ELSEVIER

Information Processing and Management 40 (2004) 161–190

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Chinese word segmentation and its effect on information retrieval

Schubert Foo [*], Hui Li

Division of Information Studies, School of Communication and Information, Nanyang Technology University, Nanyang Link, Singapore 637718

Received 28 February 2002; accepted 5 August 2002

Abstract

A set of IR experiments was carried out to study the impact of Chinese word segmentation and its effect on information retrieval (IR) at the Division of Information Studies, Nanyang Technological University, Singapore. A total of four automatic character-based segmentation approaches and a manual word segmentation approach was first carried out to obtain the word segments for indexing and to evaluate the segmentation accuracy of these automatic approaches. The IR experiments study both the influence of different document segmentation approaches on IR effectiveness and the methods used for query segmentation. Traditional data recall and precision measures were used to gauge IR effectiveness. A number of queries were selected and subjected to further detailed analysis to further explore the influence of word segmentation on IR.

The findings reveal that the segmentation approach has an effect on IR effectiveness. Better IR results are obtained by using the same method for query and document processing as this increase the probability of the query-document match. The recognition of a higher number of 2-character words generally contributes to the improvement of IR effectiveness. However, manual segmentation does not always work better than character-based segmentation as a result of the existence of longer words with more than two characters. No evidence is found that ambiguous words resulting from the segmentation process significantly affect IR.

Keywords: Chinese; Information retrieval; Word segmentation; Retrieval effectiveness

[*] Corresponding author. Tel.: +65-6790-4621; fax: +65-6791-5214.
E-mail addresses: assfoo@ntu.edu.sg (S. Foo), lihui_liusg@yahoo.com (H. Li).

0306-4573/$ - see front matter © 2002 Elsevier Ltd. All rights reserved.
doi:10.1016/S0306-4573(02)00079-1

1. Introduction

Research interest in Chinese information retrieval (CIR) has increased as a result of the large growth rate of online Chinese literature. Typically, an IR system determines the relevant documents according to the frequency of occurrence of the words of a query within the documents and corpus (Nie, Brisebois, & Ren, 1996). For English and other western languages, the identification of distinct words in the documents is trivial. However, this is much more difficult for the Chinese language, as well as many non-English languages, since Chinese text appears as a string of ideographic characters without any obvious boundary between words except for punctuation signs at the end of each sentence, and occasional commas within sentences.

Chinese text information processing therefore undergoes an essential segmentation process to break up the text into smaller linguistic units or segments, normally words (Nie et al., 1996; Wu & Tseng, 1993, 1995). These are subsequently used to create the index for query and retrieval operations. Numerous different segmentation approaches have been proposed for CIR. As the review of the related literature will show, these approaches can be basically divided into character-based and word-based approaches. Under these two basic groups, there are many alternatives, such as single-character or multiple-character segmentation, use of dictionary or statistics, or introducing linguistic knowledge for segmentation.

When applied to the information retrieval (IR) problem, the existing literature on CIR studies have been consistently shown that the IR result using single-character indexing is significantly worse than those using other segmentation techniques (Tong, Zai, Milic-Frayling, & Evans, 1996). However, no firm conclusions or agreements on the performance of multi-character approaches and word-based approaches have so far been reported. Some researchers obtained better results using bigram (2-character) methods while the others obtained better results using word-based approaches (Wilkinson, 1997). Therefore, some researchers believe that a better segmentation approach will be able to yield superior IR results (Nie et al., 1996), while others have not found any direct relationship between the segmentation approach and IR results from their experimental results (Kwok, 1997a,b). It is also evident that there has been no such systematic study that is been carried out to investigate this relationship.

Thus, this research aims to systematically investigate the relationship between the segmentation accuracy and its effect on CIR. Manual segmentation along with four types of automatic character-based segmentation approaches were used to process and index a set of test corpus comprising one month's economic news from the online Chinese People's Daily newspaper. A number of accuracy measures were defined and computed to gauge the quality of the four automatic segmentation approach in comparison with the segmented words arising from manual segmentation which was taken to represent the ideal segmentation case.

