The Impact Analysis of Language Differences on an Automatic Multilingual Text Summarization System

Fu Lee Wang
Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong, People's Republic of China. E-mail: flwang@cityu.edu.hk

Christopher C. Yang
Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong, People's Republic of China. E-mail: yang@se.cuhk.edu.hk

Based on the salient features of the documents, automatic text summarization systems extract the key sentences from source documents. This process supports the users in evaluating the relevance of the extracted documents returned by information retrieval systems. Because of this tool, efficient filtering can be achieved. Indirectly, these systems help to resolve the problem of information overloading. Many automatic text summarization systems have been implemented for use with different languages. It has been established that the grammatical and lexical differences between languages have a significant effect on text processing. However, the impact of the language differences on the automatic text summarization systems has not yet been investigated. The authors provide an impact analysis of language difference on automatic text summarization. It includes the effect on the extraction processes, the scoring mechanisms, the performance, and the matching of the extracted sentences, using the parallel corpus in English and Chinese as the tested object. The analysis results provide a greater understanding of language differences and promote the future development of more advanced text summarization techniques.

Introduction

As current research has indicated, automatic text summarization helps to resolve the problem of information overload by extracting the essence of retrieved documents obtained by the information retrieval systems or search engines. Users may determine whether the information of a document meets their information needs by evaluating the summary instead of browsing through the whole document in the retrieval result. Consequently, filtering of irrelevant documents becomes more efficient.

In addition to the problem of information overload, the problem of language boundaries becomes critical. This situation has become more apparent as the information in languages other than English has been increasing significantly on the World Wide Web. To address this predicament, many techniques for automatic text summarization systems have been developed in the last decade. Summarization systems for different languages, such as French (Lehmam, 1999), German (Reimer & Hahn 1990), Chinese (Chen, Kuo, Huang, Lin, & Wung, 2003), Japanese (Kataoka, Masuyama, & Yamamoto, 1999), and Korean (Myaeng & Jang, 1999), have also been explored. Unfortunately, most of these systems are monolingual summarization systems. Although there are some multilingual summarization systems (Cowie, Mahesh, Nirenburg, & Zajaz, 1998; Ogden, et al., 1999), these systems have not yet been tested on parallel corpus. Therefore, the experimental results of these summarization systems only reflect the general performance. However, they do not reflect the impact of the language differences on the summarization techniques.

In earlier related studies, it was determined that grammatical and lexical differences do have a significant effect on text processing (Yang & Li, 2003). For example, a word in one language can be translated into one or more words in another language. In some cases, a word in one language may not be translated or may be translated in different ways in another language at different times. Conceptual alternation may also occur. In addition, the grammatical differences between languages, such as tense, aspect, and voice, can have a significant effect on text processing. Therefore, it is understandable to presume that applying the same summarization technique to both Chinese and English documents will produce different results. Clearly, it is of the utmost

Accepted March 9, 2005

 ● Published online 1 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20330

importance to analyze the way in which the summarization process is affected by the language differences. Here we apply the fractal summarization technique on a parallel corpus in both English and Chinese. The summarization results in English and Chinese are compared directly.

The fractal summarization technique adopted in this study was developed based on the fractal theory (Yang & Wang, 2003a, 2003b). In the process of fractal summarization, crucial 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 fractal theory. User evaluation has been conducted and these results have shown that fractal summarization outperforms the traditional summarization without requiring exploration of the hierarchical structure of the documents. Development of the fractal summarization technique has been based on the statistical approach. Therefore, it can be easily applied to documents in any language without major modification. However, each language has its own unique features, and the language difference may significantly affect the summarization process of the documents (Wang & Yang, 2003).

In the following sections, a brief introduction of the fractal summarization technique is provided. In addition, we present a detailed discussion of the parallelism of English and Chinese parallel documents and of the performance of automatic text summarization on English and Chinese documents. This is followed by further analysis of the matching of sentences and the content in the English and Chinese summaries. Correlation of the ranking of the sentences in the English and Chinese summaries obtained by the summarization techniques and the effect of compression ratio will also be discussed.

Automatic Summarization

As earlier research has shown, automatic summarization systems extract the most important information from source documents. Traditionally, automatic text summarization systems extract the sentences from the source documents based on their significance to the documents (Edmundson, 1969; Luhn, 1958). The summarization systems calculate the weights of different extraction features for each text unit, and then the weights are combined as the overall weight of the text unit. Subsequently, the text units with the highest weight will be extracted. The text units extracted are concatenated as a summary.

The usual practice is for automatic summarization to be divided into three stages (Hovy & Lin, 1997; Sparck-Jones, 1999). The first stage is to convert the source document into some internal representation that can be used for further analysis. The next stage is to identify the key features of a document and to transform the source document into summary representation. The last stage is to generate the summary in a specific format. There are two main approaches to this process, namely *abstraction* and *extraction*. The abstraction summarization systems analyze the information contents extracted and then regenerate the data as a summary (McKeown, Robin, & Kukich, 1995). However, it has been shown that 80% of the sentences in man-made abstracts are closely matched with sentences in the original document (Kupiec, Pedersen, & Chen, 1995). Therefore, most summarization systems generate summary by concatenating text units extracted from the source document sequentially (Edmundson, 1969; Kupiec, Pedersen, & Chen, 1995; Luhn, 1958). These are known as *extraction summarization systems*. This paradigm transforms the automatic summarization problem to ranking of sentences in the source document, and the sentences at the top-ranking level will be extracted and then concatenated as a summary.

In automatic summarization, the system usually extracts the text units with more salient and nonredundant information contents from the source document. These are then aggregated and compressed into a specific format. Although a considerable amount of text-unit extraction approaches have been proposed to perform automatic text summarization; the extraction features are usually classified into groups according to the level of processing.

