Segmenting Chinese Unknown Words by Heuristic Method

Christopher C. Yang and K. W. Li

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong

Abstract. Chinese text segmentation is important in Chinese text indexing. Due to the lack of word delimiters in Chinese text, Chinese text segmentation is more difficult than English text segmentation. Besides, the segmentation ambiguities and the occurrences of out-of-vocabulary words (i.e. unknown words) are the major challenges in Chinese segmentation. Many research works dealing with the problem of word segmentation have focused on the resolution of segmentation ambiguities. The problem of unknown word identification has not drawn much attention. In this paper, we propose a heuristic method for Chinese test segmentation based on the statistical approach. The experimental result shows that our proposed heuristic method is promising to segment the unknown words as well as the known words. We have further investigated the distribution of the errors of commission and the errors of omission caused by the proposed heuristic method and benchmarked the proposed heuristic method with our previous proposed technique, boundary detection.

1. Introduction

Text segmentation is defined as the segmentation of texts into linguistic units, normally words [8]. The problem can be formally defined as [6]

$$\arg\max_{W_i} P(W_i \mid C) = \arg\max_{W_i} P(W_i)P(C \mid W_i) / P(C) \quad (1)$$

where $C=c_1c_2\ldots c_m$ is an input character string, and
$W_i=w_1w_2\ldots w_n$ is a possible word segmentation.

Since $P(C|W_i)$ equals to 1 and P(C) is a constant, the above formulation can be simplified to

$$\arg\max_{W_i} P(W_i \mid C) = \arg\max_{W_i} P(W_i) \quad (2)$$

Unlike English, many Asian languages (e.g. Chinese, Japanese and Thai) do not have delimiters of words as spaces or punctuation marks. As a result, segmenting

Chinese text is a more difficult task than segmenting English text. The techniques to segment Chinese character sequences into words can be divided into three categories: (a) statistical approach [2][7][6], (b) lexical rule-based approach [9], and (c) hybrid approach based on statistical and lexical information.

The segmentation ambiguities and the occurrences of out-of-vocabulary words (i.e. unknown words) are the most challenging problems in Chinese text segmentation. Many research works dealing with the problem of word segmentation have focused on the resolution of segmentation ambiguities. The problem of unknown word identification has not drawn much attention. The unknown words are diverse, including personal names, organization names and their abbreviations. The unknown words are defined as the words which are not found in the lexicon [1], but they provide more precise and comprehensive meaningful terms for information retrieval. Lai and Wu [4] referred the unknown words or phrases as phrase-like units (PLU) that can be combinations of words in the lexicon or some meaningless characters. Due to the ever-changing nature of language, no general lexicon can be comprehensive. New words or phrases are created everyday. In this paper, we propose a heuristic method based on the statistical approach to segment Chinese text. In particular, the capability of the proposed technique to segment unknown words is investigated.

The followings are some types of unknown words that frequently occur:

1. *Abbreviation* (acronym): Abbreviations are difficult to be identified since their morphological structures are very irregular. Their affixes reflect the conventions of the selection meaningful components [1]. However, the affixed of abbreviations are common words which are least informative for indicating the existence of unknown words. For example, 中大 (CU) is the abbreviation of 香港中文大學 (The Chinese University of Hong Kong).
2. *Proper nouns*: Proper nouns can be classified into 4 subcategories: 1) name of people; 2) name of place; 3) name of organization; 4) specific terms, e.g. 大腸桿菌 , 脫氧核糖核酸, 可卡因, 目眩神迷. Certain key words are indicators for each different subcategory. For instance, there are about 100 common surnames which are prefix characters of Chinese personal names. The characters, such as 道, 區, 路, frequently occur as suffixes of the names of places. However, names such as 香港仔, 黃竹坑, 粉嶺 are hard to identify.
3. *Derived words*: The derived words have affix morphemes which are strong indicators of unknown words, e.g. 電腦化.
4. *Compounds*: The compounds are very productive type of unknown word. Nominal and verbal compounds are easily formed by combining two or more words, e.g. 邊境禁區, 胡琴演奏會, 爆竊集團.

