Automatic Summarization of Chinese and English Parallel Documents

Fu Lee Wang* and Christopher C. Yang†

*: Department of Computer Science
The City University of Hong Kong
†:Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
yang@se.cuhk.edu.hk

Abstract. *As a result of the rapid growth in Internet access, significantly more information has become available online in real time. However, there is not sufficient time for users to read large volumes of information and make decisions accordingly. The problem of information-overloading can be resolved through the application of automatic summarization. Many summarization systems for documents in different languages have been implemented. However, the performance of summarization system on documents in different languages has not yet been investigated. In this paper, we compare the result of fractal summarization technique on parallel documents in Chinese and English. The grammatical and lexical differences between Chinese and English have significant effect on the summarization processes. Their impact on the performances of the summarization for the Chinese and English parallel documents is compared.*

1. Introduction

As the information available on the World Wide Web is growing exponentially, the information-overloading problem has become a significant problem. Such problem can be reduced by text summarization, but it is time consuming for human professional to conduct the summarization. Due to the huge volume of information available on line in real time, the research of automatic text summarization becomes very critical.

The information available in languages other than English on the World Wide Web is increasing significantly. In the recognition of the need for summarization systems for languages other than English, summarization systems developed for other languages, such as Korean [21], Japanese [14], and Chinese [3], etc., has been developed recently. Most of these summarization systems are monolingual system, i.e., they can process documents in one single language only. There are some multilingual summarization systems, i.e., they are capable to process document in multiple languages [4, 23]. However, the multilingual documents used for these summarization systems are not in parallel; therefore, the experimental results of these multilingual summarization systems do not reflect the impact of the languages on the results of applying the summarization techniques.

† Corresponding Author: Christopher C. Yang (yang@se.cuhk.edu.hk)

In this paper, we investigate the fractal summarization technique that is proposed based on the fractal theory [26, 27]. In fractal summarization, the important information is captured from the source text by exploring the hierarchical structure and salient features of the document. A condensed version of the document that is informatively close to the original is produced iteratively using the contractive transformation in the fractal theory. User evaluation has been conducted and shown that fractal summarization outperforms the traditional summarization without exploring the hierarchical structure of the documents. The fractal summarization technique is developed based on the statistical approach and can be applied to any languages. In this work, we apply the fractal summarization technique on a parallel corpus in English and Chinese. The summarization results in English and Chinese are compared directly.

The rest of this paper will be organized as following. Section 2 reviews the techniques in automatic text summarization. Section 3 presents the fractal summarization technique. Section 4 analyzes the difference of Chinese and English parallel document. Section 5 compares the results of Chinese and English summarization. Section 6 provides the concluding remarks and suggests some future research directions.

2. Automatic Summarization

Traditional automatic text summarization is the selection of sentences from the source document based on their significance to the document [5, 18] without considering the hierarchical structure of the document. The selection of sentences is conducted based on the salient features of the document. The thematic, location, heading, and cue phrase features are the most widely used summarization features.

- The *thematic feature* is first identified by Luhn [18]. Edmundson proposed to assign the thematic weight to keyword based on term frequency, and the sentence thematic weight as the sum of thematic weight of constituent keywords [5]. The *tfidf* (Term Frequency, Inverse Document Frequency) score is most widely used to calculate the thematic weight [24].
- The significance of sentence is indicated by its *location* [2] based on the hypotheses that topic sentences tend to occur at the beginning or in the end of documents or paragraphs [5]. Edmondson proposed to assign positive weights to sentences according to their ordinal position in the document.
- The *heading feature* is proposed based on the hypothesis that the author conceives the heading as circumscribing the subject matter of the document [5]. A heading glossary is a list consisting of all the words in headings and subheadings with positive weights. The heading weight of sentence is calculated by the sum of heading weight of its constituent words.
- The *cue phrase feature* is proposed by Edmundson [5] based on the hypothesis that the probable relevance of a sentence is affected by the presence of pragmatic words. A pre-stored cue dictionary with cue weight is used to identify the cue phrases. The cue weight of sentence is calculated by the sum of cue weight of its constituent words

Typical summarization systems select a combination of summarization features [5, 16, 17], the total sentence weight ($W_{sentence}$) is calculated as weighted sum of the weights computed by each sailent features,

$$W_{sentence}=a_1\times w_{thematic}+a_2\times w_{location}+a_3\times w_{heading}+a_4\times w_{cue}$$

where $w_{thematic}$, $w_{location}$,. $w_{heading}$ and w_{cue} are thematic weight, location weight, heading weight and cue weight of the sentence respectively; and a_1, a_2, a_3, and a_4 are positive integers to adjust the weighting of four summarization features. The sentences with sentence weight higher than a threshold are selected as part of the summary. It has been proven that the weighting of different summarization features do not have any substantial effect on the average precision [16]. In our experiment, the maximum weight of each feature is normalized to one, and the total weight of sentence is calculated as the sum of scores of all summarization features without weighting. However, the cue phrase feature is disabled for summarization of parallel document, because there is not any parallel cue phrase dictionary defined for Chinese and English currently. Besides, it does affect the performance of the summarization result by adding the cue phrase feature.

