CHINESE-WORD SEGMENTATION BASED ON MAXIMAL-MATCHING AND BIGRAM TECHNIQUES

Luk Wing Pong, Robert (陸永邦)
Department of Chinese, Translation and Linguistics
City Polytechnic of Hong Kong
Email: CTRWPL92@CPHKVX.CPHK.HK

ABSTRACT

One of the most simple and accurate Chinese-word segmentation technique is maximal-matching. However, its performance depends on the coverage of the list of words which are usually derived from a general dictionary. When it is directly applied to segment technical articles instead of general news articles, the error rate degraded significantly from 1.2% (as in the literature) to 15%. This is an important problem in two respect. First, usually the domain-specific terms are not readily available on computer. These terms have to be entered manually by expert or they can be detected automatically from thematic corpora. Second, if corpus analysis is applied to supplement information for the design and development of text processing systems, these analysis depend on the correct word segmentation of these corpora of technical articles. In this paper, we propose to combine the maximal-matching and bigram techniques in Chinese-word segmentation for detecting words in thematic corpora, where both techniques overcome each other's short coming. The Hong Kong Basic Law was selected as a representative technical article for evaluation because it has a fair amount of technical terms, compound nouns and names. The segmentation performances of the maximal-matching, bigram and the combined techniques are compared. The combined technique was able to achieve 33% improvement in segmentation performance and identify 33% of the terms in the Basic Law.

I. INTRODUCTION

Thematic corpora are compiled to enlarge the scale of the sub-language approach [1] in text processing systems (e.g. machine translation and text retrieval), on the one hand to supplement empirical information for the design of these systems and on the other hand to evaluate these systems with authentic data. However, techniques for corpus analysis may not be appropriate for thematic corpora because they were developed to analyze general articles which are sampled across various domains. By contrast, thematic corpora sample articles of a specific domain and these articles tend to be technical, for example, constructing a machine translation system for financial reports or text retrieval system for constitutional law. An important case in point is Chinese-word segmentation which is an elementary stage of any Chinese text processing systems. For general corpora, it has been acknowledged that names and proper nouns constitute major errors in word segmentation [2,3] even though the amount of segmentation error is small, typically between 1% and 2%. However, the amount of error may increase significantly with thematic corpora since technical terms and compound nouns are more likely to occur for technical articles. Although domain-specific dictionaries (e.g. dictionary on computers) are available, the representativeness of these entries have to be evaluated using the thematic corpora. These entries are usually entered manually because current OCR technologies are not as cost-effective as professional typists, given that errors occur frequently when character size changes as in many dictionaries. An alternative is to extract a tentative list of technical terms or proper nouns from thematic corpora and the list is verified using a dictionary or extended manually using a concordance program [4].

Extracting the list is similar to detecting proper nouns and technical terms as in text retrieval for English where strongly associated words are grouped together, depending on their co-occurring frequencies or mutual information within a specified context [5]. Syntactic patterns are also used to eliminate improbable cases [6]. Detection for Chinese is simpler than for English since Chinese terms and proper nouns tend to be a sequence of consecutive characters. Detection of two-character words have already been reported in [7,8] but technical terms and proper nouns usually have more than 2 characters, particularly those that are translated. We are unaware of any report in the literature about the effectiveness of detecting two-character words in improving the segmentation performance.

The aim of this paper is to address the problem of word detection for improving the performance of Chinese word segmentation of thematic corpora. We combined both maximal-matching [9] and bigram [7]

techniques because they complement each other and they are relatively inexpensive and simple to implement compared with the relaxation [10], adaptive statistical [11] and competitive neural network [12] techniques. In addition, the combined technique does not need to estimate positions of error occurrence for each thematic corpus as for the adaptive statistical technique.

The basic idea of the combined technique is to use a list of words to match with the input clause from left-to-right. A new segmentation and matching position is found at the end of the longest matched word. Since segmentation errors are due to the coverage of the list of words, typically the terms tend to be over-segmented into smaller ones. For example, the term, 中華人民共和國 (People's Republic of China), is over segmented in the following clause: /根據/中/華/人民/共/和/國/憲法/第/三十一/條/的/規定/，/.

To reduce the amount of over-segmentation, adjacent single-character words are grouped using the bigram technique after maximal-matching. These characters are combined if their mutual information (MI) or co-occurrence frequencies (CF) are greater than a threshold. MI and CF are estimated from the thematic rather than a general corpus because the estimated values should be biased to the thematic corpus. The CF is estimated as the frequency of occurrence of character A immediately before B (i.e. f(A,B)) whereas MI is estimated as log (p(A,B) / (p(A) * p(B)) where p(A,B) (= f(A,B)/N where N is the sum of all CF's) is the estimated probability of character A occurring immediately before B, p(A) and p(B) are the estimated probabilities of character A and B in the corpus, respectively. The threshold is defined by the top N% of the MI or CF distributions. Typically, the MI distribution appears symmetrical and unimodal (figure 1) but the CF distribution decreases with increasing CF and log CF. Although the bigram technique can be modified to identify words of arbitrary length, many non-words are detected. By combining with maximal-matching, the bigram technique can only operate in certain parts of the thematic corpus, reducing the number of non-words detected.