- Surface-level approaches use salient features of a document to extract the important information content. These surface-level approaches were widely adopted in the early systems in the 1950s and 1960s (Edmundson, 1969; Luhn, 1958). Many features have been identified as key features; they include such features as the thematic, the location, the background, the cue-phrase, etc.
- Entity-level approaches build an internal representation for text units and their relationships. These approaches tend to use graph topology to determine which text unit is important. The entity-level approach includes the following features: similarity, proximity, co-occurrence, co-reference, the lexical chain, connectivity, the syntactic relation, the logical relation, etc. The entity-level approach was not developed until the 1970s. In recent years, some new approaches have been developed. However, some of the newer systems require human interaction, for example the building of a parse tree. Therefore, all of the summarization cannot be done automatically.
- Discourse-level approaches model the global structure of the text and its relation to communicative goals. Recently developed, the discourse-level is representative of the new research directions. The discourse-level approaches require an in-depth knowledge of linguistics and therefore are not practical in implementation. These extraction features at discourse-level include the rhetorical structure, a thread of topics, the format of the document, etc.

Many of the automatic text summarization systems have been implemented based on the different extraction features. They are evaluated by different measurements based on different corpus. As a result, these results cannot be compared directly. However, they do have common elements. All the systems identify an upper limit for the precision of the summarization system, the performance of the system will grow faster with the addition of extraction features and they will

DOI: 10.1002/asi

reach their upper boundary after three or four extraction features have been added. After that point is reached, further additions of extraction features will not improve the precision, whereas this factor will sometimes even decrease the aggregated precision of the system (Kupiec et al., 1995). It has been proven that the extraction of 20% of sentences from the source document can be as informative as the full text of a document (Morris, Kasper, & Adams, 1992). However, summaries with a high compression ratio are more useful (Teufel & Moens, 1997, 1998). As high-compression-ratio summarization systems do not have the opportunity to extract all the relevant sentences from the source document, precision is the only relevant measurement for performance with most summarization systems (Kupiec et al., 1995; Teufel & Moens, 1997, 1998).

Traditionally, automatic text summarization systems extract the sentences from the source documents based on their significance to the document (Edmundson, 1969; Luhn, 1958) without considering the hierarchical structure of the document. As summarization systems reach their upper bound of performance, most of the summarization systems adopt only three or four extraction features (Kupiec et al., 1995; Teufel & Moens, 1997, 1999). Data has shown that the thematic, location, heading, and cue-phrase features in the surface-level are the most widely used summarization features.

- Originally, the thematic feature was identified by Luhn (1958). Further improvement was introduced by Edmundson who proposed assigning the thematic weight to keywords based on term frequency and the sentence thematic weight as the sum of thematic weight of constituent keywords (Edmundson, 1969). The *tfidf* (term frequency, inverse document frequency) score is currently the most widely used approach to calculate the thematic weight of keywords (Salton & Buckley, 1988).
- The significance of a sentence is indicated by its location (Baxendale, 1958) based on the hypotheses that the topic sentences tend to occur at the beginning or in the ending of documents or paragraphs (Edmundson, 1969). Edmondson proposed to assign positive location weight to a sentence according to its ordinal position inside the document.
- The heading feature has been proposed based on the hypothesis that the author conceived the heading as circumscribing the subject matter of the document (Edmundson, 1969). A heading glossary is a list consisting of all the words in headings and subheadings with positive weights. The heading weight of a sentence is calculated by the sum of heading weight of its constituent words.
- The cue phrase feature as proposed by Edmundson (1969) was based on the hypothesis that the probable relevance of a sentence is affected by the presence of some pragmatic words, such as "conclusion." A prestored cue dictionary with cue weight has been used to identify the cue phrases. The cue weight of a sentence is calculated by means of a total derived from the cue weight of constituent words.

Typically, summarization systems select a combination of summarization features (Edmundson, 1969; Jones, 2001; Lam-Adesina & Lin & Hovy, 1997) and the total sentence weight ($W_{sentence}$) is calculated as weighted sum of the weights computed by each of the salient features, i.e.,

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

where $w_{thematic}$, $w_{location}$, $w_{heading}$ and w_{cue} are the 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 the four summarization features.

Those sentences with sentence weight higher than a threshold value are selected as part of the summary. It has been proven that the weighting of different summarization features does not have any substantial effect on the average precision in some information retrieval application (Lam-Adesina & Jones, 2001). In the present study, some experiments have been conducted to investigate the impact of weighting with summarization features on overall precision of automatic summarization. The experimental results indicate that the summarization system with equal weighting of summarization features performs best. In the current experiment, the maximum value of each summarization feature has been normalized to one, and the total weight of sentence calculated as the sum of scores of all summarization features without weighting. However, the cue-phrase feature was disabled for summarization of parallel documents, as currently there appears to be lack of a parallel cue-phrase dictionary defined for Chinese and English. Fortunately, the addition of the cue-phrase features does not significantly affect the performance of the summarization results.

Fractal Summarization Based on Fractal Theory

Previously, many summarization models have been proposed. None of the earlier models was based entirely on the document structure. Consequently, they do not take into account the fact that the human abstractors extract sentences according to the hierarchical document structure. In the literature, document structure has been described as *fractals*. In the past, fractal theory has been widely applied in the area of digital image compression (Barnsley & Jacquin, 1988; Jacquin, 1993), which is similar to text summarization in the sense that they both extract the most important information from the source and reduce the complexity of the source (Mani, 2001; Yang & Wang, 2003b). The fractal summarization model represents the first effort to apply fractal theory to document summarization. This system generates the summary by a recursive deterministic algorithm based on the iterated representation of a document.

Fractal Theory and Fractal View for Controlling Information Displayed

As previous studies have shown, fractals are mathematical objects that have a high degree of redundancy (Mandelbrot, 1983). These objects are made of transformed copies of

DOI: 10.1002/asi

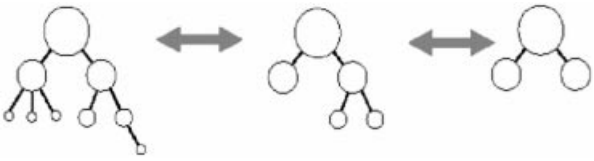

FIG. 1. Fractal view of a tree at different abstraction levels.

themselves or part of themselves. Mandelbrot (1983) was the first person to investigate the fractal geometry and subsequently developed the fractal theory. He applied his theory to the measurement of the length of the British coastline. This is a well-known example that depends on the measurement scale. The larger the scale, the smaller is the value of the length of the coastline and the higher the abstraction level. For example, the British coastline includes bays and peninsulas. Bays include sub-bays and peninsulas include sub-peninsulas. By using fractals to represent these structures, abstraction of the British coastline can be generated with different degrees of abstraction.