3. Fractal Summarization

Many summarization models have been proposed previously. None of the models are entirely developed based on document structure, and they do not take into account of the fact that the human abstractors extract sentences according to the hierarchical document structure. Document structure can be described as fractals. In the past, fractal theory has been widely applied in the area of digital image compression, which is similar to the text summarization in the sense that they both extract the most important information from the source and reduce the complexity of the source. The fractal summarization model is the first effort to apply fractal theory to document summarization. It generates the summary by a recursive deterministic algorithm based on the iterated representation of a document.

3.1 Fractal Theory & Fractal View for Controlling Information Displayed

Fractals are mathematical objects that have high degree of redundancy [19]. These objects are made of transformed copies of themselves or part of themselves. Mandelbrot was the first person who investigated the fractal geometry and developed the fractal theory [19]. In his well known example, the length of the British coastline depends on measurement scale. The larger the scale is, the smaller value of the length of the coastline is and the higher the abstraction level is. The British coastline includes bays and peninsulas. Bays include sub-bays and peninsulas include sub-peninsulas. Using fractals to represent these structures, abstraction of the British coastline can be generated with different abstraction degrees.

A physical tree is one of classical example of fractal objects. A tree is made of a lot of sub-trees; each of them is also tree. By changing the scale, the different levels of abstraction views are obtained (Fig. 1). The idea of fractal tree can be extended to any logical tree. The degree of importance of each node is represented by its fractal

value. The fractal value of focus is set to 1. Regarding the focus as a new root, we propagate the fractal value to other nodes with the following expression:

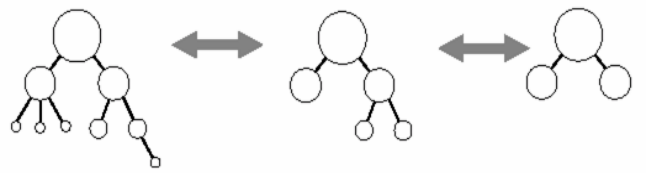

Fig. 1. Fractal View for Logical Tree at Different Abstraction Level.

$$Fv_{root} = 1$$
$$Fv_{child\ node\ of\ x} = C\ Fv_x / N_x^{1/D}$$

where Fv_x is the fractal value of node x; C is a constant between 0 and 1 to control rate of decade; N_x is the number of child-nodes of node x; and D is the fractal dimension.

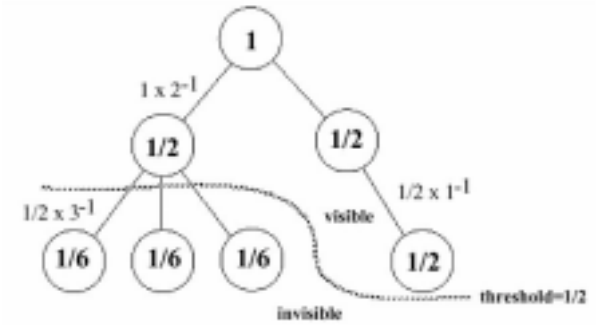

Fig. 2. An Example of the Propagation of Fractal Values.

Fractal view is a fractal-based method for controlling information displayed [15]. Fractal view provides an approximation mechanism for the observer to adjust the abstraction level and therefore control the amount of information displayed. At a lower abstraction level, more details of the fractal object can be viewed. A threshold value is chosen to control the amount of information displayed, the nodes with a fractal value less than the threshold value will be hidden (Fig. 2). By changing the threshold value, the user can adjust the amount of information displayed.

3.2 Fractal Summarization

Many studies of human abstraction process has shown that the human abstractors extract the topic sentences according to the document structure from the top level to the low level until they have extracted sufficient information [6, 11]. Advance summarization techniques take the document structure into consideration to compute the probability of a sentence to be included in the summary. However, most traditional automatic summarization models consider the source document as a

sequence of sentences but ignoring the structure of document. *Fractal Summarization Model* is proposed to generate summary based on document structure [26].