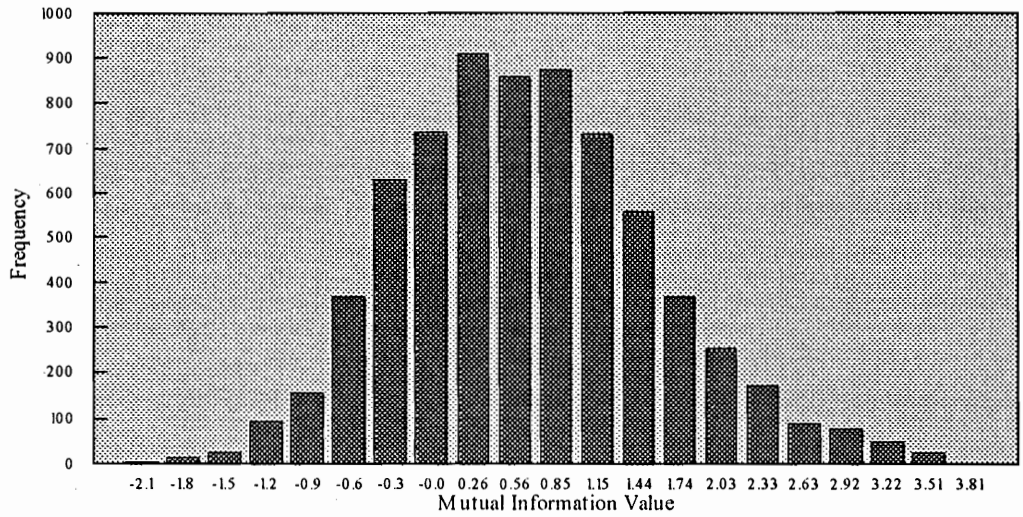

Figure 1: Frequency distribution of MI estimated from the Hong Kong Basic Law. Note that the distribution appears like a normal curve as found in large corpora and the mean position is slightly larger than zero (i.e. two characters are independent).

In the rest of this paper, we discuss how the segmentation programs are evaluated. Next, we compare the performance of maximal-matching between words extracted from a general corpus and those extracted from the manually-segmented text. We show that the bigram technique is equivalent to the nearest-neighbor (NN) clustering. We report the effect of adjusting the threshold (or percentage quartile) on its segmentation and word-identification performances. Finally, we report the result of the combined technique.

II. EVALUATION

We chose the Hong Kong Basic Law [13] as our test data because it

(a) contains a fair amount of proper names, technical terms and compound nouns (see figure 2);
(b) is a typical technical text (in law) which is suitable for machine translation and corpus analysis for humanities and law research;
(c) is large enough for estimating the mutual information in data-exploration since the

distribution of mutual information appear like a normal curve as found in large corpora (figure 1);

(d) is manually segmented and accessible on the computer [14].

第十一條
根據中華人民共和國憲法第三十一條，
香港特別行政區的制度和政策，
包括社會、經濟制度，
有關保障居民的基本權利和自由的制度，
行政管理、立法和司法方面的制度，
以及有關政策，
均以本法的規定爲依據。
香港特別行政區立法機關制定的任何法律，
均不得同本法相抵觸。

Figure 2: An extract of the Hong Kong Basic Law. Terms can already be found, such as the Hong Kong Special Administrative Zone, People's Republic of China and Economic System.

The Basic Law (Figure 2) is different from a general corpus, such as the PH corpus [4] of general news articles. Only 1059 different words appeared in both the Basic Law (2,028 different words) and the PH corpus (42,613 different words), representing 52% and 2.5% overlap, respectively. Figure 2 shows the variation of percentages of word overlap in the Basic Law with the length of the word (i.e. the number of characters). Single-character and two-character words have relatively high percentages of overlap compared with longer words because technical terms and compound nouns tend to be long (> 2 characters), particularly with terms that originated or translated from foreign languages like English, (e.g. aspirin as 阿司匹靈 or tort as 民事過失).

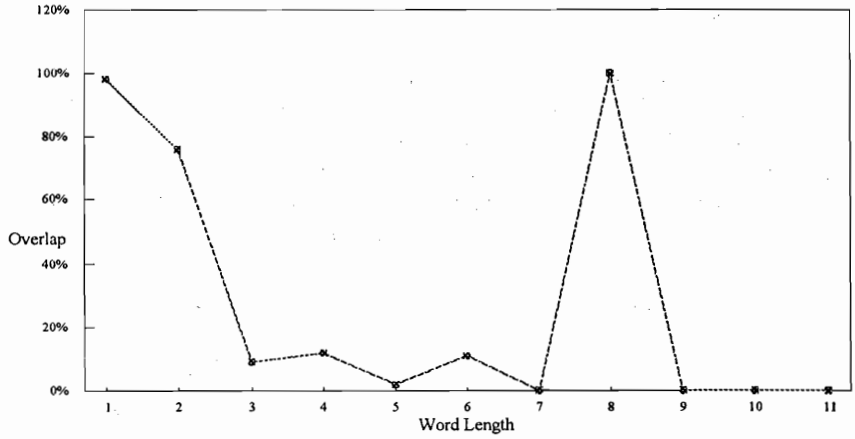

Figure 2: Percentages of words that occur in both the Hong Kong Basic Law and the PH corpus. Note that the percentages of the 8-character words are not accurate because there is only a single 8-character word in the Hong Kong Basic Law.

The Basic Law was segmented by a graduate student in applied linguistics. She was told to segment words as long as possible. For example, the compound noun, 香港特別行政區 (Hong Kong Special Administrative Zone), is considered as a single word rather than its constituents, /香港/特別/行政/區/ (/Hong Kong/ Special/ Administrative/ Zone/). Segmentation markers are placed before and after 的 (de) whenever it is used as an adjective suffix, for example /中國/的/領土/ (/China/ de/ Territories/), since 的 is very productive, creating many new words if it is combined with its adjective root. Unlike [15], numbers are grouped together as a unit, for example /一九八四/年/十二/月/十九/日/ (/1984/Year/December/Month/Nineteenth/Day/).

