Error Analysis of Chinese Text Segmentation using Statistical Approach

Christopher C. Yang
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
+852 2609 8239
yang@se.cuhk.edu.hk

K. W. Li
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
kwli@se.cuhk.edu.hk

ABSTRACT
Chinese text segmentation is important for the indexing of Chinese documents, which has significant impact on the performance of Chinese information retrieval. The statistical approach overcomes the limitations of the dictionary based approach. The statistical approach is developed by utilizing the statistical information about the association of adjacent characters in Chinese text collected from the Chinese corpus. Both known words and unknown words can be segmented by the statistical approach. However, errors may occur due to the limitation of the corpus. In this work, we have conducted the error analysis of two Chinese text segmentation techniques using statistical approach, namely, boundary detection and heuristic method. We define two types of errors, errors of commission and errors of omission. It is found that the amount of the errors of commission is significantly more than the amount of the errors of omission. Further analysis of these errors is presented. Such error analysis is useful for the future development of automatic text segmentation of Chinese text or other text in oriental languages. It is also helpful to understand the impact of these errors on information retrieval system in digital libraries.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – Indexing methods. H.3.7. [Information Storage and Retrieval]: Digital Libraries

Keywords
Text Segmentation, indexing, statistical approach, multilingual processing.

1. INTRODUCTION
Digital libraries in languages other English have been developed in many countries. Information retrieval is a major component in any digital libraries, where indexing is an inevitable process. 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.

Dictionary based approach and statistical approach are two major approaches in Chinese text segmentation. In dictionary based approach, a lexicon containing a large number of Chinese words is pre-defined and heuristic methods are utilized to match against the lexicon to segment a Chinese sentence [22]. Maximum matching [6] is a typical heuristic method. The maximum matching groups the longest initial sequence of characters, which matches a dictionary entry, as a word starting from the beginning or the end of a sentence. When a word is found, it continues at the next character until all the characters in the sentence have been covered. An alternative of maximum matching is minimum matching. However, the performance of the dictionary based approach highly depends on the completeness of the coverage of the adopted lexicon. It is impossible to include all the Chinese words (known words and unknown words) in a dictionary since the set of words is open-ended [1]. Building a lexicon is also expensive and time consuming. In addition, a Chinese character sequence is always segmented in the same way regardless of context by the lexical rule based approach [7].

In the statistical approach, the lexical statistics of the Chinese characters in corpora are utilized to mark the boundaries of words [8]. It overcomes the limitations of the dictionary-based approach. Sproat and Shih [15] extract bi-grams with the highest value of mutual information in the sentence recursively until no other bi-grams can be extracted. Sproat and Shih’s technique was later evaluated by Chen et al. [1] and found that it performed better than the maximum matching. However, Sproat and Shih’s technique can only segment uni-grams and bi-grams. Chieh [3][4] developed a PAT tree based method for segmentation. All of the lexical patterns without limitation of pattern length are first extracted. A mutual information based filter algorithm is then utilized to filter out the character strings in a PAT tree. The performance is good but building a PAT tree is time consuming and large space overhead is required. Yang et al. [19] developed a boundary detection technique. It detects the segmentation points by a threshold and the abrupt changes in the values of mutual information between adjacent characters in a Chinese sentence. High accuracy and efficiency is achieved and, it is not restricted to segment only uni-grams and bi-grams.

A major advantage of the statistical approach is the capability to segment unknown words (out-of-vocabulary words). The unknown words are defined as the words which are not found in
the lexicon [2], but they provide more precise and comprehensive meaningful terms for information retrieval. Lai and Wu [9] referred the unknown words or phrases as phrase-like units (PLU) that can be combinations of words in the lexicon or some meaningless characters. The unknown words are diverse, including personal names, organization names and their abbreviations. Due to the ever-changing nature of language, no general lexicon can be comprehensive. New words or phrases are created everyday. Some typical examples of unknown words are i) abbreviation, e.g. 中文大学 (The Chinese University of Hong Kong), ii) proper nouns, e.g. name of organization, and iii) derived words, and iv) compound words.

Two statistical based Chinese text segmentation techniques have been developed by Yang et al. [19] [21], namely, boundary detection and heuristic method. Due to the limitation of the lexical statistics collected from the Chinese corpus, errors may occur in segmentation. In this work, error analysis is conducted to investigate the distributions and properties of two major types of errors, errors of commission and errors of omission.