Fractal summarization is developed based on the fractal theory, the important information is captured from the source document by exploring the hierarchical structure and salient features of the document. A condensed version of the document that is informatively close to the original is produced iteratively using the contractive transformation in the fractal theory. Similar to the fractal geometry applying on the British coastline where the coastline includes bays, peninsulas, sub-bays, and sub-peninsulas, large document has a hierarchical structure with several levels, chapters, sections, subsections, paragraphs, sentences, terms, words and characters. A document can be represented by a hierarchical structure (Fig. 3). However, a document is not a true mathematical fractal object since a document cannot be viewed in an infinite abstraction level. The smallest unit in a document is character; however, neither a character nor a word will convey any meaningful information concerning the overall content of a document. The lowest abstraction level in our consideration is a term. A document is considered as *prefractal* that are fractal structures in their early stage with finite recursion only [7].

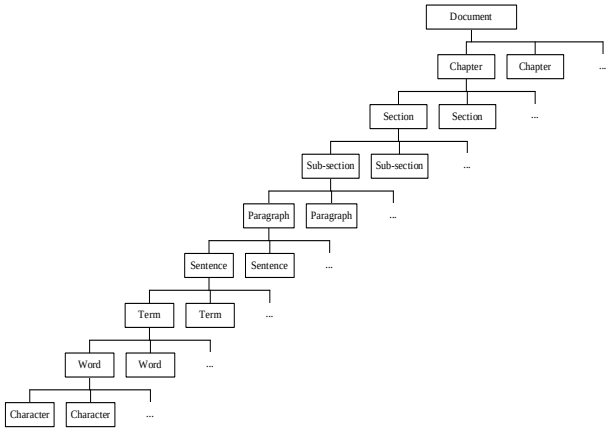

Fig. 3. Prefractal Structure of Document.

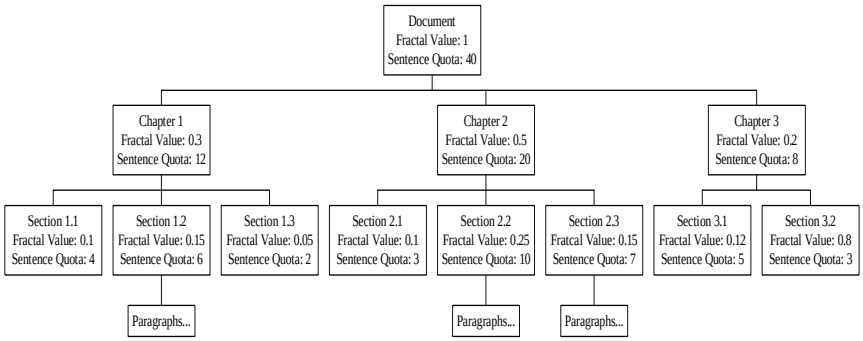

Fig. 4. An Example of Fractal Summarization Model.

The Fractal Summarization Model applies a similar technique as fractal view and fractal image compression [1]. An image is regularly segmented into sets of non-overlapping square blocks, called range blocks, and then each range block is subdivided into sub range blocks, until a contractive mapping can be found to represent this sub range block. The Fractal Summarization Model generates the summary by a simple recursive deterministic algorithm based on the iterated representation of a document. The original document is represented as fractal tree structure according to its document structure. The weights of sentences under a range block are calculated by the traditional summarization methods described in Section 2. The fractal value of root node is 1 and the fractal values of the child node are propagated according to the sum of sentence weight under the child nodes.

$$Fv_{root} = 1$$

$$Fv_{child\ node\ r\ of\ x} = Fv_x \times \frac{\sum_{sentences\ under\ r} Sentence\ weight}{\sum_{sentences\ under\ x} Sentence\ weight}$$

Given a document, users provide compression ratio to specific the amount of information displayed. The *compression ratio* of summarization is defined as the ratio of number of sentences in the summary to the number of sentences in the source document. The summarization system computes the number of sentences to be extracted as summary accordingly and the system assigns the number of sentences to the root as the quota of sentences. The quota of sentences is allocated to child-nodes by propagation, i.e., the quota of parent node is shared by its child-nodes directly proportional to the fractal value of the child-nodes. The quota is then iteratively allocated to child-nodes of child-nodes until the quota allocated is less than a threshold value and the range-block can be transformed to some key sentences by traditional summarization methods (Fig. 4). A threshold value is the maximum number of sentences can be extracted from a range block. If the quota is larger than the threshold value, the range block must be divided into sub-range block.