The segmentation performance is measured by comparing the automatically and manually segmented clauses of the same text. Given that clause i has N_i characters, the maximum amount of segmentation errors is N_{i-1}. Thus, the normalized segmentation error for every clause is $NE_i = E_i/N_i$ where E_i is the amount of segmentation errors. The mean segmentation error is defined as follows where there are k clauses: $E = (\Sigma_i NE_i) / k$.

A single state transducer is used to determine E_i. Initially, E_i is set to zero and the transducer begins at the left-most position on both clauses. The transducer has only two types of actions depending on whether the input symbols are segmentation markers or not. If one clause has segmentation marker but the other does not, then the transducer increments E_i by one and it advances beyond the position of the segmentation marker but remains at the same position on the other clause. If the transducer encounters either segmentation markers or none on both clauses, then there are no segmentation errors and it moves to the following positions of both clauses.

The amount of over-segmentation can be measured simply as the number of automatic segmentation markers, $N_{a,i}$, minus the number of manual segmentation markers, $N_{m,i}$, in clause i, excluding the markers at the beginning and the end of the clause. The over-segmentation is normalized to $N_{m,i}$ because that is the desired number of segmentation. The mean over-segmentation is defined as: $O = [\Sigma_i (N_{a,i}-N_{m,i})/N_{m,i}] / k$.

If O is positive, then there are over-segmentations and vice versa. If O is zero, then the amount of manual segmentation is approximately the same as the automatic. However, we need to examine E to determine whether these segmentations are accurate.

Apart from segmentation, the bigram technique identifies new words. The identification performance (C/B) can be measured as the percentage of identified words, $W_{b,i}$, in the list of different words, W_b, extracted from the Basic Law. However, this measure does not indicate whether the bigram is detecting words that are not in the Basic Law. Thus, another percentage (C/T) defines the ratio between the identified words that are in the Basic Law and the number of identified words.

III. MAXIMAL-MATCHING TECHNIQUE

Maximal-matching depends on the coverage of the dictionary. We compared the list of words extracted from the Basic Law and its subset that also occurred in the PH corpus. The first list achieve a low segmentation error of 1.2% compared with 15% using the second list. Consequently, the clause accuracy for the first list (82%) is much better than the second (28%). This is quite surprising since the words extracted from the PH corpus are derived from a general Chinese dictionary [16] of about 56,000 words. The first list under-segments the Basic Law (i.e. 3%) where as the other over-segments (6%).

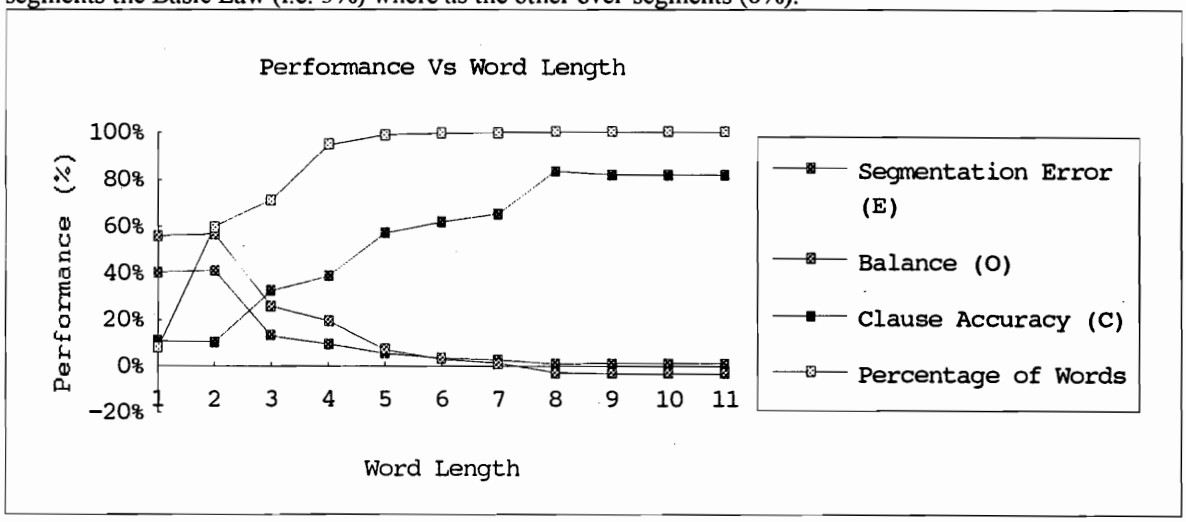

Figure 3: Variation of segmentation performance with the different words extracted from the Hong Kong Basic Law, up to certain lengths. Key: E represent the segmentation error, O measures the over-segmentations, C is the clause accuracy and W is the amount of words up to a particular length.

Since the maximal-matching uses the longest matched word, we were interested in the variation of segmentation performance with the list of words up to a particular length. The segmentation error reduces as longer words are included (Figure 3). However, the addition of 2-character words do not in general improve the segmentation performances of the single-character words. Longer words have to be identified for

significant improvement. The amount of over-segmentation reduces near to zero when words of length up to 7 characters are used in segmentation. Longer words will moderately make maximal-matching under-segments but improving the clause accuracy significantly.