2 Statistical Approach

In the statistical approach of Chinese text segmentation, a text corpus is analyzed and the context specific information about syntactic structures and usage of words are obtained. Association formulae are utilized to measure the association among

adjacent characters in Chinese text. The most popular association formulae are mutual information (MI) and significance estimation (SE).

Mutual information (MI):

Mutual information is first adopted by Sproat and Shih [7] to measure the association between two adjacent characters

$$MI(a,b) = \log_2 \frac{\mathrm{N}\ freq(ab)}{freq(a)\,freq(b)} \tag{3}$$

where a and b are Chinese characters, N is the size of corpus, *freq(ab)* is the frequency of occurrence of the character string *ab*, and *freq(a)* and *frea(b)* are the frequencies of occurrence of *a* and *b*, respectively.

Significant Estimation (SE):

Significant estimation is adopted by Chien [2] to measure the association of n characters where n can be any values greater than two.

$$SE(c) = \frac{freq(c)}{freq(a) + freq(b) - freq(c)} \tag{4}$$

where c is a string with n characters, a and b are two overlapping substrings of c with n-1 characters.

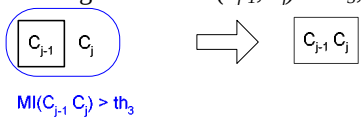
If $c=c_jc_{j+1}c_{j+2}$,

$$SE(c_jc_{j+1}c_{j+2}) = \frac{freq(c_jc_{j+1}c_{j+2})}{freq(c_jc_{j+1}) + freq(c_{j+1}c_{j+2}) - freq(c_jc_{j+1}c_{j+2})} \tag{5}$$

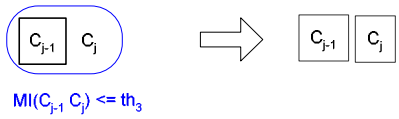
3 Heuristic Method

We propose a heuristic method with five rules to segment Chinese text using mutual information and significance estimation of all bi-grams and tri-grams, respectively, in the corpus. Given a Chinese strings with n characters, $c_1\ c_2 \ldots c_j \ldots c_n$, every character is initialized as a unigram. The heuristic method begins from the second character, c_2, and determines the matching rules. When a rule is matched, c_j is combined with the previous n-gram(s) to form a longer n-gram or remain as a unigram. If there is no other rule can be matched, the next character, c_{j+1}, will be considered. This process is repeated until the last character, c_n, is reached.

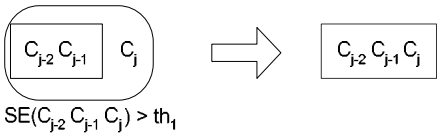
Rule 1:
If c_{j-1} is a unigram and $MI(c_{i-1}, c_i) > th_3$, c_{j-1} and c_j are combined as a bi-gram.



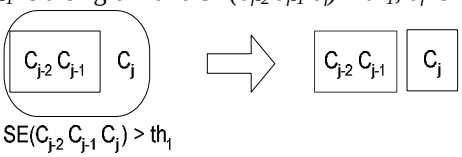
If c_{j-1} is a unigram and $MI(c_{j-1}, c_j) \leq th_3$, c_{j-1} and c_j remain as unigrams.



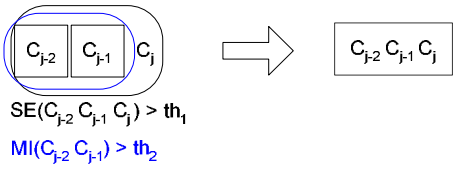
Rule 2:
If $c_{j-2}\ c_{j-1}$ is a bi-gram and $SE(c_{j-2}\ c_{j-1}\ c_j) > th_1$, $c_{j-2}\ c_{j-1}$ and c_j are combined as a tri-gram.