Fig. 4 demonstrates an example of fractal summarization model. The fractal value of the root is 1, and the system extract 40 sentences from the root node. The system then allocates the sentence quota to the child nodes directly proportion to the fractal value of child node. The fractal value and sentence quota will be prorogated to the grandchild nodes. For example, the Section 1.2 receives quota of 6 sentences which is higher than threshold value, therefore the system will extend the node in paragraph levels. However, the Section 1.1 and 1.3 receive a quota less than 5 sentences, therefore the system directly extract sentence at section level. The detail of the Fractal Summarization Model is shown as the following algorithm:

Fractal Summarization Algorithm

1. Choose a Compression Ratio.
2. Choose a Threshold Value.
3. Calculate the Sentence Number Quota of the summary.
4. Divide the document into range blocks.
5. Transform the document into fractal tree.
6. Set the current node to the root of the fractal tree.
7. Repeat

```
7.1       For each child node under current node,
              Calculate the fractal value of child node.
          7.2     Allocate Quota to child nodes in proportion
          to fractal values.
7.3       For each child nodes,
              If the quota is less than threshold value
                  Select the sentences in the range block by
                  extraction
              Else
                  Set the current node to the child node
                  Repeat Step 7.1, 7.2, 7.3
8.  Until all the child nodes under current node are
    processed
```

3.3 Experimental Result

It is believed that a full-length text document contains a set of subtopics [12] and a good quality summary should cover as many subtopics as possible. Experiment of fractal summarization and traditional summarization has been conducted on Hong Kong Annual Report 2000 [26], the traditional summarization model without considering the hierarchical structure of the documents extracts most of sentences from few chapters. However, the fractal summarization model extracts the sentences distributively from each chapter. The fractal summarization model produces a summary with a wider coverage of information subtopic than traditional summarization model. A user evaluation has been conducted to compare the performance of the fractal summarization and the traditional summarization without using the hierarchical structure of documents. The results show that all subjects consider the summary generated by fractal summarization method as a better summary. The fractal summarization can achieve up 91.25% precision and 87.125% on average, but the traditional summarization can achieve up to maximum 77.50% precision and 67% on average.

4. Comparison of Chinese and English Parallel Documents

Parallel documents are popular in places with multilingual culture, such as Québec, Hong Kong, Singapore and many other European countries. Hong Kong had its bilingual culture since it was a British colony more than a century ago. The official languages are Chinese and English, therefore a lot of documents are written in Chinese and English using covert translation [26]. For example, most of the documents released by the government have both Chinese and English versions. The documents are written by experienced bilingual linguists, and therefore the quality of the documents can be assured. In this section, we investigate the characteristics in the parallelism of Chinese and English parallel documents.

4.1 Indexing

In informational retrieval and processing, indexing is one of the most important research issues, searching and retrieval of information is impossible without proper indexes. The information content of a document is determined primarily by the frequency of the terms in the document; therefore, the indexing of document is the process to transform a document into a vector of terms with its frequency or other related score. In fact, the process is much complicated since the terms in a document are not properly marked-up. English indexing includes several steps, i.e., lexical analysis, stop-wording, stemming, and index terms selection [8, 13]. The techniques for English indexing are considerably more mature than Chinese indexing. Due to the lack of word delimiters (such as spacing in English), Chinese text segmentation is more difficult. Besides, there are ambiguities in Chinese text segmentation. Different ways of segmenting a Chinese sentence may lead to different meanings [10, 25]. There are three major approaches in Chinese text segmentation: a) statistical approach, b) lexical rule-based approach, and c) hybrid approach based on statistical and lexical information [22].

4.2 Parallelism of Chinese and English Parallel Documents

Parallel Corpus is defined as a set of document pairs that are aligned based on their parallelism. Parallel corpus can be generated by overt translation or covert translation [26]. Due to the grammatical and lexical differences between different languages, words in one language may be translated into one or more words in another language or may not be translated at all. There is probably more than one way to translate a word in one language into another language. However, a pair of parallel documents is always parallel in terms of their information contents.

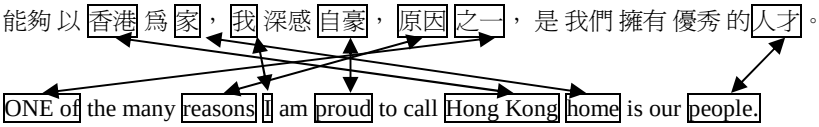

Fig. 5. Reordering of Equivalent Terms in Chinese and English sentences.

The sentence structures for a pair of parallel documents in two languages are different due to the grammatical and lexical differences in languages. For example, as shown in Fig. 5, the orderings of terms in two languages are not the same. The structures of sentence can also be changed. Several sentences in one language can be merged into one sentence in another language. It is also possible to mix the content of several sentences in one language together to form a number of sentences in another language. In order to study the alignment of sentences in two languages, we have analyzed the mapping of sentence in Hong Kong Annual Report 2000. 85% of the sentences in two languages are one-to-one mapping. 7.3% of the sentences are one-to-two mapping between Chinese and English and 6.1% of the sentences are two-to-one mapping between Chinese and English. The one-to-one, two-to-one and one-to-

two sentence mappings totally yield more than 98%. Most sentences contain tow or less sub-sentences in the other language (Table 1).