IV. BIGRAM TECHNIQUE

Previous work [7,8] grouped two adjacent characters as a two-character word. However, we showed that detecting only two-character words hardly improve segmentation performance in the last section. Thus, we extended the idea to detect words of arbitrary length by grouping any two adjacent characters in the text if their MI or CF is greater than a threshold. This is equivalent to NN clustering [17] which defines a distance matrix between characters in a clause. Distances between two non-adjacent characters are infinite because overlap grouping is avoided. Thus, it is sufficient to know the distance between adjacent characters. Distances of the same character at the same position must have zero distances. A pre-defined threshold can cut the dendrogram into a sequence of subtrees which represent detected words of the clause.

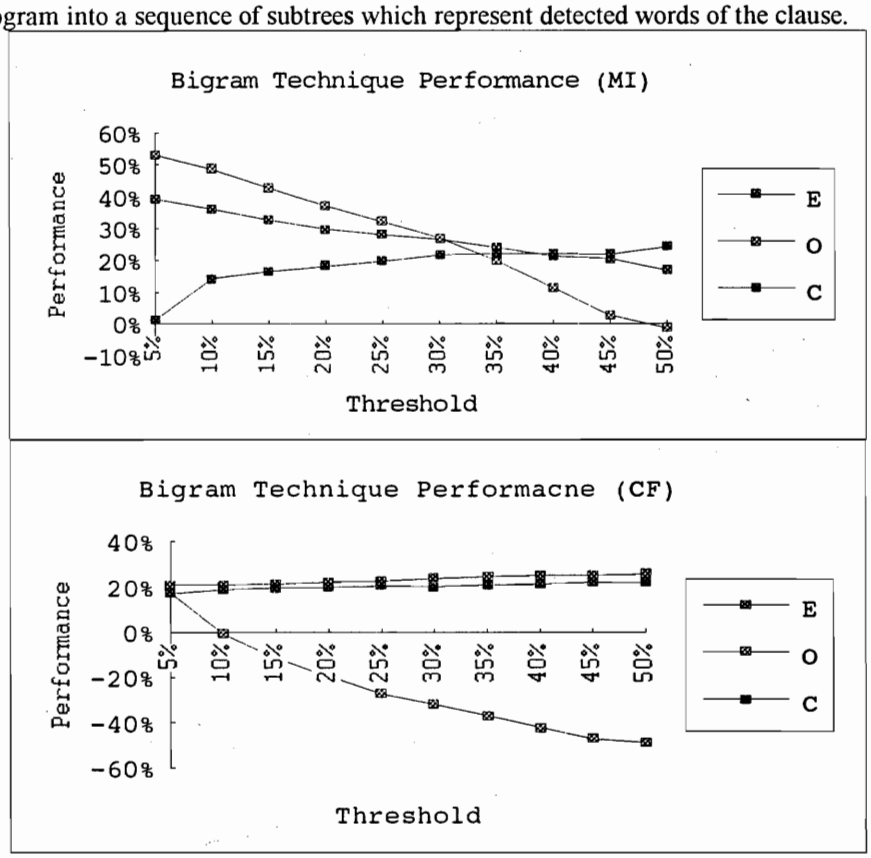

Figure 5: Variation of segmentation performance between distances defined as the MI (a) and CF (b), with respect to different percentage quartiles. Key: E, O and C are the segmentation error, over-segmentation and clause-accuracy, respectively. The suffix m and b indicate the performance measure of MI and the CF, respectively.

The MI is used to define the distance between two adjacent characters and it is estimated from a thematic corpus and not from the a general corpus. The threshold is defined in terms of the top N% quartile since it becomes the N% significance level if the distribution is normal. Apart from MI, the distance can be defined in terms of the CF. Figure 5 shows variation of the segmentation performance between distances defined in terms of the MI and the CF, with respect to different percentage quartiles.

Using MI, the segmentation error reduces steadily from 40% to 17%, just over 50% error reduction. The amount of over-segmentation is about zero when the percentage quartile is about 48% and the clause-accuracy rose from almost zero to about 27%. Note that the accuracy rose dramatically between 5% and 10%. Using CF, the segmentation error and the clause accuracy do not vary dramatically with the percentage quartile. The amount of segmentation error is about the same as the one using MI and the clause-accuracy is only 4% lower than MI. The amount of over-segmentation quickly becomes under-segmentation when the percentage quartile reaches beyond 10%. In summary, the oc-occurrence frequencies are more robust to the

variation of the percentage quartile than using MI but at a cost of lower clause-accuracy and segmentation error.