In the next few sections, we first introduce the association formulae in statistical approach of Chinese text segmentation and the segmentation techniques. The definitions of errors are then presented. Results of error analysis and discussion will be provided at the end.

2. STATISTICAL APPROACH IN CHINESE TEXT SEGMENTATION

The statistical approach of the Chinese text segmentation is developed based on the lexical statistics collected from the Chinese corpus. Association formulae are utilized to measure the association among adjacent characters in Chinese text. Mutual information (MI) and significance estimation (SE) are two popular association formulae. Their formulations are given as bellow

\[
MI(a, b) = \log_2 \frac{N \cdot freq(ab)}{freq(a) \cdot freq(b)}
\]

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 \(freq(b)\) are the frequencies of occurrence of a and b, respectively.

\[
SE(c) = \frac{freq(c)}{freq(a) + freq(b) - freq(c)}
\]

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

2.1 Boundary detection

The boundary detection technique uses the mutual information to determine the segmentation points in a Chinese sentence. Thresholding and abrupt changes of the values of mutual information are utilized for the detection of segmentation points as illustrated in Figure 1. The abrupt changes are characterized as valleys and the bowl shape curves.

![Figure 1. Detection of segmentation points by boundary detection technique](image-url)
2.2 Heuristic Method

Yang et al. has recently developed the heuristic method with five rules to segment Chinese text using mutual information and significance estimation [21]. Given a Chinese string 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.

The five rules are given as below:

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

\[
\begin{array}{c}
\text{\( \bullet \) } \ c_j \ \ c_j \\
\text{MI(\( c_{j-1}, c_j \)) > th_3}
\end{array}
\]

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.

\[
\begin{array}{c}
\text{\( \bullet \) } \ c_j \\
\text{MI(\( c_{j-1}, c_j \)) \leq th_3}
\end{array}
\]

**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.

\[
\begin{array}{c}
\text{\( \bullet \) } \ c_{j-1} \ c_j \\
\text{SE(\( c_{j-2} \ c_{j-1} \ c_j \)) > th_1}
\end{array}
\]

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

\[
\begin{array}{c}
\text{\( \bullet \) } \ c_j \\
\text{SE(\( c_{j-2} \ c_{j-1} \ c_j \)) \leq th_1}
\end{array}
\]

**Rule 3:**
If \( c_{j-2} \ c_{j-1} \) 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.

\[
\begin{array}{c}
\text{\( \bullet \) } \ c_{j-1} \ c_j \\
\text{SE(\( c_{j-2} \ c_{j-1} \ c_j \)) > th_1}
\end{array}
\]

If \( c_{j-2} \ c_{j-1} \) 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.

\[
\begin{array}{c}
\text{\( \bullet \) } \ c_j \\
\text{SE(\( c_{j-2} \ c_{j-1} \ c_j \)) \leq th_1}
\end{array}
\]

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

\[
\begin{array}{c}
\text{\( \bullet \) } \ c_{j-1} \ c_j \\
\text{SE(\( c_{j-2} \ c_{j-1} \ c_j \)) > th_1}
\end{array}
\]

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

\[
\begin{array}{c}
\text{\( \bullet \) } \ c_j \\
\text{SE(\( c_{j-2} \ c_{j-1} \ c_j \)) \leq th_1}
\end{array}
\]

**Rule 5:**
If \( \ldots \ c_{j-2} \ c_{j-1} \ c_j \) is a n-2 gram 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.

\[
\begin{array}{c}
\text{\( \bullet \) } \ c_j \\
\text{SE(\( c_{j-2} \ c_{j-1} \ c_j \)) \leq th_1}
\end{array}
\]

\[
\begin{array}{c}
\text{\( \bullet \) } \ c_j \\
\text{MI(\( c_{j-2} \ c_{j-1} \)) \leq th_2}
\end{array}
\]

\( th_1, th_2, \) and \( th_3 \) are thresholds determined experimentally.

2.3 Precision and Recall

In our previous experiment using our corpus of Hong Kong local Chinese new articles and HKSAR government press releases with about 2000 documents, it is found that the precision and recall for boundary detection are 0.809 and 0.818, respectively, and the precision and recall for heuristic method are 0.897 and 0.919, respectively. The error rates for boundary detection and heuristic method are 0.195 and 0.105, respectively.