A tree is one of the classical examples of fractal objects. A tree consists of numerous subtrees or branches and each one is a tree as well. By changing the scale, the different levels of abstraction views are obtained (Figure 1). Clearly, the idea of a tree can be extended to include any hierarchical structure. A fractal tree is a tree structure where the degree of importance of each node is represented by its fractal value. The fractal value of a root is set to 1, and the fractal value is propagated to other nodes with the following expression:

$$\begin{cases} Fv_{root} = 1 \\ Fv_{child\ node\ r\ of\ x} = \frac{C\,Fv_x}{N^{D^{-1}}} \end{cases}$$

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. The fractal dimension describes the complexity of the fractal object. For example, if there are many child nodes in a tree, then it is more complex, and therefore it has a higher value of fractal dimension. To simplify this discussion, both C and D are considered to be 1 in the experiment. Giving a different value to the fractal dimension will change the fractal value of each node. This leads to the magnification effect of certain nodes. The effect of other values of fractal value dimension will be investigated in future research.

It has been established that the fractal view is a fractal-based method for controlling information displayed (Koike, 1995). A 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 and the nodes with a fractal value less than the threshold value will be invisible to the observer (Figure 2). By changing the threshold value, the observer can adjust the amount of information displayed.

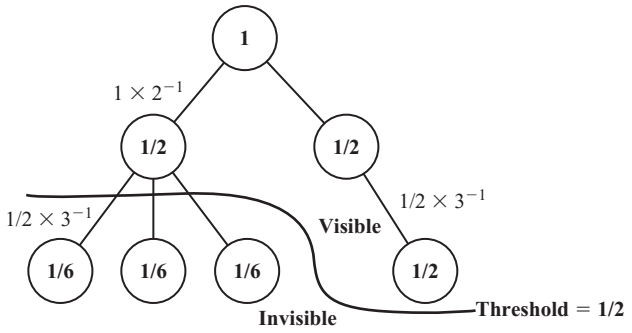

FIG. 2. An example of the propagation of fractal values.

Fractal Summarization

Many studies of the human abstraction process have shown that human abstractors extract the topic sentences according to the document structure from the upper level to the lower level until they have extracted sufficient information (Endres-Niggemeyer, Maier, & Sigel, 1995; Glaser & Strauss, 1967). Advance summarization techniques consider the document structure 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 tend to ignore the structure of the document. The fractal summarization model has been proposed to generate a summary based on the hierarchical structure of document (Yang & Wang, 2003a, 2003b).

Development of fractal summarization is based on fractal theory, one in which the more 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 fractal theory. Using the example of fractal geometry as applied to the measurement of the British coastline where the coastline includes bays, peninsulas, sub-bays, and sub-peninsulas, a large document also has a hierarchical structure with several levels, chapters, sections, subsections, paragraphs, sentences, terms, words, and characters. Similarly, a document can be represented by a hierarchical structure (Figure 3). However, a document is not a true mathematical fractal object because a document cannot be viewed in an infinite abstraction level. The smallest unit in a document is a character. However, neither a character nor a word will convey any meaningful information concerning the overall content of a document. The lowest abstraction level for consideration in this study is a term. A document is considered as prefractal, that is fractal structures in their early stages with finite recursion only (Feder, 1988).

As previous studies have shown, the fractal summarization model is developed based on techniques from the fractal view and fractal image compression (Barnsley & Jacquin, 1988; Jacquin, 1993). In image compression, an image is regularly segmented into a set of nonoverlapping square blocks, called *range blocks* and then each range block

DOI: 10.1002/asi

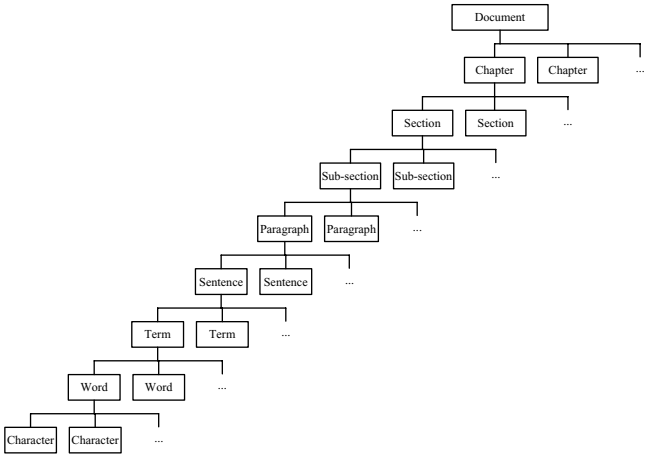

FIG. 3. The prefractal structure of a large document.

is subdivided into subrange blocks, until a contractive mapping can be found to represent this subrange 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 a fractal tree structure according to its document structure. The weight of specific sentences under a range block is calculated by the traditional summarization methods. The fractal value of root node is 1 and the fractal values of the child node are propagated according to the sum of sentence weights of sentences under the child nodes.

$$\begin{cases} Fv_{root} = 1 \\ Fv_{child\ node\ r\ of\ x} = Fv_x \times \dfrac{\sum Sentence\ weight_{sentences\ under\ r}}{\sum Sentence\ weight_{sentences\ under\ x}} \end{cases}$$

In the above formula, the sum of sentence weights of sentences under a range block includes all the sentences under the subrange blocks that are child nodes of the range block. For example, the sentence weight of range block x includes all the sentences under range block r, because the range block is a subrange block under x.