Using these approaches, a total of five different indices were therefore created for the IR experiments. A total of 20 queries were used in the experiments. Similarly to the segmentation carried out individual documents of the corpus, the queries were segmented prior to the matching and retrieval process. These were segmented using both the manual and corresponding automatic approaches so that two main groups of IR experiments were actually conducted. This allowed the relationship between query segmentation and document segmentation on IR effectiveness to be assessed. The traditional IR effectiveness measures of data recall and data precision was computed and contrasted against segmentation approaches. Statistical analysis was applied to explore the correlation between the segmentation accuracy and IR effectiveness. In order to probe further, a number of queries were identified and examined in detail to reveal the cause and effect of the retrieval results thereby providing a more thorough understanding of the retrieved results. From all these experimental data results, it became possible to derive a set of conclusions.

The rest of the paper is organised as follows. Following a review of related literature, the methodology used for the study is presented. This describes the automatic approaches that were used to create the index and query segmentation for the IR experiments. A number of accuracy measures were defined and computed for these different approaches. The setting of the IR experiments and results from the experiments are subsequently reported using the measures of data recall and precision. A number of queries were used and analysed in detail to aid the explanation of the results arising from different document segmentation approaches, query segmentation approaches, and the effect of the existence of ambiguous word segments. The paper concludes with a summary of the pertinent findings and suggestions for future work.

2. Review of related literature

2.1. Chinese segmentation for IR

The basic approaches of Chinese segmentation can be roughly divided into two groups, namely, character-based approaches and word-based approaches as shown in Fig. 1.

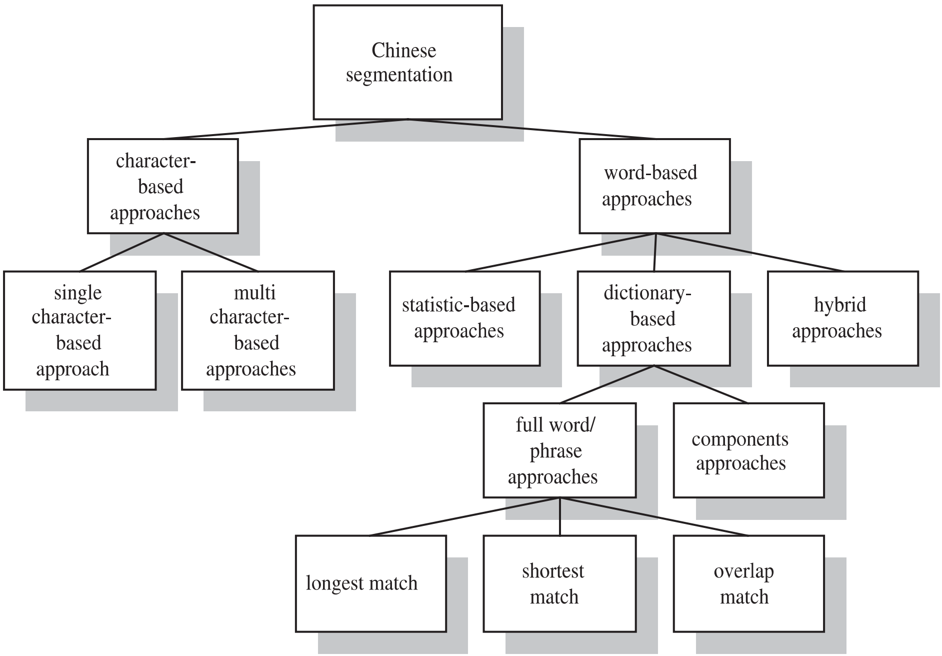

Chinese segmentation
character-based approaches
word-based approaches
single character-based approach
multi character-based approaches
statistic-based approaches
dictionary-based approaches
hybrid approaches
full word/phrase approaches
components approaches
longest match
shortest match
overlap match

Fig. 1. Basic approaches of Chinese segmentation.

2.1.1. Character-based approaches

Character-based approaches can be defined as purely mechanical processes that extract certain number of characters (to form a string) from texts. According to the number of characters extracted, character-based approaches can be further divided into single character-based approach and multi-character-based approaches.

Single character-based approach divides Chinese texts into single characters and is the simplest method to segment Chinese text. The majority of today's CIR systems generally do not employ single character-based approach as the main segmentation approach although some research groups have obtained encouraging results by only using this approach (e.g. Huang & Robertson, 1997; Nie, Chevallet, & Bruandet, 1997; Smeaton & Wilkinson, 1996—refer to Table 1).