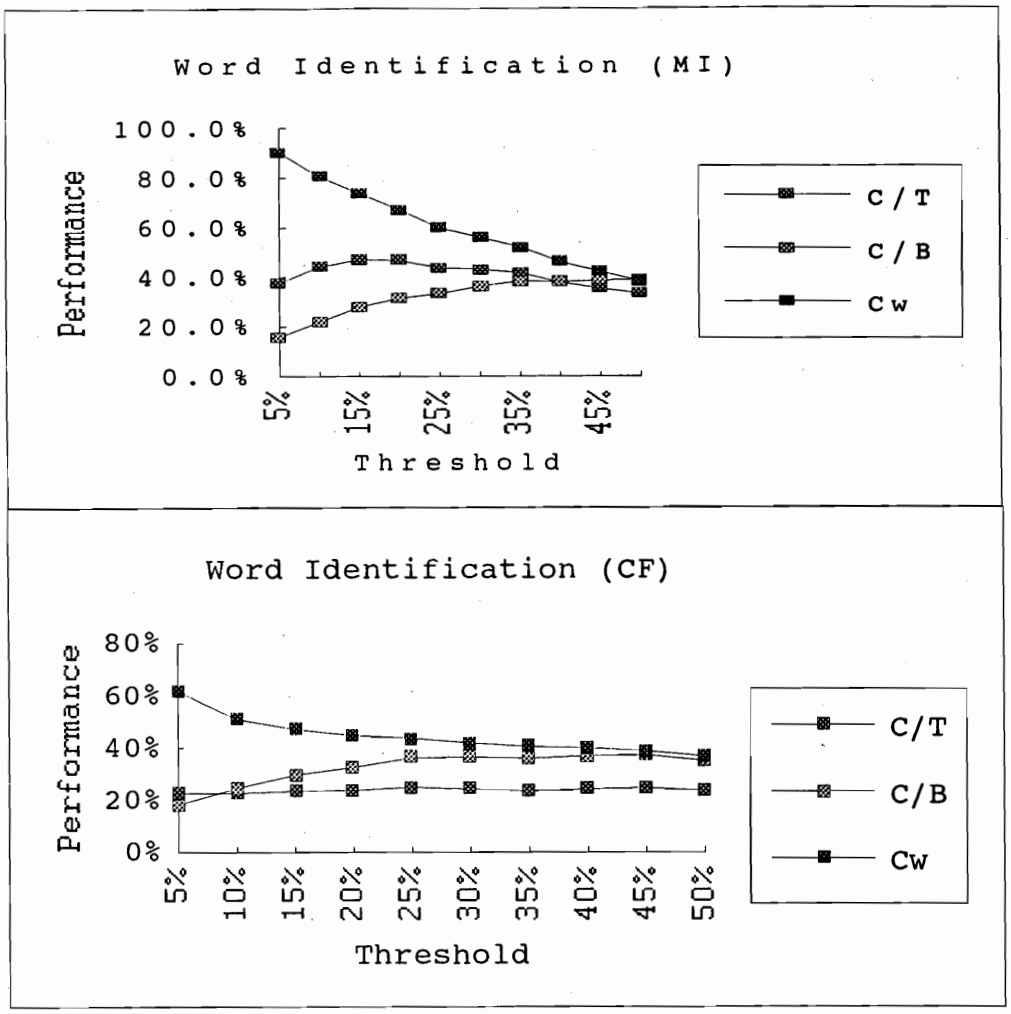

Figure 6: Word-identification performances with different values of percentage quartile for MI (a) and CF (b). Suffixes m and b indicate that NN clustering is carried out using MI and CF, respectively. Key: C/T is the percentage of words in the Basic Law that are identified, C/B is the percentage of identified words in the Basic Law and Cw is the percentage of identified words that are in either the Basic Law or the PH corpus.

Figure 6 shows the word-identification performance with respect to different percentage quartiles. Using MI, the percentages of words in the Basic Law that are identified (i.e. C/T measure) rose to a maximum when the percentage quartile is about 15% where almost 50% of the identified words are words in the Basic Law. A decrease in percentages as the percentage quartile increases imply that there are more words identified but less of them are in the Basic Law. The C/B measures the percentage of identified words that are in the Basic Law. This measure increases steadily but slowly flattened. The Cw measures the percentages of identified words in either the Basic Law or the PH corpus, indicating that the identified words are recognized Chinese words. Here, Cw decreases when the percentage quartile increases, indicating more non-recognized Chinese words are identified.

Using CF, the percentage of words in the Basic Law that are identified do not vary significantly with the percentage quartile. The C/B and Cw measure increase and decrease, respectively, where both asymptotically tend to 36%. Non recognized words (30 characters) are usually longer than those identified using MI (11 characters at the maximum). The word-length distribution of the identified words using bigram frequencies are skewed where as the distribution using MI appears like the distribution of words in the Basic Law. In summary, the CF is more robust than the MI in word-identification but the former can yield almost 100% more correct word-identification than the latter.

The bigram technique achieved similar segmentation performances (i.e. E = 15-17% and C = 28%) to the maximal-matching using words from a general dictionary [16]. The maximal-matching tends to over-segment but the bigram technique can potentially under-segment, depending on the percentage quartile.

V COMBINED TECHNIQUE

The combined technique applies maximal-matching to the text and then using the bigram technique to group single-character words. The identified words are combined with the existing list of words which are used by the maximal-matching to segment the given text again. Figure 7 shows the segmentation performance of the combined technique using MI or CF.