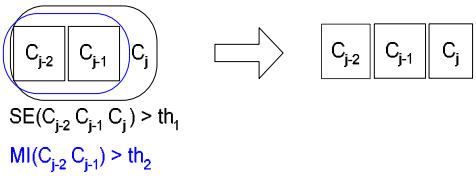
If $c_{j-2}\ c_{j-1}$ is a bi-gram and $SE(c_{i-2}\ c_{i-1}\ c_i) \leq th_1$, c_i is remained as a unigram.



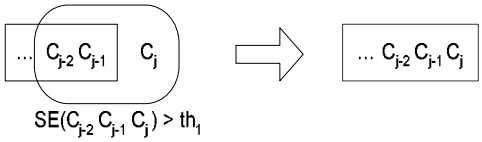
Rule 3:
If c_{j-2} and c_{j-1} are unigrams, $SE(c_{j-2}\ c_{j-1}\ c_j) > th_1$, and $MI(c_{j-2}\ c_{j-1}) > th_2$, $c_{j-2,}\ c_{j-1}$ and c_j are combined as a tri-gram.



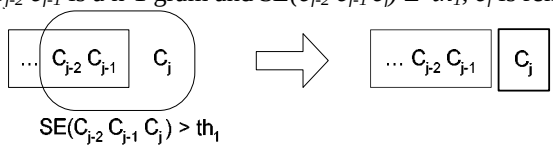
If c_{j-2} and c_{j-1} are unigrams, and $SE(c_{j-2}\ c_{j-1}\ c_j) \leq th_1$ or $MI(c_{j-2}\ c_{j-1}) \leq th_2$, c_j is remained as a unigram.



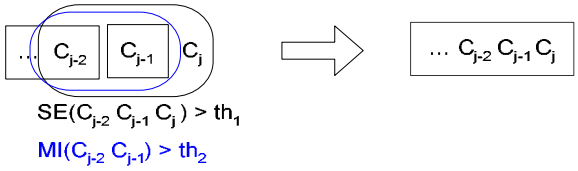
Rule 4:
If …c_{j-2} c_{j-1} is a n-1 gram and $SE(c_{j-2}\ c_{j-1}\ c_j) > th_1$, …$c_{j-2}$ c_{j-1} c_j are combined as a n-gram.



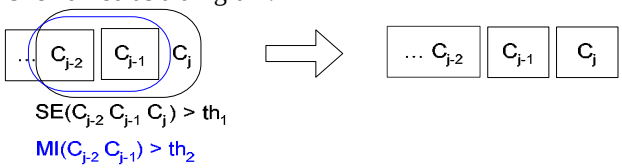
If …c_{j-2} c_{j-1} is a n-1 gram and $SE(c_{j-2}\ c_{i-1}\ c_i) \leq th_1$, c_i is remained as a unigram.



Rule 5:
If …c_{j-2} is a n-2 gram, c_{j-1} is a unigram, $SE(c_{j-2}\ c_{j-1}\ c_j) > th_1$, and $MI(c_{j-2}\ c_{j-1}) > th_2$, …$c_{j-2}$ c_{j-1} c_j are combined as a n-gram.



If …c_{j-2} is a n-2 gram, c_{j-1} is a unigram, and $SE(c_{j-2}\ c_{j-1}\ c_j) \leq th_1$, or $MI(c_{j-2}\ c_{j-1}) \leq th_2$, c_j is remained as a unigram.



th_1, th_2, and th_3 are thresholds determined experimentally.

3.1. Example

Given a Chinese sentence, 財政司司長發表財政預算案的日子, the mutual information for all bi-grams and the significant estimation for all tri-grams are as follows:

Bi-gram	財政	政司	司司	司長	長發	發表	表財	財政
MI	6.83	4.27	4.49	4.69	0.51	3.80	-1.42	6.83
Bi-gram	政預	預算	算案	案的	的日	日子		
MI	3.43	7.65	5.79	0.73	0.82	6.43		

Tri-gram	財政司	政司司	司司長	司長發	長發表
SE	0.253	0.540	0.806	0.005	0.006
Tri-gram	財政預	政預算	預算案	算案的	案的日
SE	0.128	0.340	0.463	0.103	0.007

th_1, th_2, and th_3 are 0.3, 3.0, and 1.0.