Table 1. Mapping of Sentences between Hong Kong Annual Report 2000 Chinese Version and English Version.

Number of Mappings (Percentage)		No. of Sentences in HKAR 2000 (English Version)					
		1	**2**	**3**	**4**	**5**	**6**
No. of Sentences in HKAR 2000 (Chinese Version)	**1**	6288 (85.00%)	540 (7.30%)	31 (0.42%)	3 (0.04%)	0	1 (0.01%)
	2	448 (6.06%)	43 (0.58%)	9 (0.12%)	0	0	0
	3	19 (0.26%)	6 (0.08%)	2 (0.03%)	2 (0.03%)	0	0
	4	2 (0.03%)	1 (0.01%)	0	2 (0.03%)	0	0
	5	0	0	0	0	0	0
	6	1 (0.01%)	0	0	0	0	0

Table 2. Statistics of Keyword 'Hong Kong' and '香港' in the Hong Kong Annual Report 2000 and its tfidf score in the first aligned sentence in the Hong Kong Annual Report 2000 .

Text Unit	Term frequency		Text block frequency		No of Text Block		*tfidf* Score	
	English	Chinese	English	Chinese	English	Chinese	English	Chinese
Document-Level	1217	1704	1	1	1	1	1217.00	1704.00
Chapter-Level	70	94	23	23	23	23	70.00	94.00
Section-Level	69	93	247	257	358	358	105.95	137.47
Subsection-Level	16	26	405	445	804	804	31.83	48.19
Paragraph-Level	2	3	787	893	2626	2632	5.48	7.68
Sentence-Level	1	1	1113	1357	9098	7976	4.03	3.56

In addition to the alignment of sentence, keywords may or may not appear in both of the aligned English and Chinese sentences and the length of keywords in English and Chinese are not necessary the same. Such problem has significant impact on the thematic weight of keywords utilized in the summarization techniques. Table 2 shows the overall statistics of "Hong Kong" and "香港" in the bilingual Hong Kong Annual Report 2000. The term frequency and text block frequency of "Hong Kong" and "香港" at different levels of the parallel corpus are significantly different. The total frequency of "Hong Kong" is 1217, and the total frequency of "香港" is 1704. The frequency of "香港" is much higher than that of "Hong Kong". In addition, the

measurements of the length of keywords in English and Chinese are different. It highly affects the computation of *tfidf* scores of the English and Chinese terms. As a result, the *tfidf* scores for a pair of equivalent keywords in two languages are usually significantly different. However, they are positively correlated. Table 3 shows the correlation of the *tfidf* scores of "Hong Kong" and "香港" at different levels of the Hong Kong Annual Report 2000.

Table 3. Correlation of *tfidf* Scores of "Hong Kong" and "香港" in the Hong Kong Annual Report 2000 at Different Document Levels.

Document Levels	Correlation of Keyword 'Hong Kong' and Keyword '香港'
Chapter-Level	0.8456
Section-Level	0.8588
Subsection-Level	0.7574
Paragraph-Level	0.4147

Table 4. Length of Sentences in Hong Kong Annual Report 2000 (English Version and Chinese Version).

	Length of Sentence in Chinese (No. of Characters)	Length of Sentence in English (No. of Words)
Lower Limit	2	2
Lower Quartile	24	15
Median	33	21
Upper Quartile	45	29
Upper Limit	215	128
Mean	36.16	23.14
Standard Deviation	17.30	11.10

The measurements of sentence length in Chinese and English sentences are different. The Chinese text is character based and the sentence length is measured by number of characters. However, the English text is word based and the sentence length is measured by number of words. One English word usually consists of several Chinese characters; therefore the number of characters in Chinese sentences is usually more than the number of words in English sentences. The statistics of sentence lengths of the bilingual Hong Kong Annual Report 2000 in Chinese and English is show in Table 4. There is no significant difference in dispersion of sentence length in two languages, the standard deviation of sentence length in both languages is about half of the arithmetic mean of the sentence length. The difference of sentence length and the sentence alignment in two languages may affect the sum of *tfidf* score of terms in sentences. The sum of *tfidf* score of terms in sentence of two languages is shown in Table 5. It is shown that there is no significant difference in the dispersion of the sum of *tfidf* score of terms in sentences in two languages.