Given a document, users provide a compression ratio to specify the amount of information displayed. The compression ratio of summarization is defined as the ratio of the 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 the 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 (Figure 4). A threshold value is the maximum number of sentences that can be extracted from a range block. If the quota is larger than the threshold value, the range block must be divided into a subrange block. It has been proven that for summarization by extraction of a fixed number of sentences, the optimal length of summary is three to five sentences (Goldstein, Kantrowitz, Mittal, & Carbonell, 1999). For the present study, the default value of the threshold has been specified as five.

Figure 4 demonstrates an example of fractal summarization model. The fractal value of the root is 1 and the system extracts 40 sentences from the root node. The system then allocates the sentence quota to the child nodes directly proportional to the fractal value of child node. The fractal value and sentence quota will be propagated to the grandchild nodes. For example, Section 1.2 receives a quota of six sentences. As this is higher than the threshold value, the system will extend the node in paragraph levels. However, Sections 1.1 and 1.3 receive a quota less than five sentences. Therefore, the system directly extracts sentences at the section level.

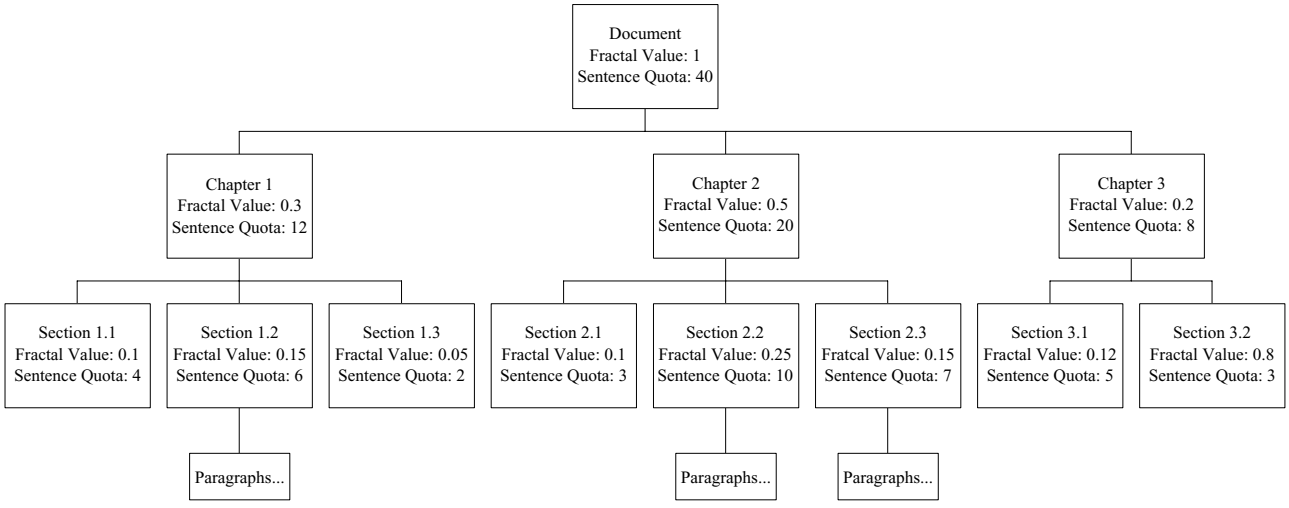

FIG. 4. An example of fractal summarization model.

DOI: 10.1002/asi

The details of the fractal summarization algorithm are shown below:

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 a 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 the fractal values.
   7.3. For each child node,
        If the quota is less than the threshold value:
            Select the sentences in the range block by extraction
        Else
            Set the current node to the child node;
            Repeat Steps 7.1., 7.2., 7.3.
8. Until all the child nodes under the current node are processed.

Experimental Result

It has been established that a full-length text document contains a set of subtopics (Hearst, 1993) and that a good quality summary should cover as many subtopics as possible. Experiments on fractal summarization and traditional summarization have been conducted on the *Hong Kong Annual Report 2000* (Yang & Wang, 2003a, 2003b), using the traditional summarization model without considering the hierarchical structure of the documents, extracting most of the sentences from a few chapters. However, the fractal summarization model extracts the sentences distributively from each chapter. The results indicated that the fractal summarization model produced a summary with a wider coverage of information subtopic than a traditional summarization model. A user evaluation has been conducted to compare the performance of the fractal summarization with that of the traditional summarization without using the hierarchical structure of documents. These results show that all subjects consider the summary generated by fractal summarization method to be a better summary. The fractal summarization can achieve up to 91.25% precision and 87.125% on average, but the traditional summarization only can achieve up to a maximum of 77.50% precision and 67% on average. Experimental results showed that the fractal summarization model outperforms the traditional summarization at 99% confidence level (Wang & Yang, 2003).

Parallelism Chinese and English Parallel Documents

Parallel documents are available in many cities around the world that have multilingual cultures, such as Quebec, Hong Kong, and Singapore as well as cities in many European countries. In the case of Hong Kong, as a British colony for more than a century, there has been a bilingual culture. Consequently, the official languages are Chinese and English and many important documents are written in Chinese and English, using covert translation (Yang & Li, 2003). For example, most of the documents released by the government have both Chinese and English versions. As these documents are written by experienced bilingual linguists, the quality can be assured. In this section, the characteristics in the parallelism of Chinese and English parallel documents will be examined.

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 (Yang & Li, 2003). Because of 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.

The sentence structure for a pair of parallel documents in two languages is different due to the grammatical and lexical differences in languages. For example, as shown in Figure 5, the orderings of terms in two languages are not the same. The structure of a 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. As shown in Figure 6, the three sentences in the Chinese version are merged into one sentence in English. Moreover, the keyword "香港" (equivalent to "Hong Kong") is missing in the English sentence.

Taking the *Hong Kong Annual Report 2000* as an example (as illustrated in Table 1), there are 7,976 sentences in the Chinese version, but there are 9,098 sentences in the English version. There is strong evidence to show that the mappings of English and Chinese sentences are not necessary one-to-one. Sometimes, one sentence in one language will be split into several sentences in another language. In the worst

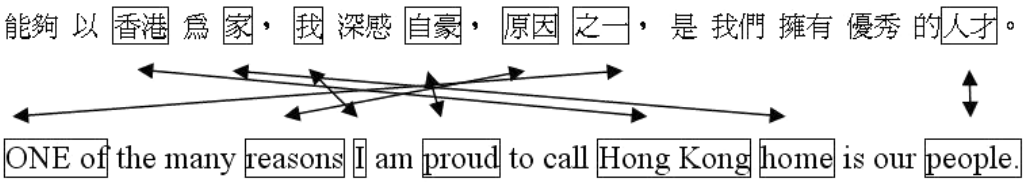

FIG. 5. The reordering of equivalent terms in Chinese and English sentences.