Using Rule 1, *MI(*財政*)* > th_3, [財] and [政] are combined as a bi-gram, [財政].
Result: [財政]
Using Rule 2, *SE(*財政司*)* ≤ th_1, [司] is remained as a unigram.
Result: [財政][司]
Using Rule 5, *SE(*政司司*)* > th_1 and *MI(*政司*)* > th_2, [財政], [司] and [司] are combined to form [財政司司].
Result: [財政司司]
Using Rule 4, *SE(*司司長*)* > th_1, [財政司司] and [長] are combined as [財政司司長].
Result: [財政司司長]
Using Rule 4, *SE(*司長發*)* ≤ th_1, [發] is remained as a unigram.
Result: [財政司司長][發]
Using Rule 5, *SE(*長發表*)* ≤ th_1, [表] is remained as a unigram.
Result: [財政司司長][發][表]
Using Rule 1, *MI(*發表*)* > th_3, [發] and [表] are combined as a bi-gram, [發表].
Result: [財政司司長][發表]
Using Rule 2, *SE(*發表財*)* ≤ th_1, [財] is remained as a unigram.
Result: [財政司司長][發表][財]
Using Rule 5, *SE(*表財政*)* ≤ th_1, [政] is remained as a unigram.
Result: [財政司司長][發表][財][政]
Using Rule 1, *MI(*財政*)* > th_3, [財] and [政] are combined as a bi-gram, [財政].
Result: [財政司司長][發表][財政]
Using Rule 2, *SE(*財政預*)* ≤ th_1, [預] is remained as a unigram.
Result: [財政司司長][發表][財政][預]

Using Rule 5, *SE(*政預算*)* > th_1 and *MI(*政預*)* > th_2, [財政], [預] and [算] are combined as [財政預算].

Result: [財政司司長][發表][財政預算]

Using Rule 4, *SE(*預算案*)* > th_1, [財政預算] and [案] are combined as[財政預算案].

Result: [財政司司長][發表][財政預算案]

Using Rule 4, *SE(*算案的*)* ≤ th_1, [的] is remained as a unigram.

Result: [財政司司長][發表][財政預算案][的]

Using Rule 5, *SE(*案的日*)* ≤ th_1, [日] is remained as a unigram.

Result: [財政司司長][發表][財政預算案][的][日]

Using Rule 1, *MI(*的日*)* ≤ th_3, [日] is remained as a unigram.

Result: [財政司司長][發表][財政預算案][的][日]

Using Rule 3, *SE(*的日子*)* ≤ th_1, [子] is remained as a unigram.

Result: [財政司司長][發表][財政預算案][的][日][子]

Using Rule 1, *MI(*日子*)* > th_3, [日] and [子] are combined as a bi-gram, [日子].

Result: [財政司司長][發表][財政預算案][的][日子]

In this example, 財政司司長 and 財政預算案 are unknown words that are compound words formed by the combination of known words, 財政司 and 司長, and 財政 and 預算案, respectively. 發表 and 日子 are known words. It shows that the proposed heuristic method can successfully segment both the unknown words which is impossible to be identified by dictionary.

4 Experiments

Experiments are conducted to measure the precision, recall and error rate of the segmentation results using the Hong Kong local Chinese news articles and HKSAR government press releases as corpus. The formulations of precision, recall, and error rate are given as below:

$$\text{Precision} = \frac{c}{n} \qquad \text{Recall} = \frac{c}{N} \qquad \textit{Error rate} = \frac{e}{N}$$

where *c* is the number of words that are correctly segmented,
e is the number of words that are incorrectly segmented,
n is the number of words segmented,
and *N* is the number of words in the corpus.