We have also analyzed the sum of *tfidf* score of the constituent terms in a sentence against its sentence length. The correlation coefficient of sentence length and the sum of the *tfidf* score of terms in Chinese sentences is 0.62, which means there is a weak positive correlation. On the other hand, the correlation coefficient of sentence length and the sum of the *tfidf* score of terms in English sentences is 0.52. The correlation in English document is even weaker than Chinese document. Therefore, the relationship

between sentence length and the sum of the *tfidf* score of sentences is not strong. However, since a longer sentence tends to have a larger sum of the *tfidf* score, the longer sentence will have a higher probability to be extracted by the summarization techniques as part of a summary.

Table 5. Sum of *tfidf* Score of Terms in Sentences of Hong Kong Annual Report 2000 (Chinese and English Version).

	HKAR 2000 (Chinese Version)	HKAR 2000 (English Version)
Lower Limit	0	0
Lower Quartile	263.74	231.20
Median	464.09	471.09
Upper Quartile	767.16	803.91
Upper Limit	6313.61	8120.42
Mean	584.43	599.10
Standard Deviation	478.57	527.55

5 Comparison of Summarization of Chinese and English Parallel Documents

The comparison of the summaries in two languages produced by the same summarization technique can help us to understand the impact of the grammatical and lexical difference of languages on the summarization result. In our experiment, the fractal summarization has been applied to the Chinese and English parallel documents. In this section, we present the comparison of the intersection of the summaries in two languages, and the precision of the summary generated in two languages.

The comparison of the number of sentences extracted by the fractal summarization technique in each chapter of the Chinese and English version of Hong Kong Annual Report 2000 is shown in Fig. 6. Roughly, the distributions of the number of sentences extracted from chapters in two languages are similar. The correlation of the number of sentence extracted from each chapter in the Chinese and English documents is 0.9353. It shows that they are highly positive correlated.

Although, the number of sentence extracted from chapters in two languages are very similar, they may extract different sets of sentences from the chapters. In order to compare the matching of sentences extracted in the Chinese and English summaries, we define three types of sentence matching:

- A direct match is the case when a one-to-one sentence mapping is identified from the sentences in the Chinese and English summaries.
- A partial match is the case when a one-to-many or many-to-many sentence mapping is identified from the whole sentence or the partial sentence in the Chinese and English summaries.
- An unmatch is the case when a sentence is extracted as the summary in one language, but none of its equivalent sentences is extracted in the summary of the other language.

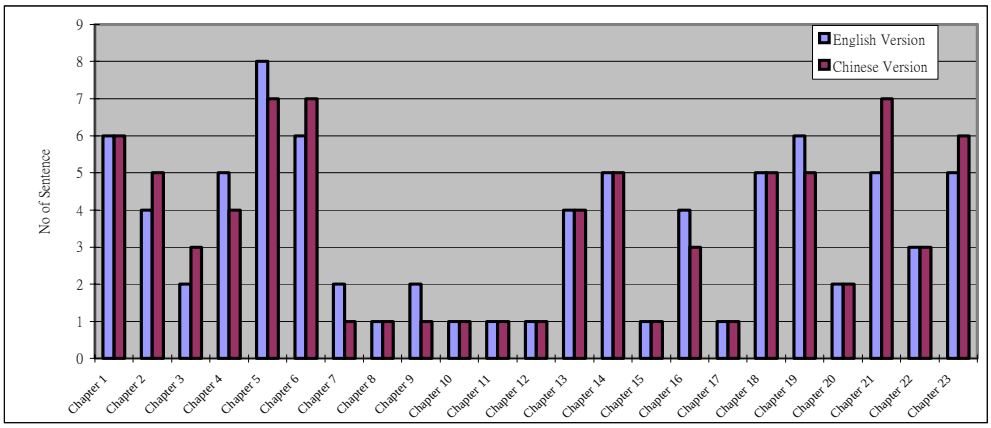

Fig. 6. No. of Sentences Extracted from Hong Kong Annual Report 2000 by Fractal Summarization with 1% Compression Ratio.

Table 6. Sentences Matching in Fractal Summaries of HKAR 2000 (Chinese and English Version) with 1% Compression Ratio.

Type of Sentence Match	Percentage
Direct Match	16.28%
Partial Match	9.30%
Unmatch	74.42%

As shown in Table 6, the intersection of extracted sentences in Chinese and English summaries is very small. The sum of direct match and partial match only corresponds to 25% of the matching of sentences, and the rest are unmatched. As presented in Table 1, the majority of the sentences in the pair of parallel documents can be aligned by one-to-one, one-to-two, or two-to-one mappings. However, applying the same summarization technique individually on the English and Chinese document produces significantly different set of sentences in the summaries. It reflects that the grammatical and lexical differences of the languages have significant impact on the summarization processes.