The formulations of the metrics, precision, recall, and error rate are given as below:

\[
\text{Precision} = \frac{c}{n}
\]

\[
\text{Recall} = \frac{c}{N}
\]

\[
\text{Error rate} = \frac{e}{N}
\]

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.
It is also found that there is not any significant difference between
the precision on segmenting known and unknown words. Using
boundary detection, the precision of segmenting known words and
unknown words are 0.812 and 0.801, respectively. Using the
heuristic method, the precision of segmenting known words and
unknown words are 0.897 and 0.918, respectively. It shows that
the statistical approach is able to segment the known words as
well as the unknown words.

The heuristic method is performing better than the boundary
detection because it uses more statistical information. The
heuristic uses both significance estimation of tri-grams and mutual
information of bi-grams, however, the boundary detection only
uses the mutual information of bi-grams. Both heuristic method
and boundary detection have reasonably high performance.

In the next sections, analysis of errors is conducted to investigate
what types of errors are frequently produced by the statistical
approach. Such error analysis is useful to identify the problems
in the process of automatic text segmentation and the impact of
the errors on the information retrieval systems.

3. ERROR ANALYSIS
3.1 Types of Errors
The errors in Chinese text segmentation can be categorized into
two major types of errors, i) errors of commission and ii) errors of
omission [5]. 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.

3.2 Error of Commission
The errors committed by the automatic segmentation techniques
produce two segments that are not supposed to exist. These
segments can be known words, unknown words, or not even a
word. Based on these possibilities, we further categorized the
errors of commission into six types of errors, EC1 to EC6, as follows:

EC1. A word is incorrectly segmented into two segments.
    Both segments are not words.
    For example, an unknown word 可卡因 (cocaïne) can be
    incorrectly segmented to 可 卡 and 因. Both 可 卡 and 因
    are not words.

EC2. A word is incorrectly segmented into two segments.
    Both segments are known words.
    For example, a compound word 運輸 系統
    (transportation system) can be incorrectly segmented to
    two simple words 運輸 (transport) and 系統 (system).

EC3. A word is incorrectly segmented into two segments.
    Both segments are unknown words.
    For example, the unknown word 廣 州 香 港 快 線
    (Guangzhou-Hong Kong Express Rail) can be
    incorrectly segmented into two unknown words, 廣 州
    (Guangzhou) and 香 港 快 線 (Hong Kong Express Rail).

EC4. A word is incorrectly segmented into two segments.
    One segment is a known word and another segment is
    not a word.
    For example, 現代化 (modernize) can be incorrectly
    segmented to a simple word 現代 (modern) and a single
    character 化 先生 们 (gentlemen) can be incorrectly
    segmented to a simple word 先生 (Mr.) and a pluralizing
    suffix 們. A Chinese name 陳 明 珠 can be incorrectly
    segmented to 陳 (a Chinese surname but not a word) and
    a simple word 明 珠 (Pearl). An unknown word 吳 主 席
    (Chairman Ng) can be incorrectly segmented to 吳 (a
    Chinese surname but not a word) and a simple word 主
    席 (chairman).

EC5. A word is incorrectly segmented into two segments.
    One segment is an unknown word and another segment is
    not a word.
    For example, the unknown word 豐 田 碼 (Toyota) can
    be incorrectly segmented to an unknown word 豐 田
    (Toyota) and a segment 品牌 (brand). An unknown word
    非 欧 盟 (non-Euro) can be incorrectly segmented as a
    segment 非 (non) and an unknown word 欧 盟 (Euro).
    A name of a person 馮 約翰 (John Bond, chairman of
    HSBC Holdings) can be incorrectly segmented as a
    segment 馮 (a Chinese surname) and 約 翰 (Chinese
    translation of John).

EC6. A word is incorrectly segmented into two segments.
    One segment is a known word and another segment is
    an unknown word.
    For example, the unknown word 指 數 (Nasdaq Index) can be
    incorrectly segmented to an unknown word 指 數 (the Chinese
    translation of Nasdaq) and a known word 指 數 (index). The
    unknown word 銀 行 (Standard Chartered PLC)
    can be incorrectly segmented to an unknown word 銀
    行 (Standard Chartered) and a known word 業
    (bank).

