (over 150 million Chinese characters) of synchronous textual data gathered from various Chinese speech communities, including Beijing, Hong Kong, Taipei, and others. We run this system in two tracks in the Second International Chinese Word Segmentation Bakeoff, with Backward Maximal Matching (right-to-left) as the primary mechanism. We also explored the use of a number of supplementary features offered by the large dictionary in post-processing, in an attempt to resolve ambiguities and detect unknown words. While the results might not have reached their fullest potential, they nevertheless reinforced the importance and usefulness of a large dictionary as a basis for segmentation, and the implication of following a uniform standard on the segmentation performance on data from various sources.

1 Introduction

Our team has participated in two tracks of the ACL SIGHAN-sponsored Second International Chinese Word Segmentation Bakeoff, namely Academia Sinica open (ASo) and Peking University open (PKo). The production segmentation system we used draws heavily on a large dictionary derived from processing a very large amount of synchronous textual data. In Section 2, our segmentation flow for the current Bakeoff will be described, and in Section 3, the results will be evaluated and analysed. Errors will be analysed and implications discussed in Section 4, followed by a conclusion in Section 5.

2 Segmentation Framework

The major resource of our segmentation system is a large dictionary. In the following, we describe the main segmentation mechanism based on maximal matching, and other supplementary features for post-processing attempted in the current Bakeoff.

2.1 Dictionary-based Segmentation

The primary mechanism of segmentation makes use of a large dictionary derived from processing a large amount (over 150 million Chinese characters) of synchronous textual data, mostly printed news, gathered from various Chinese speech communities, including Beijing, Hong Kong, Taipei, and others, following a uniform segmentation standard. The dictionary has now grown to a size of over 800,000 word types, with frequencies of each entry being tracked closely. For this Bakeoff, additional items from the respective training data were also included in the existing dictionary for segmentation. Thus unsegmented texts will first go through a process of Backward Maximal Matching (BMM) segmentation equipped with the combined dictionary.

2.2 Supplementary Features

2.2.1 Rule Development

According to specific divergence of the segmentation standard of each test corpus from our production standard, a set of general adaptation rules were applied to transform the texts to achieve “standard complacency” as much as possible. The adaptation rules vary in nature, depending on how intense the segmentation standard differences are between each test corpus and our own. Hence some rules are based on linguistic structures while others are based on particular treatment of elements like numerals and units.

These adaptation rules are coupled with a set of heuristic segmentation disambiguation rules derived from our long-term and extensive processing of text data. Such rules are based on BMM, and amount to around 20,000 at the time of writing. Each rule has gone through careful consideration before putting to real production use, to ensure that they produce correct results in most cases without overgeneralisation.

2.2.2 Statistical BMM/FMM Comparison and Replacement

After texts were segmented by BMM, the forward counterpart (Forward Maximal Matching, FMM) was also done for comparison, as the discrepancies between the two segmented texts often indicate potential ambiguities. Statistical information such as the frequency distributions of the segmented units in question were obtained from our large dictionary. By comparing the independent joint likelihood of the two combinations, segmented units with exceptionally low frequency are likely to be disregarded, allowing us to choose the correct segmentation. For example, in the test data, the phrase 開設有 is segmented as 開/設有 by the backward approach, whereas 開設/有 will be obtained if segmented forwardly. The latter segmented alternative, 開設/有, is more likely to appear in the text.

2.2.3 Unknown Word Detection

One of the most challenging issues in Chinese word segmentation is the treatment of unknown words which can be further divided into two categories: new words (NWs) and named entities (NEs). In our treatment of unknown words, a slight distinction was made between Chinese NEs and other NWs including foreign names. The detection processes are similar but statistical data were gathered from different portions of our textual data. When a sequence of single characters is hit, windows of two and three characters (only nominal morphemes were considered) were extracted to form “potential NE/NW candidates”. The likelihood of these characters being monosyllabic words (i.e. out-word) and that of being part of multi-syllabic words (i.e. in-word) were compared to make the best guess whether they should be combined or segmented.