DOI: 10.1002/asi

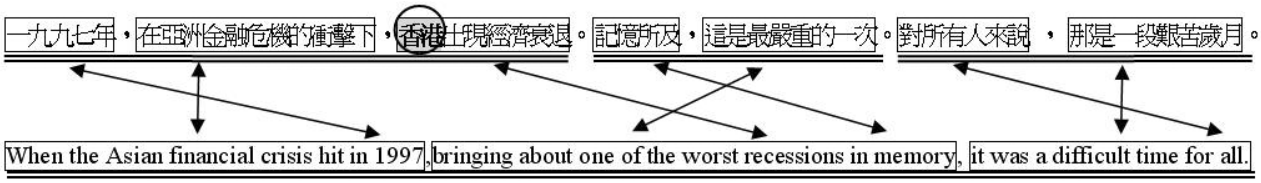

FIG. 6. An example of sentence alignment and missing keywords in Chinese and English parallel documents.

TABLE 1. The number of text blocks in the *Hong Kong Annual Report 2000.*

	HKAR 2000 (English version)	*HKAR 2000* (Chinese version)
No. of chapters	23	23
No. of section	358	358
No. of subsections	804	804
No. of paragraphs	2626	2,632
No. of sentence	9098	7,976

case, several sentences will be mixed together and rephrased as several sentences in another language. As shown in Table 1, the document structures of a pair of parallel documents are the same at a high abstraction level. However, at a lower abstraction level, they exhibit a significant difference in their structure.

To study the alignment of sentences in two languages, an analysis has been done of the mapping of sentences in *Hong Kong Annual Report 2000* (Table 2). Results indicate that 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 6.06% 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 in total yield more than 98%. Most sentences contain two or less subsentences in the other language.

In addition to the alignment of sentences, 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. A problem such as this has significant impact on the thematic weight of keywords utilized in the summarization techniques. Considering the previous example in Figure 6, the term "香港" appears in the Chinese sentence, but the equivalent term "Hong Kong" does not appear in the English sentence. Table 3 shows the overall statistics of "Hong Kong" and "香港" in the bilingual

TABLE 2. The mapping of sentences of the *Hong Kong Annual Report 2000.*

No. of mappings (%)		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 3. Statistics and *tfidf* score of keyword "Hong Kong" and "香港" in the first aligned sentence in the *Hong Kong Annual Report 2000.*

	Term frequency		Text block frequency		Number of text block		*tfidf* Score	
Text unit	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

DOI: 10.1002/asi

TABLE 4. The 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	0.8456
Section	0.8588
Subsection	0.7574
Paragraph	0.4147

Hong Kong Annual Report 2000. For clarification, the term *frequency* is the frequency of the term in the text unit; *text block frequency* is the number of text blocks that contain the term, and number of text block is the number of text blocks in the corpus.

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 1,217, and the total frequency of "香港" is 1,704. 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 has considerable effect on the computation of *tfidf* scores of the English and Chinese terms. As a result, the *tfidf* scores for a pair of equivalent terms in two languages are usually significantly different. However, they are positively correlated. Table 4 shows the correlation of the *tfidf* scores of "Hong Kong" and "香港" at different levels of the *Hong Kong Annual Report 2000*. At a lower level of abstraction, the difference of the-matic weights between a pair of equivalent terms becomes more significant.

As the data indicates, 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 shown in Table 5. There is no significant difference in dispersion of sentence length in the 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 sentences of the two languages is shown in Table 6. These results indicate that there is no significant difference in the dispersion of the sum of *tfidf* score of terms in sentences in the two languages.

TABLE 5. The length of sentences in the *Hong Kong Annual Report 2000.*

	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
SD	17.30	11.10

TABLE 6. The sum of the *tfidf* score in sentences of the *Hong Kong Annual Report (HKAR) 2000*.

	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
SD	478.57	527.55

An analysis has been done for 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 weakly positive correlation. Whereas the correlation coefficient of sentence length and the sum of the *tfidf* score of terms in English sentences is 0.52. The correlation in the English document is even weaker than that of the Chinese document. Therefore, the relationship between sentence length and the sum of the *tfidf* score of sentences is not strong. However, because 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.

Comparison of Summarization of Chinese and English Parallel Documents

The comparison of the summaries in the two languages produced by the same summarization technique can facilitate understanding of the impact of the grammatical and lexical difference of languages on the summarization result. In the present experiment, the fractal summarization has been applied to the Chinese and English parallel documents. In this section, a comparison is presented of the sentence matching of the summaries, the precision of the summaries, the effect of the compression ratio, and difference of matched and unmatched sentences.

Sentence Matching in the Summaries

As previously discussed, automatic summarization systems rank the sentences in descending order of sentence weight and the sentences at the upper levels are extracted as

DOI: 10.1002/asi

TABLE 7. Pearson's correlation and Spearman's correlation of the sentences in the *Hong Kong Annual Report 2000*.

	Pearson's correlation coefficient	Spearman's correlation coefficient
Mean	0.54	0.52
Max	0.68	0.66
Min	0.41	0.40
SD	0.08	0.08

a summary. To analyze the extraction process, the correlation of sentence weight and their ranking in the two languages are measured (Kendall & Gibbons, 1990; Harman, 1992). The Pearson's correlation coefficient is used to measure the correlation of the sentence weights and the Spearman's rank correlation coefficient is used to measure the correlation of the sentence ranking in the document. Table 7 presents the result of the correlation analysis. The results indicate that the Pearson's correlation coefficient and the Spearman's correlation coefficient are very similar. They are moderately positively correlated. In other words, the sentences extracted in the two languages should be moderately similar.