3.3 Error of Omission
A segmentation point can be omitted by the automatic
segmentation techniques to form a string of characters that is
supposed to be decomposed into two segments. These segments
can be known words or unknown words. We further categorize
the errors of omission into three types of errors, EO1 to EO3, as follows:

EO1. The incorrect segment includes two unknown words.
    For example, the incorrect segment 和 黃 長 實
    (Hutchison Whampoa and Cheung Kong) should be
    segmented to and 黃 (the Chinese abbreviation of
    Hutchison Whampoa Ltd. and 黃 埠 有 限 公 司) and
    長 實 (the Chinese abbreviation of Cheung Kong
    (Holdings) Limited 長 江 實 業 (集團) 有 限 公 司.
    Hutchison Whampoa and Cheung Kong are two
    independent organizations but closely related.
    Therefore, 和 黃 and 長 實 are frequently placed together
    without delimiters. It is difficult to determine the
segamentation points between them by statistical approach.

EO2. The incorrect segment includes two known words.
For example, the incorrect segment 污染造成(pollution causes) should be segmented as two simple words 污染 (pollution) and 造成(cause).

EO3. The incorrect segment includes a known word and an unknown word.
For example, the incorrect segment 攻打伊拉克(attack Iraq) should be segmented as a known word 攻打 (attack) and an unknown word 伊拉克(Iraq).

3.4 Result of Error Analysis
Based on the definitions of different types of errors in Chinese text segmentation, we conduct the analysis of errors for the boundary detection and heuristic method as described in Section 2. It is found that there are significant more errors of commission than errors of omission for both boundary detection and heuristic method. The distributions of errors of commission and errors of omission are provided in Table 1.

Table 1. The Distribution of the Errors of Commission and the Errors of Omission

<table>
<thead>
<tr>
<th></th>
<th>Errors of Commission</th>
<th>Errors of Omission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boundary Detection</td>
<td>98.13 %</td>
<td>1.87 %</td>
</tr>
<tr>
<td>Heuristic Method</td>
<td>94.23 %</td>
<td>5.77 %</td>
</tr>
</tbody>
</table>

In the Chinese corpus with about 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.

The distributions of EC1 to EC6 and EO1 to EO3 errors for boundary detection and heuristic methods are given in Table 2 and Table 3.

Table 2. Distribution of EC1 to EC6

<table>
<thead>
<tr>
<th></th>
<th>Boundary Detection</th>
<th>Heuristic Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC1</td>
<td>19.32%</td>
<td>16.58%</td>
</tr>
<tr>
<td>EC2</td>
<td>38.26%</td>
<td>23.72%</td>
</tr>
<tr>
<td>EC3</td>
<td>18.72%</td>
<td>31.78%</td>
</tr>
<tr>
<td>EC4</td>
<td>9.40%</td>
<td>9.03%</td>
</tr>
<tr>
<td>EC5</td>
<td>5.09%</td>
<td>11.76%</td>
</tr>
<tr>
<td>EC6</td>
<td>9.21%</td>
<td>7.13%</td>
</tr>
</tbody>
</table>

Table 3. Distribution of EO1 to EO3

<table>
<thead>
<tr>
<th></th>
<th>Boundary Detection</th>
<th>Heuristic Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>EO1</td>
<td>51.26%</td>
<td>23.81%</td>
</tr>
<tr>
<td>EO2</td>
<td>22.76%</td>
<td>29.07%</td>
</tr>
<tr>
<td>EO3</td>
<td>25.98%</td>
<td>41.72%</td>
</tr>
</tbody>
</table>

4. DISCUSSION
In the error analysis, we find that there are only a few percentages of errors of omission produced by both boundary detection and heuristic method. The majority of the errors are errors of commission. Most of the errors by the investigated statistical approach are errors generated by determining wrong segmentation points. If we further investigate the errors of commission, some of the errors are acceptable in some text processing. It is because the two segments produced by the errors of commission can both be words, either known words or unknown words. These known words or unknown words can still be processed except that the phrases (mostly compound words) that formed by these known and unknown words can no longer be captured. However, the errors of omission are mostly not acceptable in any kinds of text processing. It is because the strings of characters that generated by missing the segmentation points are mostly not words. These strings of characters are supposed to be segmented into two words, either known words or unknown words.

To further understand the impact of these errors on information retrieval systems, we need to investigate the distribution of EC1 to
EC6. EC2, EC3, and EC4 are errors of commissions that segment words into two segments where both or either one are known words or unknown words. As discussed earlier, these errors are acceptable in some text processing. Among the errors of commission, EC2, EC3, and EC4 constitute 66.38% and 64.53% for boundary detection and heuristic method, respectively. EC5 and EC6 are errors of commissions that segment words into two segments where one segment is either a known word or an unknown word and another segment is not a word. Among the errors of commission, EC5 and EC6 constitute 14.3% and 18.89% for boundary detection and heuristic method, respectively. One of the two segments produced by these errors can still be useful in some text processing. EC1 are errors of commissions that segment words into two segments where both segments are not words. These segments cannot be used in any kinds of text processing. Among the errors of commission, EC1 constitutes 19.32% and 15.48% for boundary detection and heuristic method, respectively.