For NE detection, the in-word statistics was based on all the multi-syllabic named entities in the Taipei portion from our dictionary and the out-word statistics on the rest of it. The in-word frequency of a given character is thus the number of times the character appears within a multi-syllabic named entity. The in-word probability is the in-word frequency divided by the total number of times the character appears in all our textual data. The independent joint in-word and out-word probabilities were computed and compared for each candidate, which would be combined as a word if the in-word probability is greater than the out-word probability and the first character in the candidate is within a list of Chinese surnames, again collected from all textual data.

For NW detection, the in-word statistics was based on all the multi-syllabic words in our dictionary. For every newly combined word, neighbouring prefixes and suffixes (according to those provided in the segmentation standard) were also detected and combined, if any. A list of foreign names and all the characters appearing in them was also extracted from our dictionary. When a new word is detected, its neighbouring words would be scanned and would be combined if they are within this foreign name list, thus enabling the identification of names like 芬妮亞當.

3 Results and Analysis

The results of the different stages of segmentation are shown in Table 1 and Table 2.

In both test corpora, the primary dictionary-based segmentation alone has achieved a significant percentage (over 95% in recall and over 90% in precision). This exemplifies that the rich vocabulary we have offers a useful resource for language engineering, and provides a solid platform for further enhancement.

Post-processing with supplementary features from the dictionary shows consistent incremental improvement in segmentation. The scores (F-measure) due to FMM and BMM with heuristic rules demonstrate a relatively substantial gap at the very beginning, largely because of the heuristic rules developed and accumulated through the precise and systematic processing of our sizable textual data.

Operation	R	P	F	R_{OOV}	R_{IV}
FMM only	0.953	0.903	0.927	0.658	0.966
BMM, plus heuristic rules	0.960	0.915	0.937	0.661	0.974
Comparison and replacement	0.964	0.921	0.942	0.663	0.978
Unknown word detection	0.966	0.931	0.948	0.715	0.977
Official results*	0.943	0.931	0.937	0.531	0.962

Table 1: Results for ASo using combined dictionary

Operation	R	P	F	R_{OOV}	R_{IV}
FMM only	0.957	0.928	0.942	0.842	0.964
BMM, plus heuristic rules	0.967	0.947	0.957	0.849	0.974
Comparison and replacement	0.969	0.951	0.960	0.851	0.977
Official results*	0.952	0.951	0.951	0.784	0.962

Table 2: Results for PKo using combined dictionary

The performance of unknown word detection can be seen from the leap in R_{OOV} after the operation. It increases remarkably from 0.663 to 0.715, offsetting the fall in R_{IV}, which drops by 0.001 and this may be due to the concatenation of some monosyllabic morphemes which are supposed to be independent words.

The results of the comparison between FMM and BMM are summarized in Table 3 and Table 4. A noticeable drawback of such comparison is that some phrases will be mis-segmented in either direction. For example, the phrase 絲毫不覺得 will be segmented backwardly into 絲/毫不/覺得 but 絲毫/不覺/得 forwardly. The correct segmentation, 絲毫/不/覺得, cannot be attained in both cases. Hence as an experiment, for any combination of five characters which are segmented into 1/2/2 pattern by BMM and 2/2/1 pattern by FMM, the 2/1/2 pattern will be also tested against the overall probabilities. For the former example, 絲毫/不/覺得 will override the other two.

Table 3 shows that the number of correct replacements from FMM is 399 in the AS test corpus, combining the gain from the reshuffling of 5-character strings, the total is 408. Since the default choice is the BMM segmented texts, the sum 408 is the total gain from this BMM/FMM comparison, while 77 correct segmented texts have been mis-replaced, the gain/loss ratio is 5.30. This means that our system only loses 1 correct segmentation in exchange of gaining 5.3 correct ones.

Likewise in the case of the PK test corpus in Table 4, the gain/loss ratio is 4.67. The ratio is smaller than that for the AS test corpus. It is thus evident that the comparison and replacement by means of BMM and FMM offers a substantial achievement in the accuracy of the segmentation process.