Since the percentage of matching sentences in the Chinese and English summaries is low, we have further investigated if there are significant differences in the content of the summaries. It is found that the content of the summaries are very close although the sentences are not exactly matched. Sentences covering similar content are extracted in the Chinese and English summaries. Table 7 presents the summaries extracted from Chapter One of the Hong Kong Annual Report 2000. Six sentences are extracted in the both of the Chinese and English summaries. C2 is a direct match of E1 and C4 is a direct match of E2. However, no matching can be found between C1, C3, C5, C6 and E3, E4, E5, E6. When we pay attention in the content of C1, C3, C5, C6 and E3, E4, E5, and D6, we find that they cover very similar content. They are all conveying similar messages about “Hong Kong as an international financial and business centre”, “transportation and communication infrastructure”, “the Pearl River Detla”, and “Hong Kong relationship with China” .

Table 7. The Chinese and English Summaries Extracted from Chapter One of the Hong Kong Annual Report 2000

C1. 香港所具備的特質，加上蓬勃的經濟、法治的自由社會、國際商貿和旅遊中心的地位、完善的運輸和電訊基建，以及龐大的國際社會，全都是代表“國際都會”的典型標記。
C2. 不過，我們明白，要香港脫穎而出，成爲國際都會，我們必須持續改進，提升香港的生活質素，例如積極保護環境、推廣藝術文化等。
C3. 香港是世界第十大貿易體系，主要由於香港是通往中國內地的門戶。
C4. 一九七八年，鄧小平先生推行“門戶開放”政策，這個轉變令香港廠商有機會擴展業務，進軍內地市場，間接幫助香港發展爲今天全球最重要的商貿金融中心之一。
C5. 其次，我們打算與廣東當局加強合作，推廣香港國際機場和貨櫃港口，促進香港與珠江三角洲的貿易往來。
C6. 我們也會繼續鞏固香港作爲亞太區中心和中國門檻的地位，力求實現目標，把香港建設爲亞洲國際都會。
E1. “We do, however, recognise that we have to advance further in improving the quality of life in Hong Kong, for example in environmental protection and arts and culture, if we are to compete as a world city.”
E2. The change brought about by Deng Xiaoping's 'open-door' policy in 1978 gave Hong Kong manufacturers an opportunity to expand and migrate across the boundary and their success has helped make Hong Kong one of the world's most remarkable trade and financial centres.
E3. “Hence, China's accession to the WTO will mean further enhancement of Hong Kong's position as an international financial and business centre, a transportation and communication hub, a centre for professional services and our traditional role as a gateway to the Mainland.”
E4. “Hong Kong's close economic relationship with the Mainland, and in particular with the rest of the Pearl River Delta, puts Hong Kong in a unique position.”
E5. “Thirdly, we will encourage Hong Kong companies to co-operate with their Pearl River Delta partners to establish logistics centres and to promote Hong Kong's logistics capabilities.”
E6. “In part, the study indicated that Hong Kong's dominant position stems from its political and legal stability, proximity to major markets (Hong Kong is within five hours flying time of half the world's population), excellent infrastructure, its dense network of financial and professional service firms and the quality of its local management.”

We further investigate if there is any significance difference in the performance of the fractal summarization technique on different languages in terms of precision. The precision of a summary is computed as follow:

$$\frac{\textit{no. of senetnces accepted by the user as part of the summary}}{\textit{no. of sentences in the summary}}$$

A user evaluation with ten subjects is conducted The average precision of English summary is 85.125% and the average precision of Chinese summary is 85.25%. The highest precisions of summaries in two languages are both 91.25%. There is no substantial difference in precision of summaries in Chinese and English.

As a conclusion, we find that the sentences extracted in the Chinese and English summaries are significantly. However, the performances of the summaries in terms of precision are very close. In addition, the content of the extracted sentences in the summaries are similar although they are directly matched. These evidences show that the grammatical and lexical differences between languages have significant effect on the extraction of sentences in their summaries. However, the overall performances of the summaries do not have any significant differences.

6. Conclusion

Automatic text summarization is important as the information overloading problem becomes serious on the World Wide Web due to the exponential growth of information in real time. Information available in languages other than English on the World Wide Web is growing significantly. Techniques for processing or summarizing English documents only are not able to satisfy the needs of Internet users. It is desire to determine if the existing techniques can perform in English and other languages. In this paper, we have investigated the impact of the grammatical and lexical differences of English and Chinese on the fractal summarization techniques. The performances of the fractal summarization on English and Chinese parallel documents are also investigated. It is found that the differences of the languages have significant effect on the extraction processes of sentences for summarizing English and Chinese documents. However, the overall performances of the summarization in English and Chinese are similar. The content covered in the summaries is similar although the sentences extracted may not be matched.