Multi-character-based (or *N*-gram) approaches segment texts into strings containing two (bigram), three or more characters. Compared with the single character approach, the multi-character approaches consistently yield superior CIR results (Kwok, 1997a,b). Since 75% of all available and commonly used Chinese words are made up of two characters (Wu & Tseng, 1993), a popular approach is bigram approach that segments a linear Chinese sentence ABCDEF into AB, CD, EF and generates most of the correct Chinese words in a piece of text. A variation is the overlapping bigram that segments a sequence of Chinese sentence ABCDEFG into AB BC CD DE EF FG. Similar results for using these two approaches for CIR have been reported (Tong et al., 1996). Another group of researchers from the Swiss Federal Institute of Technology (ETH) implemented another variant of bigram approach by introducing a stoplist to partially segment sentences before applying the bigram segmentation to create the index (Mateev, Munteanu, Sheridan, Wechsler, & Schuble, 1997). In doing so, the size of the vocabulary was reduced significantly but they found that the IR performance suffered no loss compared with pure bigram approach.

In applying these character-based approaches to IR, the most obviously advantage is its simplicity and ease of application. This in turn lead to other advantages of reduced costs and minimal overheads in the indexing and querying process. As such, multi-character-based approaches, especially the bigram approaches, have been found to be practical options that are implemented in many CIR systems.

2.1.2. Word-based approaches

Word-based approaches, as the name implies, attempts to extract complete words from sentences. They can be further categorised as statistics-based, dictionary-based and hybrid approaches.

Statistics-based approaches rely on statistical information such as word and character occurrence frequencies in a set of preliminary training data. As such, this approach is significantly dependent on a training corpus so that the index terms produced are thus more sensitive and useful for particular fields or subject areas that are similar to the training corpus (Nie et al., 1996). In practice, purely statistical approaches have not been very popular. Work that have reported using such an approach include those of Sproat and Shih (1990, 1996) and the University of Massachusetts (Allan et al., 1996; Smeaton & Wilkinson, 1996). In the case of the latter, this approach was applied to query segmentation but not for document segmentation.

Dictionary-based approach is commonly used in most current systems utilising the word-based approach for text segmentation. In this approach, segmented texts are matched against a

Table 1
Comparison of segmentation approaches in IR experiments

Research team	Document segmentation	Linguistic knowledge	Query segmentation	Superior approach
Royal Melbourne Institute of Technology (Royal Melbourne Institute of Technology, 1996, 1997; Smeaton & Wilkinson, 1996; Wilkinson, 1997)	Character-based (single character, overlapping bigram) Word-based (longest match)		Character-based	Word-based combination best
University of Massachusetts (Allan et al., 1996, 1997; Smeaton & Wilkinson, 1996; UM, 1996, 1997; Wilkinson, 1997)	Character-based (single character)		Character-based (single character) Word-based (statistic-based)	Character-based
Information Technology Institute (Information Technology Institute, 1997; Ngo & Lai, 1996; Wilkinson, 1997)	Character-based			
CLARITECH Corporation (Claritech Corporation, 1996; Smeaton & Wilkinson, 1996; Tong et al., 1996)	Character-based (overlapping bigram) Word-based (dictionary-based approaches/heuristics)	Simple morphological rules, and heuristics about people's names	Same as document process	Word-based
City University (Beaulieu et al., 1996; City University, 1996; Huang & Robertson, 1997; Smeaton & Wilkinson, 1996)	Character-based (single character) Word-based (longest match)		Word-based	Character-based
George Mason University (George Manson University, 1996; Smeaton & Wilkinson, 1996)	Character-based (single character)			
University of California, Berkeley (He, Xu, Chen Meggs, & Gey, 1996; Smeaton & Wilkinson, 1996; University of California, Berkeley, 1996)	Word-based (longest match/overlap match)			
Cornell University (Cornell University, 1996; Smeaton & Wilkinson, 1996)	Character-based (bigram)			

(*continued on next page*)

Table 1 (*continued*)