For the purpose of this study, fractal summarization has been conducted on the bilingual version of the *Hong Kong Annual Report 2000*. The comparison of the number of sentences extracted by the fractal summarization technique in each chapter of the report is shown in Figure 7. The document is analyzed by chapters; therefore, the maximum value, minimum value, mean, and the standard score of correlation coefficient for the chapters are reported. The distributions of the number of sentences extracted from chapters in the two languages are similar. The correlation of the number of sentences extracted from each chapter is 0.9353. Results show that they are highly positively correlated.

Although the number of sentences extracted from the chapters in the two languages is very similar, the system may extract different sets of sentences from the chapters. To compare the matching of sentences extracted in the Chinese and English summaries, three types of sentence matching are defined.

TABLE 8. The sentences matching in fractal summaries of *HKAR 2000* (Chinese and English version) with a 1% compression ratio.

Type of sentence match	%
Direct match	16.28
Partial match	9.30
Nonmatch	74.42

- 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.
- A nonmatch 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.

As shown in Table 8, the sentence matching in the Chinese and English fractal summaries of the *Hong Kong Annual Report 2000* has been analyzed. The intersection of sentences extracted in Chinese and English summaries is very small. The sum of direct match and partial match only corresponds to 25% of sentences in the summaries, and the rest are unmatched. As indicated in Table 2, 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 documents produces a significantly different set of sentences in the summaries. This reflects the fact that grammatical and lexical differences of the languages have significant impact on the summarization processes.

The percentage of matching sentences in the Chinese and English summaries is low; hence, further investigation has taken place to determine whether there are any significant differences in the content of the summaries. The results confirm that the content of the summaries are very close although the sentences are not exactly matched. Sentences

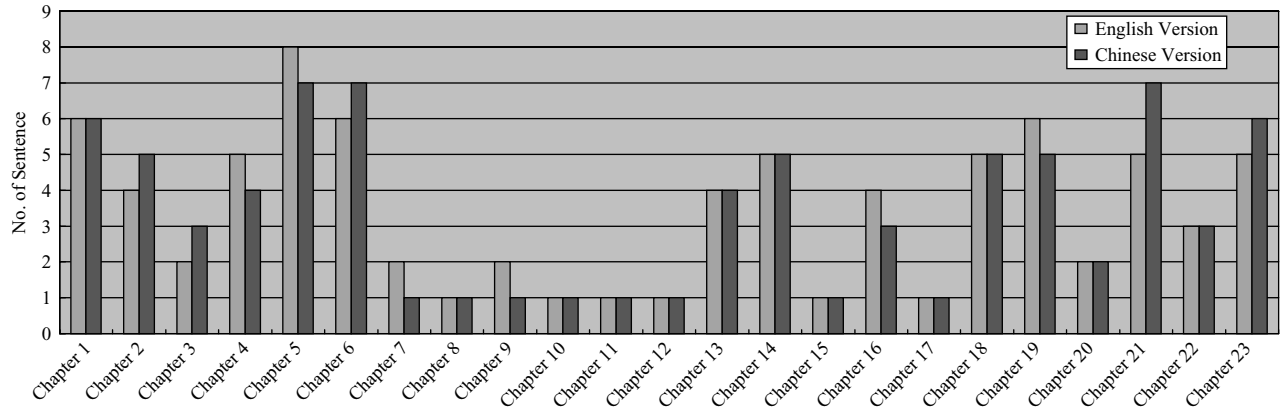

FIG. 7. The number of sentences extracted from the *Hong Kong Annual Report 2000* by fractal summarization with 1% compression ratio.

DOI: 10.1002/asi

TABLE 9. The Chinese and English summaries extracted from chapter 1 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, recognize 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."

covering similar content are extracted in the Chinese and English summaries. Table 9 presents the summaries extracted from chapter 1 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 identified between C1, C3, C5, C6 and E3, E4, E5, E6. With careful analysis, it was found that the content of C1, C3, C5, C6 and E3, E4, E5, E6, covers very similar material. They all convey similar messages about "Hong Kong as an international financial and business centre," "transportation and communication infrastructure," "the Pearl River Delta," and "Hong Kong relationship with China."

Analysis of Precision

The sentences extracted convey similar content; however, of particular interest to this study is whether there is any significant difference in their performance in terms of precision. A user evaluation with 10 subjects was conducted. All the subjects were Chinese–English bilingual college graduates. Because of limitation of time, the full version of the *Hong Kong Annual Report* was not presented to the subjects. As all the subjects reside in Hong Kong and they have a general knowledge about the contents of the *Hong Kong Annual Report*, they can judge the quality of the summaries. Both Chinese and English fractal summaries of the *Hong Kong Annual Report 2000* were presented to each subject in random order and the subjects were asked to accept or reject each sentence in the summaries based on if they would select the sentence as part of summary of the document. The precision of a summary was computed as the ratio of sentences accepted by a user to the total number of sentences in the summary, i.e.,

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

The result of the evaluation is presented in Table 10. The average precision of the English summary is 85.125% and the average precision of the Chinese summary is 85.25%. The highest precisions of summaries in the two languages are both 91.25%. The *p*- value in a two-tailed *t* test is 0.9059, therefore, there is no substantial difference in the precision of summaries between Chinese and English.

TABLE 10. The precision of summaries of *HKAR 2000* by fractal summarization with a 1% compression ratio.

User ID	Fractal summarization model *HKAR 2000* (English version)	Fractal summarization model *HKAR 2000* (Chinese version)
User 1	81.25%	83.75%
User 2	85.00%	82.50%
User 3	80.00%	80.00%
User 4	85.00%	90.00%
User 5	88.75%	87.50%
User 6	81.25%	82.50%
User 7	91.25%	87.50%
User 8	86.25%	91.25%
User 9	85.00%	81.25%
User 10	87.50%	86.25%

DOI: 10.1002/asi

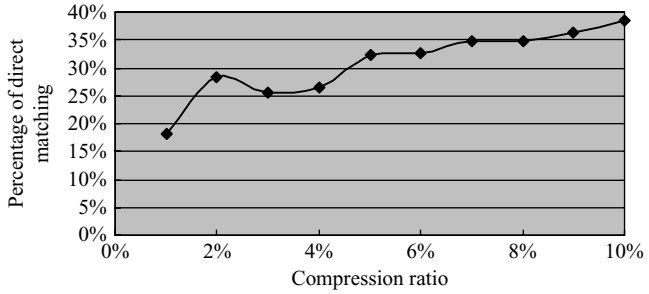

FIG. 8. The percentage of direct matching of sentences extracted in English and Chinese summaries versus the compression ratio.