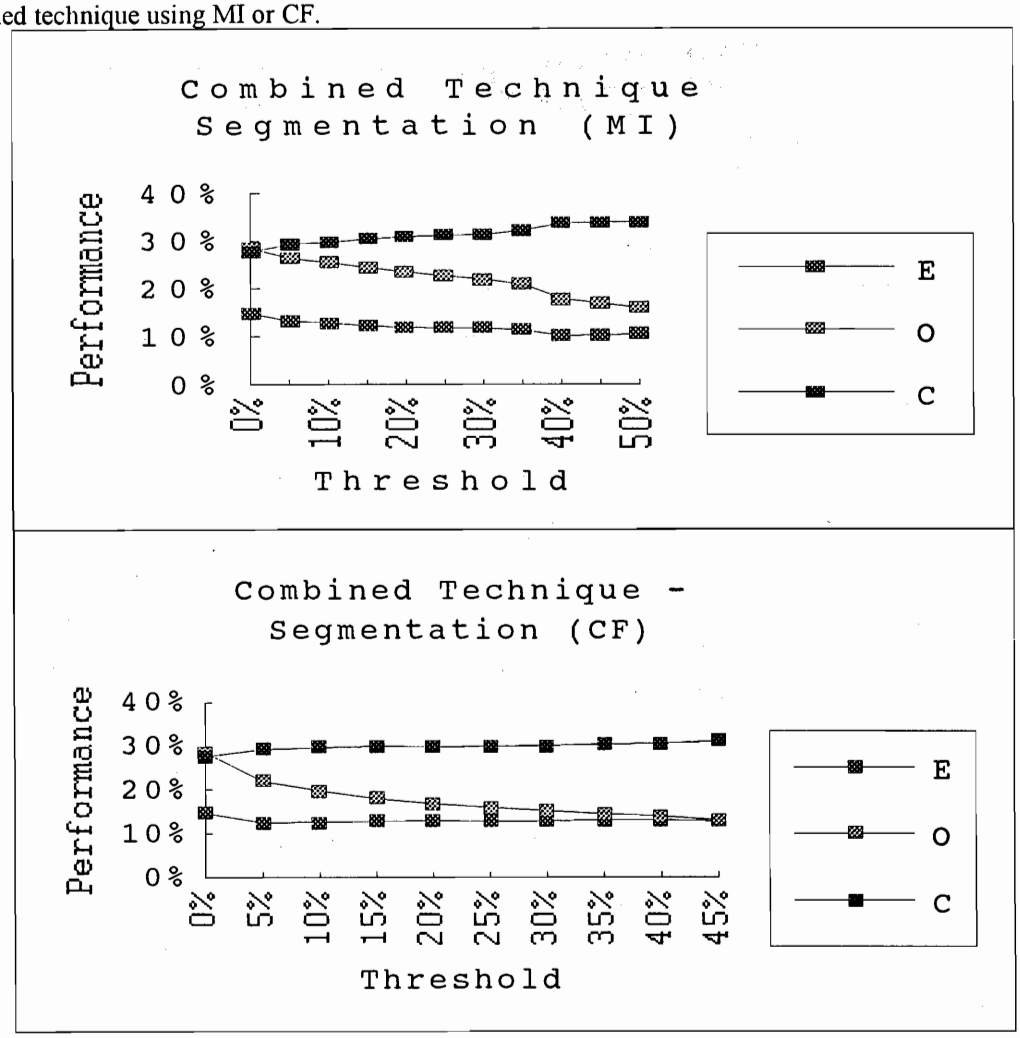

Figure 7: Segmentation performance of the combined technique using. Key: E, O and C are the segmentation error, over-segmentation and clause accuracy, respectively. The following letter "m" and "b" indicate that MI and CF are used, respectively.

Using MI (figure 7), the combined technique can reduce the segmentation error from 15% to 10% (i.e. error reduction of 33%) when the percentage quartile is at 40%. The amount of over-segmentation is reduced by 12% (from 28% to 16%) and the clause accuracy is increased by 5% (i.e. 18% improvement). Using CF, the segmentation error is better than using the MI, only when the percentage quartile is 5%. Otherwise, the segmentation error varies little with the percentage quartile. Reduction of over-segmentation is larger than that using MI but the clause accuracy is not as high as that using MI. In summary, the segmentation performance using MI is better than that using CF.

The amount of correct word identifications are all higher than 50% because the maximal-matching technique uses a dictionary of 52% overlap with the Basic Law (Figure 8). Although the difference between percentages of identified words in the Basic Law is small (C/Tm versus C/Tb) between MI and CF, the other two measures have pronounced difference. In both measures, the MI achieves better word identifications than CF where we expect the percentage of correct identification chosen from the identified words is 80% for

MI, almost independent from the percentage quartile. In addition, the percentage of identified words that are recognized Chinese words remain above 85%, decreasing with increasing values of the percentage quartile. A dramatic increase in word-identification occur in the first 5% quartile and subsequent variation in performance vary less than the first 5%.