In the corpus of our experiment, there are totally 2000 documents. The number of known words and the number of unique known words are 317,386 and 44,189, respectively. The number of unknown words and the number of unique unknown words are 108,296 and 30,792, respectively. Many of the unknown words are names of persons, events, organizations, and technical terms. The average frequency of the known words and unknown words are 7.18 and 3.52, respectively. The average

frequency of the known words is double of the average frequency of the unknown words.

4.1 Performance and Benchmarking

In this section, we present the performance of the proposed heuristic methods and benchmark its performance with that of our previous proposed technique, boundary detection [9]. The boundary detection is developed to detect the segmentation points in a Chinese sentence based on the abrupt changes on the mutual information between the adjacent bi-grams [9]. It is shown by experiment results that the boundary detection technique is efficient and effective. It only requires the mutual information but not the significance estimation.

Table 1 and Table 2 present the experimental result of the boundary detection and heuristic method for unknown words and known words, respectively.

Table 1 – Performance of Boundary Detection and Heuristic Method for segmenting unknown words

Algorithms	Precision	Recall	Error rate
Boundary Detection	0.801	0.752	0.187
Heuristic Method	0.918	0.903	0.081

Table 2 – Performance of Boundary Detection and Heuristic Method for segmenting known words

Algorithms	Precision	Recall	Error rate
Boundary Detection	0.812	0.841	0.195
Heuristic Method	0.897	0.919	0.105

Table 3 presents the overall experimental result of boundary detection and heuristic method for both known words and unknown words and their efficiency.

Table 3 – Overall Performance of Boundary Detection and Boundary Heuristic Method

Algorithms	Precision	Recall	Error rate	Processing Time used in seconds
Boundary Detection	0.809	0.818	0.193	89.6
Heuristic Method	0.902	0.915	0.099	254.5

The result shows that the heuristic method is promising to segment the unknown words as well as the know words. Besides, the heuristic method is significantly better than the boundary detection in terms of precision and recall. It can be explained by the fact that the heuristic method utilizes both mutual information of bi-grams and significant estimation of tri-grams to form n-grams, but the boundary detection only utilize the mutual information on bi-grams to determine the segmentation points. Given more statistical information, the heuristic method is able to achieve higher performance. On the other hand, the heuristic method is significantly more time consuming than the boundary detection because it involves the computation of both

mutual information and significant estimation. The rules matching in heuristic matching is also more computational expensive than identifying segmentation point in boundary detection.

4.2 Analysis of Error

The errors arose from using the boundary detection and heuristic method were analyzed based on the methods suggested by Dai et al. [3][4]. The errors in Chinese text segmentation can be divided into two types: a) errors of commission and b) errors of omission. The errors of commission refer to the segments that are identified by the automatic text segmentation techniques as words but in fact they are not. The errors of omission are words that are not identified by the automatic text segmentation techniques to be words but in fact they are. Table 4 presents the distribution of the errors of commission and the errors of omission caused by the boundary detection and the heuristic method.

Table 4. The Distribution of the Errors of Commission and the Errors of Omission

	Errors of Commission	Errors of Omission
Boundary Detection	93.5 %	6.5 %
Heuristic Method	84.9%	15.1 %

The result shows that both of the boundary detection and the heuristic method have significantly more percentage of the errors of commissions. Comparing between the boundary detection and the heuristic method, the heuristic method has relatively less percentage of the errors of commission and relatively more percentage of the errors of omission. It is due to the fact that the boundary detection determines the segmentation points purely by the abrupt changes of mutual information of adjacent Chinese characters. However, the abrupt changes may not always correspond to segmentation points especially when the statistical information provided by the corpus is not sufficient. Therefore, the boundary detection may identify segmentation points that indeed do not exist.

The error of commission can be further categorized into two types:

E1. A word is incorrectly segmented into segments that are simple words
For example, 中文大學 can be incorrectly segmented to [中文] and [大學]. [中文] and [大學] are both simple words

E2. A word is incorrectly segmented into segments such that one or both of the segments is not a word.
For example, 財政司司長 can be incorrectly segmented to [財政] and [司司長]. [財政] is a word but [司司長] is not a word.