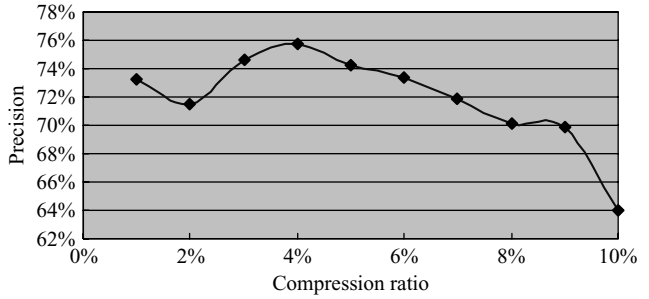

FIG. 9. The precision of English and Chinese summaries versus the compression ratio.

Effect of the Compression Ratio on Sentence Matching and Precision of Summaries

The sentence matching and precision of summaries has been compared when the compression ratio of the summaries is 1%. In this subsection, analysis will show how sentence matching and precision of summaries is affected by changing the compression ratio. As summaries with high compression ratio are more useful (Teufel & Moens, 1997), only the summaries with compression ratio less than 10% have been analyzed.

Figures 8 and 9 show the percentage of direction matching and precision at different compression ratios, respectively. As the compression ratio of summaries increases, the number of sentences extracted in the summaries increases as proportional to the number of sentences in the documents and the percentage of direct matching of the sentences in English and Chinese summaries increases. However, the precision of the summaries decreases. As difference sets of sentences are identified as top-ranking sentences, therefore the percentage of direct match is not high when the compression ratio is low. As the compression ratio increases, more and more middle-ranking sentences are included; it is apparent that the percentage of direct match increases. However, the middle-ranking sentences are less significant than the top-ranking sentences, therefore the precision decreases as the compression ratio increases.

Precision of Direct Matched Sentences and Unmatched Sentences

The precision of both of the English and Chinese summaries are considerably high although the percentage of direct matching of sentence is low; thus, it is important to establish the degree of precision in the direct matched and unmatched sentences in the English and Chinese summaries.

TABLE 11. The precision of direct matched and unmatched sentences in English and Chinese summaries.

	Chinese summaries		English summaries	
	Direct matched	Unmatched	Direct matched	Unmatched
Precision	77.39%	63.32%	75.00%	67.49%

As the summarization features are affected by language differences, the percentage of the direct match of the sentence extracted is not high. An experiment was conducted to investigate the precision of direct matched sentences and unmatched sentences. Fractal summarization was applied to parallel documents in these two languages independently at a 1% compression ratio. For comparison, the sentences extracted in the two languages were matched by human professionals. An experiment consisting of 10 subjects was conducted to compare the precision of direct matched and unmatched sentences in the summaries. The result is shown in Table 11. Results show that the precision of the direct matched sentences is significantly higher than that of the unmatched sentences by 14% and 7.5%, respectively, in both the Chinese and English summaries. In other words, if a sentence is identified as significant in both languages, this sentence is more significant. Therefore, the precision can be improved by combining sentence weights of parallel sentences in the two languages, i.e., the sentence score is adjusted based on the sentence score in another language. Whereas the precision of the unmatched sentences is still acceptable at around 65%, which is considerably higher, therefore the summarization of a document independently for each language will provide good results.

Conclusion

Clearly, automatic text summarization has become important as the information-overload problem becomes serious on the Web. This can be attributed to the exponential growth of information in real-time. Information available in languages other than English on the Web is growing significantly. Techniques for processing or summarizing English documents are not able to satisfy the needs of Internet users. Consequently, there is an urgent need to determine whether the existing techniques can perform in English and other languages.

In this article, we have investigated the impact of the grammatical and lexical differences of English and Chinese on fractal summarization techniques. The performances of the fractal summarization on English and Chinese parallel documents were also examined. Based on these results, it was determined that the differences of the languages have a

DOI: 10.1002/asi

significant effect on the sentence score of the sentence in the two languages, and the ranking of sentence is therefore affected. It was also apparent that the sentences extracted in the Chinese and English summaries are significantly different. This evidence clearly illustrates that the grammatical and lexical differences between languages have a significant effect on the extraction of sentences in their summaries. As a result, more efforts are required to investigate the impact of language difference on the design of multilingual information retrieval systems.

Although the extracted sentences are different, apparently the essence of the document is not compromised. 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 not direct matched. In addition, the precision of the direct matched sentences is significantly higher than that of the unmatched sentences. Given this insight, the precision of summaries can be improved by considering the sentence scores of the parallel sentences in two languages. Conversely, large numbers of documents have been written in more than one language because of increased international cooperation. In the future development of summarization systems, it is very important to study the way in which features in different languages can be combined to improve the overall performance of the systems.

Acknowledgment

This project was supported by the Earmarked Grant for Research from the Hong Kong Research Grant Council, 4335/02E.

References

Barnsley, M.F., & Jacquin, A.E. (1988). Application of recurrent iterated function systems to images. In Proceedings of SPIE Visual Communications and Image Processing '88 (pp. 122–131). Bellington, WA: SPIE.

Baxendale P. (1958). Machine-made index for technical literature—An experiment. IBM Journal of Research and Development, 2(4), 354–361.

Chen, H.H., Kuo, J.J., Huang, S.J., Lin, C.J., & Wung, H.C. (2003). A summarization system for Chinese news from multiple sources. Journal of the American Society for Information Science and Technology, 54(3), 1224–1236.