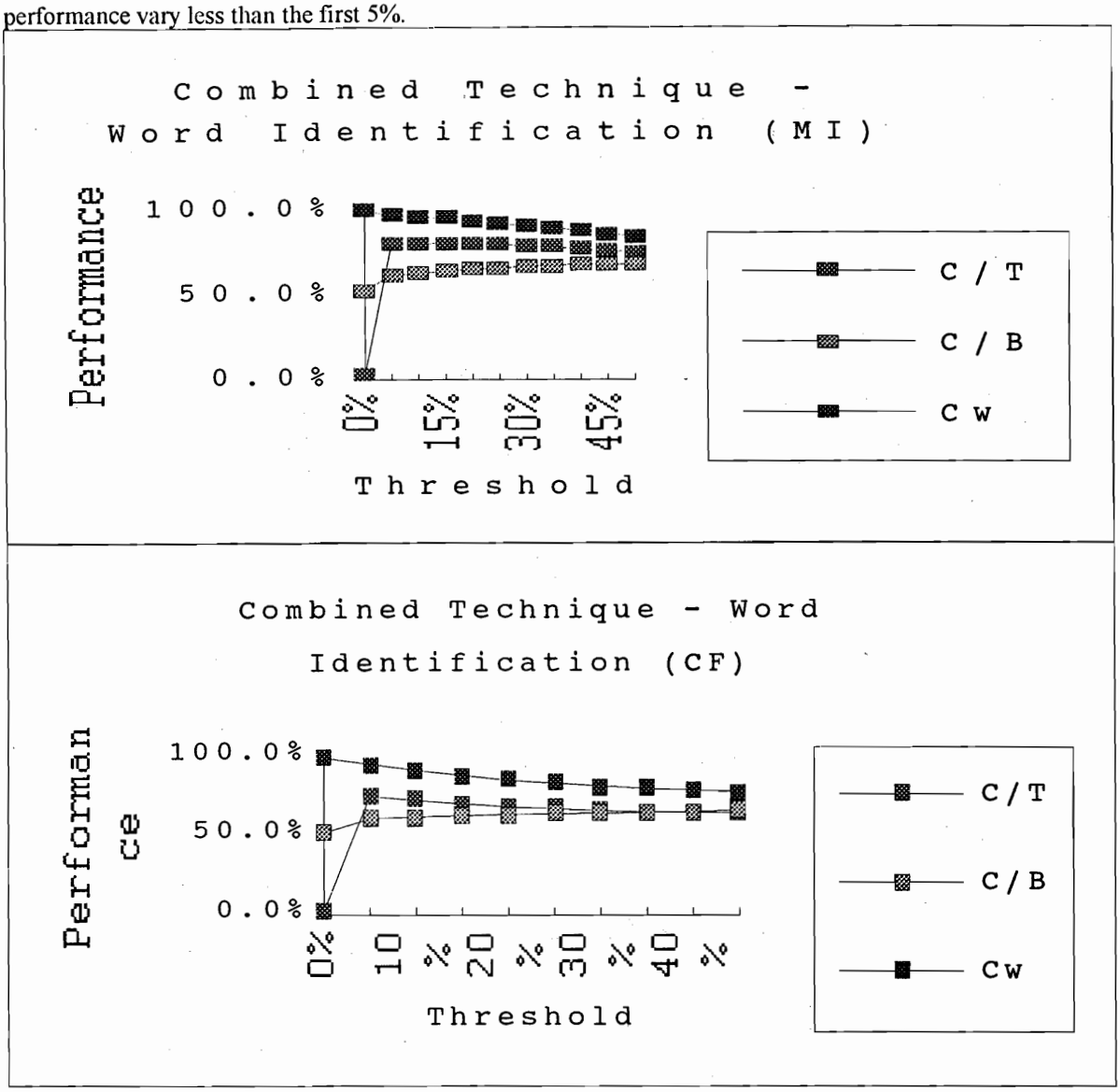

Figure 8: Word-identification performances with respect to different values of the percentage quartile. Key: C/T and C/B represent the percentage of correct identification with respect to the number of identified words and words in the Basic Law, respectively. Cw represents the percentage of identified words that are either in the Basic Law or the PH corpus. The following letter "m" and "b" denote using MI and CF in word segmentation and identification, respectively.

Since the initial list of words used in maximal-matching has 52% overlap with the Basic Law, the measure C/B is not representative. We re-calculated the percentages according to the following normalization formula: dC/T = (C/T - 52%) / (100% -52%). Figure 9 shows that the MI can detect 33% of the remaining words in the Basic Law that are not in the initial word list of maximal-matching. The performance of MI is consistently better than that achieved using CF. When the percentage quartile is 0%, it is equivalent to using only the maximal-matching technique (i.e. no word identification). Again, the first 5% yields a dramatic increase in performance and there is little difference between using MI and CF.

A list of words or short phrases detected by the combined technique is in the appendix where the mutual information is used and the threshold is set at the top 20%. There are no words of length greater than 6. Words of length greater than 4 are few and usually not recognized as words because of the attached verbs (e.g. 屬於). Only 2 out of 17 words of length 4 are not recognized words or phrases. There are more three-

character non-words because of attaching particles of verbs (e.g. 療) or conjunctions (e.g. 及). Function characters at the end of words are not considered to be unrecognized words because they can be detected and rectified. Detection of two character words are more reliable than three-character ones.

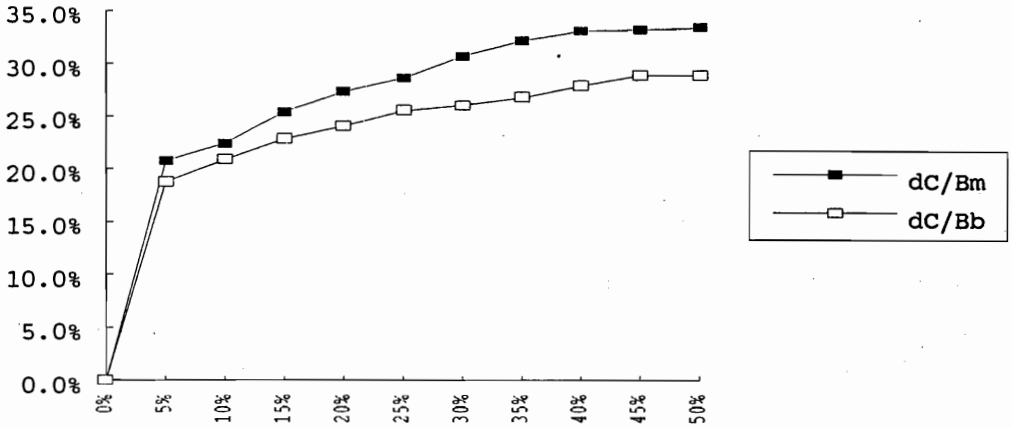

Figure 9: Normalized percentage of identified words that are in the Basic Law but not in the initial word list of maximal-matching. Key: the following letter "m" and "b" denote word segmentation and identification using MI and CF, respectively.