Table 5 presents the distribution of E1 and E2 errors in the errors of commission caused by the boundary detection and the heuristic method.

Table 5.The Distribution of E1 and E2 Errors in the Errors of Commission

	Errors of Commission	
	E1	E2
Boundary Detection	57.0%	43.0%
Heuristic Method	55.5%	44.5%

The result shows that the percentage of E1 errors is higher that that of E2 errors in the errors of commission caused by both of the boundary detection and the heuristic method. The E1 errors are hard to be avoided especially when the frequencies of both of the simple words, which are incorrectly segmented, are relatively higher than the frequency of the word formed by these simple words. Using the statistical approach, both of the boundary detection and the heuristic method take the simple words as more probable than the compound words formed by these simple words. Unless the compound words appear more frequently in the corpus, it is impossible to be avoided. It is also found that the distributions of E1 and E2 errors in the errors of commission caused by the boundary detection and the heuristic method are approximately the same. None of these two techniques have advantages in resolving either type of errors.

5. Conclusion

In this paper, we propose a heuristic method for Chinese text segmentation using statistical approach. Such method employs mutual information and significance estimation to measure the association among adjacent characters and utilizes five rules to determine the segmentation points. Experiment results show that the heuristic method is promising to segment unknown words as well as known words. The heuristic method is significantly better than our previous proposed boundary detection. Error of analysis has also been presented. It is found that the percentage of errors of commission is significantly more than the percentage of the errors of omission. In terms of the errors of commission, there are more errors caused by segmenting compound words into simple words.

References

1. Chen, K. and Bai, M., "Unknown Word Detection for Chinese by a Corpus-base Learning Method," *Computational Linguistics and Chinese Language Processing*, vol.3, no.1, February, 1998, pp.27-44.
2. Chien, L.F., "Fast and quasi-natural language search for gigabits of Chinese texts," *Research and Development in Information Retrieval, ACM-SIGIR*, Seattle, 1995, pp.112-120.
3. Dai, Y., Khoo, C. S. G. and Loh, T. E., "A New Statistical Formula for Chinese Text Segmentation Incorporating Contextual Information," *ACM SIGIR*, 1999.
4. Khoo, C. S. G., Dai, Y., and Loh, T. E., "Using Statistical and Contextual Information to Identify Two-and Three-Character Words in Chinese Text," *Journal of the American Society for Information Science and Technology*, 53(3), pp.365-377, 2002.

5. Lai, Yu-sheng and Wu, Chung-hsien, "Unknown Word and Phrase Extraction Using a Phrase-Like-Unit-Based Likelihood Ratio", *International Journal of Computer Processing of Oriental Languages*, Vol. 13, No. 1, 2000, pp.83-95.
6. Meknavin, S., Charoenpornsawat, P., Kijsirikul, B., "Feature-based Thai Word Segmentation," *Natural Language Processing Pacific Rim Symposium* (NLPRS'97),1997
7. Sproat, R. and Shih, C., "A Statistical Method for Finding Word Boundaries in Chinese Text," *Computer Processing of Chinese and Oriental Languages*, 4, 1990, pp.336-351.
8. Wu, Zimin and Tseng, Gwyneth, "ACTS: An Automatic Chinese Text Segmentation System for Full Text Retrieval," *Journal of The American Society for Information Science*, 46(2):83-96, 1995.
9. C. C. Yang, J. Luk, S. Yung, and J. Yen, "Combination and Boundary Detection Approach for Chinese Indexing," *Journal of the American Society for Information Science, Special Topic Issue on Digital Libraries*, vol.51, no.4, March, 2000, pp.340-351.
10. Yeh, C.L. and Lee, H. J., "Rule-based Word Identification for Mandarin Chinese Sentences –A Unification Approach," *Computer Processing of Chinese and Oriental Languages*, vol.5, no.2, pp.97-118, 1991.