Cowie, J., Mahesh, K., Nirenburg, S., & Zajaz, R. (1998). MINDS-Multilingual interactive document summarization. In Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization (pp. 131–132). Menlo Park, CA: AAAI.

Edmundson, H.P. (1969). New method in automatic extraction. Journal of the ACM, 16(2), 264–285.

Endres-Niggemeyer B., Maier E., & Sigel, A. (1995). How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. Information Processing and Management, 31(5), 631–674.

Feder, J. (1988). Fractals. New York: Plenum.

Glaser, B.G., & Strauss, A.L. (1967). The discovery of grounded theory: Strategies for qualitative research. New York: Aldine de Gruyter.

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99) (pp. 121–128). New York: ACM.

Harman, D.K. (1992). Ranking algorithms. In W.B. Frakes & R. Baeza-Yates (Eds.), Information retrieval: Data structures and algorithms (pp. 363–392). Englewood Cliffs, NJ: Prentice-Hall.

Hearst, M.A. (1993). Subtopic structuring for full-length document access. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93) (pp. 56–68). New York: ACM.

Hovy, E.H., & Lin, C. (1997). Automated text summarization in SUMMARIST. In Proceedings of the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization (pp. 18–24). Morristown, NJ: ACL.

Jacquin, A.E. (1993). Fractal image coding: A review. Proceedings of the IEEE, 81(10), 1451–1465.

Kataoka, A., Masuyama, S., & Yamamoto, K. (1999). Summarization by shortening a Japanese noun modifier into expression 'A no B.' In Proceedings of the 5th Natural Language Processing Pacific Rim Symposium 1999 (NLPRS'99) (pp. 409–414). Beijing: Tsinghua University Press.

Kendall, M., & Gibbons, J.D. (1990). Rank correlation methods (5th ed.). New York: Edward Arnold.

Koike, H. (1995). Fractal views: A fractal-based method for controlling information display. ACM Transaction on Information Systems, 13(3), 305–323.

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '95) (pp. 68–73). New York: ACM.

Lam-Adesina, M., & Jones, G.J.F. (2001). Applying summarization techniques for term selection in relevance feedback. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001) (pp. 1–9). New York: ACM.

Lehmam, A. (1999). Text structuration leading to an automatic summary system: RAFI. Information Processing and Management, 35(2), 181–191.

Lin, Y., & Hovy, E.H. (1997). Identifying topics by position. In Proceedings of the Applied Natural Language Processing Conference (ANLP-97) (pp. 283–290). San Francisco: Morgan Kaufmann.

Luhn, H.P. (1958.) The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159–165.

Mandelbrot, B. (1983). The fractal geometry of nature. New York: W.H. Freeman.

McKeown, K., Robin, J., & Kukich, K. (1995). Designing and evaluating a new revision-based model for summary generation. Information Processing and Management, 31(5), 703–733.

Mani, I. (2001). Recent development in text summarization. In Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM'01) (pp. 529–531). New York: ACM.

Myaeng, S.H., & Jang, D.H. (1999). Development and evaluation of a statistically-based document summarization system. In I. Mani & M. Maybury (Eds.), Advances in automatic text summarization (pp. 61–70). Cambridge, MA: MIT Press.

Morris, G., Kasper, G.M., & Adams, D.A. (1992). The effect and limitation of automated text condensing on reading comprehension performance. Information System Research, 3(1), 17–35.

Ogden, W., Cowie, J., Davis, M., Ludovik, E., Molina-Salgado, H., & Shin, H. (1999). Getting information from documents you cannot read: An interactive cross-language text retrieval and summarization system. In Proccedings of the Joint ACM DL/SIGIR Workshop on Multilingual Information Discovery and Access (pp. 120–128). New York: ACM.

Reimer, U., & Hahn, U. (1990). An overview of the text understanding system TOPIC. In U. Schmitz, R. Schütz, & A. Kunz (Eds.), Linguistic approaches to artificial intelligence (pp. 305–320). Frankfurt: P. Lang.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24, 513–523.

Sparck-Jones, K. (1999). Automatic summarising: Factors and directions. In I. Mani & M. Maybury (Eds.), Advances in automatic text summarization (pp. 1–14). Cambridge, MA: MIT Press.

DOI: 10.1002/asi

Teufel, S., & Moens, M. (1999). Argumentative classification of extracted sentences as a first step towards flexible abstracting. In I. Mani & M. Maybury (Eds.), Advances in automatic text summarization. Cambridge, MA: MIT Press.

Teufel, S., & Moens, M. (1998). Sentence extraction and rhetorical classification for flexible abstracts. In Proceedings of the AAAI Spring Symposium on Intelligent Text Summarization (pp. 89–97). Menlo Park, CA: AAAI.

Teufel, S., & Moens, M. (1997). Sentence extraction as a classification task. In Proceedings of the Workshop of Intelligent and Scalable Text Summarization (pp. 56–68). Morristown, NJ: ACL.

Wang, F.L., & Yang, C.C. (2003). Automatic summarization of Chinese and English parallel documents. In T. Sembok, H. Zaman, H. Chen, S. Urs, & S. Myaeng (Eds.), Proceedings of 6th Internationl Conference on Asian Digital Libraries, (ICADL 2003), Digital libraries: Technology and management of indigenous knowledge for global access, Lecture Notes in Computer Science, 2911 (pp. 46–61). Springer.

Yang, C.C., & Li, K.W. (2003). Automatic construction of English/Chinese parallel corpora. Journal of the American Society for Information Science and Technology, 54(8), 730–742.

Yang, C.C., & Wang, F.L. (2003a). Fractal summarization for mobile device to access large documents on the web. In Proceedings of the 12th International Conference on World Wide Web (WWW 2003) (pp. 215–224). New York: ACM.

Yang, C.C., & Wang, F.L. (2003b). Fractal summarization: Summarization based on fractal theory. In Proceedings of the 26th Annual International ACM SIGIR Conference: Research and Development in Information Retrieval (SIGIR 2003) (pp. 391–392). New York: ACM.

DOI: 10.1002/asi