7. References

1. Barnsley M. F., and Jacquin, A. E. Application of Recurrent Iterated Function Systems to Images. In Proceedings of SPIE Visual Communications and Image Processing'88, 1001, 122-131, 1988.
2. Baxendale P. Machine-Made Index for Technical Literature - An Experiment. IBM Journal (October), 354-361, 1958.
3. Chen, H.H. and Huang, S.J. A Summarization System for Chinese News from Multiple Sources. In Proceedings of 4th International Workshop on Information Retrieval with Asia Languages, 1-7. 1999.
4. Cowie J., Mahesh K., Nirenburg S., and Zajaz R. "MINDS-Multilingual Interactive Document Summarization". In Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization. 131-132. California, USA. AAAI Press, 1998.
5. Edmundson H. P. New Method in Automatic Extraction. Journal of the ACM, 16(2) 264-285, 1968.

6. Endres-Niggemeyer B., Maier E., and Sigel A. How to Implement a Naturalistic Model of Abstracting: Four Core Working Steps of an Expert Abstractor. Information Processing and Management, 31(5) 631-674, 1995.
7. Feder J. Fractals. Plenum, New York, 1988.
8. Frakes W. Stemming Algorithms. In William B. Frakes and Ricardo Baeza-Yates, editors, Information Retrieval: data structures and algorithms, 131-160. Prentice-Hall, Englewood Cliffs, NJ (1992).
9. Gallager R., Information theory and reliable communication, 1968.
10. Gan, K.W., Palmer, M., and Lua, K.T., A Statistically Emergent Approach for Language Processing: Application to Modeling Context effects in Ambiguous Chinese Word Boundary Perception. Computational Linguistics, 531-553, 1996.
11. Glaser B. G., and Strauss A. L. The Discovery of Grounded Theory; Strategies for Qualitative Research. Aldine de Gruyter, New York, 1967.
12. Hearst M. A. Subtopic Structuring for Full-Length Document Access. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 56-68, 1993.
13. Hull D. Stemming algorithms - a case study for detailed evaluation. Journal of the American Society for Information Science, 47(1):70--84, 1996.
14. Kataoka A., Masuyama S. and Yamamoto K. Summarization by shortening a Japanese Noun Modifier into Expression 'A no B'. In Proceedings of NLPRS'99, 409-414, 1999.
15. Koike, H. Fractal Views: A Fractal-Based Method for Controlling Information Display. ACM Transaction on Information Systems, ACM, 13(3), 305-323, 1995.
16. Lam-Adesina M., and Jones G J. F. Applying Summarization Techniques for Term Selection in Relevance Feedback. In Proceedings of SIGIR 2001, 1-9, 2001.
17. Lin Y., and Hovy E.H. Identifying Topics by Position. In Proceedings of the Applied Natural Language Processing Conference (ANLP-97), Washington, DC, 283-290, 1997.
18. Luhn H. P. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 159-165, 1958.
19. Mandelbrot B. The fractal geometry of nature. W.H. Freeman, New York, 1983.
20. Mani I. Recent Development in Text Summarization. ACM CIKM'01, 529-531, Georgia, USA, 2001.
21. Myaeng S. H., and Jang D. H., 1999. Development and Evaluation of a Statistically-Based Document Summarization System, In Advances in Automatic Text Summarization (ed: Inderjeet Mani), MIT Press. 61-70.
22. Nie, J.Y., Hannan, M.L., and Jin, W., Combining Dictionary, Rules and Statistical Information in Segmentation of Chinese. Computer Processing of Chinese and Oriental Languages, 125-143, 1995.
23. Ogden W., Cowie J., Davis M., Ludovik E., Molina-Salgado H., and Shin H. Getting information from documents you cannot read: an interactive cross-language text retrieval and summarization system. Joint ACM DL/SIGIR Workshop on Multilingual Information Discovery and Access: 1999.
24. Salton G., and Buckley C. Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 24, 513-523, 1988.
25. Yang C. C., Luk J., Yung J., and Yen J. Combination and Boundary Detection Approach for Chinese Indexing. Journal of the American Society for Information Science, Special Topic Issue on Digital Libraries, 51(4), 340-351, March, 2000.
26. Yang, C. C., and K. W. Li, Automatic Construction of English/Chinese Parallel Corpora. Journal of the American Society for Information Science and Technology, 54(8), 2003, pp.730-742.
27. Yang, C. C., and Wang, F. L., Fractal Summarization: Summarization Based on Fractal Theory, Proceedings of the 26th Annual International ACM Conference (SIGIR'03), Toronto, Canada, July 28 - August 1, 2003.