	BMM	FMM	Re-shuffle	Total
Correct Replacement	1124	399	9	1532
Incorrect Replacement	281	267	0	548
Mis-replaced Correct Segmentation	77	78	/	155

Table 3: Analysis of BMM/FMM comparison for ASo

	BMM	FMM	Re-shuffle	Total
Correct Replacement	1097	254	3	1354
Incorrect Replacement	131	117	2	250
Mis-replaced Correct Segmentation	55	33	/	88

Table 4: Analysis of BMM/FMM comparison for PKo

We are aware that the performance of replacement may be improved by using probabilities of *n*-grams, conditional probabilities involving the boundary words, and perhaps by considering all possibilities of segmentations for the same string of texts, as in some other segmentation systems. On the semantic level, the overall message of a paragraph can be examined as well by gathering statistics of collocating words. The ordering of applying these algorithms, however, should be important, and how they interplay with one another will be an arena to explore.

Although we have not incorporated such enhancement measures into our system in this exercise, the dictionary can nevertheless support such extension with the necessary statistical data. All previous results are based on the first-stage of segmentation with a large dictionary. Since

[*] The reported figures differ from those computed on our platform, probably due to system differences. The official scorer program is publicly available and described in (Sproat and Emerson, 2003).

we had processed texts from different Chinese speech communities including Beijing, Hong Kong, Taipei, and others, the dictionary used for segmentation also consists of all words appearing in any of these communities. In order to investigate the effect of locality on the dictionary used in segmentation, two independent dictionaries have been generated from the Beijing portion and Taipei portion, and all the above stages were repeated for the two test corpora, with results shown below in Table 5 and Table 6.

Operation	R	P	F	R_{OOV}	R_{IV}
FMM only	0.942	0.891	0.916	0.602	0.957
BMM, plus heuristic rules	0.948	0.901	0.924	0.603	0.964
Comparison and replacement	0.951	0.907	0.929	0.605	0.967

Table 5: Results for ASo using Taipei dictionary

Operation	R	P	F	R_{OOV}	R_{IV}
FMM only	0.954	0.923	0.938	0.800	0.963
BMM, plus heuristic rules	0.969	0.941	0.955	0.814	0.978
Comparison and replacement	0.971	0.946	0.958	0.815	0.981

Table 6: Results for PKo using Beijing dictionary

The results show that dictionaries derived from specific communities alone yield slightly smaller F-measures than that derived from all places together. The largest difference lies in R_{OOV} where it is 0.605 and 0.663 for ASo and 0.815 and 0.851 for PKo, confirming the significance of adopting a large and all-rounded dictionary in word segmentation.

4 Error Analysis

We have examined the discrepancies between the gold standard files and our resultant segmented files, and it is found that the segmentation errors can be basically classified into several categories.

The errors due to standard divergence have the most impact. For example, 性/取向 is considered the correct segmentation in the AS test corpus while 性取向 is one word in our large dictionary.

Inconsistencies within the same corpus (both training and test corpora) also give rise to performance fluctuations. There are cases where the same phrase is segmented differently. For example, in the AS training corpus, both 藝術/工作/者 and 藝術/工作者 are found. Similar cases are also found in the test corpus, e.g. 性/工作/者 vs. 性/工作者.

Another factor that affects the segmentation performance over the PK corpus is encoding conversion. Our production system is based primarily on materials which are in BIG5 encoding, specifically traditional Chinese characters in the BIG5 encoding space. Since the given test data are in simplified Chinese characters, a process of encoding conversion to BIG5 is in place. Such a conversion is a one-to-many mapping and thus some original words will be distorted, influencing segmentation correctness.

5 Conclusion

We have reported our results on two open tracks of the Second International Chinese Word Segmentation Bakeoff, based on a production segmentation system, which draws heavily on a large and unique dictionary. The dictionary is derived from processing a very large amount of synchronous textual data gathered from various Chinese speech communities, based on a uniform segmentation standard. It is shown that the primary dictionary-based BMM segmentation alone contribute the most in our segmentation system, with over 95% in recall and over 90% in precision, attributable to the large size of the dictionary, although our uniform segmentation standard may not have realized its full potential given the test corpora with different and changing standards. We also explored supplementary features offered by the large dictionary in post-processing, and results incrementally improve. Hence our large dictionary derived from our uniform treatment of synchronous data provides a useful resource and provides a good platform for further extension in various aspects of language engineering.

References

Richard Sproat and Tom Emerson. 2003. *The First International Chinese Word Segmentation Bakeoff*. In proceedings of the Second SIGHAN Workshop on Chinese Language Processing, July 11-12, 2003, Sapporo, Japan.