VI. CONCLUDING REMARKS

Our implementation of maximal-matching achieved a mean segmentation error (1.2%) as low as those (1.14%) reported in [11]. However, in practice, when text processing (e.g. machine translation) is applied in a specific domain (i.e. adopt a sub-language approach), the segmentation performance is degraded, as our test data demonstrated (15% segmentation error using words from a general dictionary). The bigram technique which achieves good 2-character word identification offers little assistance to maximal matching as we showed that increasing the amount of 2-character words did not improve the segmentation significantly. We extended the bigram technique to identify words of arbitrary length and the segmentation performance was about the same as maximal matching using a general dictionary. By combining both techniques, we were able to lower the segmentation error by 33% of its degraded performance and improve the word-identification by 33% of the remaining words only in the Basic Law, depending on the percentage quartile (or threshold). The MI appear to yield better segmentation and word-identification performance than CF. However, there is little difference between the two at 5% quartile where the improvement in performance is most dramatic.

REFERENCES

[1] McNaught, J. (1993) "User needs for textual corpora in NLP", *Literary and Linguistic Computing*, **8**, n4, pp.227--234.

[2] Zhang, J-S., S. Chen, Y. Zheng, X-Z. Liu and S-J. Ke (1992) "Automatic recognition of Chinese full name depending on multiple corpus", *Journal of Chinese Information Processing*, **6**, n3, pp. 7-15.

[3] Chang, J.S., C.D. Chen and S.D. Chang (1991) "Chinese word segmentation through constraint satisfaction and statistical optimization", *Proceedings of ROC Computational Linguistics Conference*, Kenting, Taiwan (in Chinese), pp. 147-166.

[4] Guo, J. and H.C. Lam (1992) "PH: a Chinese corpus for pinyin-hanzi transcription", *Technical Report TR93-112-0*, Institute of Systems Sciences, National University of Singapore.

[5] Smadja, F. (1993) "Retrieving collocations from text: Xtract", *Computational Linguistics*, **19**, n1, pp.143--177.

[6] Salton, G. (1989) *Automatic Text Processing*, Addison-Wesley: Reading, Mass.

[7] Sproat, R. and C.L. Shih (1990) "A statistical method for finding word boundaries in Chinese text", *Computer Processing of Chinese and Oriental Languages*, **4**, n4, pp. 336-351.

[8] SU, M.S. (1993) *Private Communication*, Tsinghua University, Beijing, People's Republic of China.

[9] KIT, C., Y. LIU AND N. LIANG (1989) "On methods of Chinese automatic word segmentation", *Journal of Chinese Information Processing*, **3**, n1, pp. 13-20.

[10] FAN, C.K. AND W.H. TSAI (1988) "Automatic word identification in Chinese sentences by the relaxation technique", *Computer Processing of Chinese and Oriental Languages*, **4**, n1, pp. 33-56.

[11] CHIANG, T.H., J.S. CHANG, M.Y. LIM AND K.Y. SU (1993) "Statistical models for word segmentation and unknown word resolution", *Proceedings in ROCLING V '93*, pp. 123-146.

[12] ZU, B.Z., J. ZHANG AND Q.H. HE (1992) "The method of Chinese word segmentation based on neural network", *Journal of Chinese Information Processing*, **7**, n2, pp. 36-44.

[13] PRC GOVERNEMENT PUBLICATION, *Hong Kong Basic Law.*

[14] LUN, C.S. AND L. LING (1993) *Hong Kong Basic Law (in Big-5 Code)*, Department of Chinese, Translation and Linguistics, City Polytechnic of Hong Kong.

[15] NATIONAL STANDARD BUREAU (1988) *Contemporary Chinese language words segmentation standard used for information processing.*

[16] FU, X-L. (1987) *Xiandiao Hanyu Tunrun Cidian*, Waiyu Jiaoxue Yu Yanjiu Publishing House: Beijing, PRC.

[17] EVERITT, B. (1985) *Cluster Analysis*, Heinemann: London.

APPENDIX

The following is the list of words detected by the combined technique from the Hong Kong Basic Law. The threshold is set at the top 20% of all bigrams measured by mutual information. If the last character of an entry is a slash character (i.e. "/"), then the entry is a plausible word or short phrase. Due to space, 2-character words detected are not included here.

應課差餉租值/	不少於/	這幾個/
自然資源屬於	紫荊花/	白兩色
衡又互相配合	不低於/	刑事罪/
違反誓言而	療衛生	獲功能/
各類院校均	審計署/	學語言/
集裝箱碼頭/	及竊取	來源證/
過半數票即	準備金/	境衛生
專題小組/	約束力/	航空器/
出口配額/	屈功能	新興產/
收支平衡/	丁屋地/	將採取/
過半數票/	盡忠職	也先後/
醫療衛生/	彈劾案/	範圍及
開支標準/	星花蕊	登記冊/
廉潔奉公/	範圍內/	記錄在/
日恢復對	或瀆職	明創造
技術停降/	證明書/	標準向
鄉村屋地/	代擬稿/	興旺發
軍用船隻/	既互相	均假定/
自然資源/	原舊批	花蕊上/
顆象徵著	過半數/	西醫藥/
刑事罪犯/	姬鵬飛/	龍半島
登記標誌/	構想及	
外圈寫有/	元匯價/	
救助災害/	製造業/	
航空公	學歷等/	
預算案/	被判犯/	