Given the above analysis, we can conclude that errors generated by EC1, EO1, EO2, and EO3 produce segments that are completely useless for text processing. Among all the errors, these errors constitute 20.83% and 21.39% for boundary detection and heuristic method, respectively. Segments that are generated by EC5 and EC6 are partially useful for text processing. These errors constitute 14.03% and 17.80% for boundary detection and heuristic method, respectively. The segments that generated by EC2, EC3, and EC4 are still useful for some text processing. Among all the errors, these errors constitute 65.13% and 60.81% for boundary detection and heuristic method, respectively. Although the highest precision and recall achieved by heuristic method are 0.897 and 0.918, respectively, it does not mean that the wrong segments are completely useless for text processing. For example, even the word “香大院校” (Chinese University) can still be conducted by matching the word “中文大学” (Chinese University).

Investigating the distribution of errors of commission helps us to understand the possible reasons of generating such errors by statistical approach. It is found that EC2 and EC3 are comparatively high for both boundary detection and heuristic method. For boundary detection, EC2 and EC3 constitute 56.98% of errors among all errors of commission. For heuristic method, EC2 and EC3 constitute 55.5% of errors among all errors of commissions. These errors are easier to be committed because the wrong segments generated are both words. In these cases, a segment \( c_1 \ldots c_i p c_{i+1} \ldots c_n \) can be incorrectly segmented as \( c_1 \ldots c_i \) and \( c_{i+1} \ldots c_n \), where \( c_{i+1} \ldots c_n \) and \( c_1 \ldots c_i \) are either known words or unknown words. The association among characters \( c_1 \ldots c_i \) and \( c_{i+1} \ldots c_n \) is not strong. If \( n \) is larger than 3, mutual information and significance estimation is not able to measure the association among \( c_1 \ldots c_i, c_{i+1} \ldots c_n \). However, the associations among the characters \( c_1 \ldots c_i \) and among the characters \( c_{i+1} \ldots c_n \) are relatively stronger in \( c_1 \ldots c_i, c_{i+1} \ldots c_n \). The wrong segments, \( c_1 \ldots c_i \) and \( c_{i+1} \ldots c_n \), are formed. It is difficult to avoid EC1, EC2, and EC3 in statistical approach of automatic text segmentation unless more statistical information can be collected from a larger corpus or association formulae that considers n-grams with larger values of \( n \) is used. However, an association formula that considers longer n-grams also implies longer processing time. The processing time may indeed exponentially increase with larger value of \( n \).

![Figure 4. Distributions of Unacceptable Errors, Partially Acceptable Errors and Acceptable Errors.](image)

**5. CONCLUSION**

In this paper, we have conducted the error analysis of two Chinese text segmentation techniques using statistical approach. These techniques are boundary detection and heuristic method. We have defined two types of errors, namely, errors of commission and errors of omission. Among the errors of commission, we have categorized them into six types of errors, EC1 to EC6. EC1 to EC6 are errors with their two incorrectly decomposed segments as known words, unknown words, or not words. Among the errors of omission, we have categorized them into three types of errors, EO1 to EO3. EO1 to EO3 are errors that formed by incorrectly combining two segments. These segments are known words or unknown words. Based on our analysis, we find that the majority of errors are errors of commissions. Among all the errors, only about 20% of errors produce segments that are completely useless for any kinds of text processing. About 15% of errors produce segments that are partially useful for some text processing. The rest of them produce segments that are still useful in some text processing. Our investigation also finds that the errors of commissions that incorrectly decompose words into two known words or two unknown words (EC2 and EC3), are difficult to be avoided in automatic text segmentation using statistical approach. Similarly, it is also difficult to avoid the errors of commission that incorrectly decompose words into two segments, where both of them are not words. Using association formulae that consider
longer n-grams may be able to resolve such problems but it also increase the processing time exponentially.

6. ACKNOWLEDGMENTS
This project was supported by the Direct Research Grant of the Chinese University of Hong Kong, 2050268, and the Earmarked Grant for Research by the Hong Kong Research Grant Council, 4335/02E.

7. REFERENCES