Research team	Document segmentation	Linguistic knowledge	Query segmentation	Superior approach
Queens College, CUNY (Kwok, 1997a,b; Kwok & Grunfeld, 1996; Cuny, 1996; Smeaton & Wilkinson, 1996)	Character-based (single character, bigram) Word-based (longest match/statistic-based)	Some common language usage rules	Same as document process	Bigram performs as well as word-based combination best
Institute of Systems Science (Leong & Zhou, 1997; Wilkinson, 1997)	Character-based (bigram) Word-based (longest match)			Character-based
Swiss Federal Institute of Technology (ETH) (Wilkinson, 1997)	Character-based (Bigram)			
University of Montreal (Nie et al., 1997; Wilkinson, 1997)	Character-based (overlapping bigram) Word-based (compound words and short words)			Word-based

dictionary prior to being indexed. Different approaches use different types of dictionary. Overall, they can be identified as full word or phrase approach using a complete dictionary, and component approach using a component dictionary (Wu & Tseng, 1993). The complete dictionary contains all possible words and phrases used in Chinese texts while the component dictionary only stores word and phrase components such as morphemes and simple words.

According to the choice of match, the full word or phrase approach can be further divided into longest match (by scanning the text sequentially to match the dictionary and choosing the longest strings as index term) and shortest match (by scanning the text sequentially to match the dictionary and choosing the first matched word as index term). In addition to the two normal patterns of longest and shortest match, a third pattern using an overlap match has been proposed (He et al., 1996). In this approach, tokens generated from texts can overlap each other across the matching boundary. As a common approach, the longest match is used by most researchers (for example, Beaulieu et al., 1996; He et al., 1996; Jie, Liu, & Liang, 1991; Leung & Kan, 1996; Witten, Moffat, & Bell, 1994). At the same time, the longest match has been shown to be the most important and effective method in dictionary-based approaches (Jie et al., 1991; Leung & Kan, 1996; Wu & Tseng, 1993).

Dictionary-based approaches do have drawbacks in segmentation. For example, full word and phrase approach relies on the existence of a complete dictionary. In practice, however, it is unfeasible and unnecessary to have a really complete dictionary containing all possible words used in Chinese texts. Although the component approach have been developed for the purpose of alleviating this drawback, the word and phrase components need to be combined into complete words and phrases again using either non-linguistic or linguistic techniques (Wang, Su, & Mo, 1990; Wu & Tseng, 1993, 1995). As such, this approach is less convenient to implement. The automatic Chinese text automation system (ACTS) is an example of such an approach (Wu & Tseng, 1995).

2.1.3. Hybrid approaches

Hybrid approaches are aimed at combining different approaches and taking into account the strength of various techniques. They usually combine statistic-based and dictionary-based approaches in an attempt to merge the benefits of general and domain-specific knowledge. It successful use had been reported by a number of researchers (Kwok, 1997a; Nie et al., 1996; Smeaton & Wilkinson, 1996; Nie, Gao, Zhang, & Zhou, 2000). Although the hybrid approaches take advantage of different approaches to obtain more accurate segments, these are achieved against the expense of more complex processing time, disk space and cost requirements.

2.2. Outstanding problems in Chinese segmentation

Although numerous approaches have been proposed and attempted over the years, no one single approach has been adopted as the de-facto standard for IR operations. In essence, each approach may work effectively in some cases but work poorly in others. Two segmentation problems remain largely unresolved. These are associated with the resolution of ambiguous segments and unknown words resulting from the segmentation process. As the name implies, ambiguous segments are resulting segments that are ambiguous or incorrect in the context of the sentence been segmented. Two forms of ambiguity are possible (Dai, 1997; Liu, 1994; He et al., 1996). The first is the result of an inter-cross ambiguity string (ABC can be segmented into AB|C or A|BC) and the second is the result of a combination ambiguity string (AB can be a word, A can be a word and B can be a word).

Unknown words can arise since existing dictionaries cannot possibly exhaustively cover all the words used in Chinese text as new words and constantly produced and new occurrences of proper nouns (e.g. names of people, organisations, places) cannot be identified by dictionaries. Moreover, it is likely to be cost ineffective to support such a huge dictionary. Nie's experiments have shown that existence of unknown words have adversely affected IR effectiveness (Nie et al., 1996). In his subsequent research, an attempt was made to use a NLP analyzer to recognize such unknown words and add these to the index. This resulted in a marginal increase in precision (Nie et al., 2000).

2.3. Query segmentation and document segmentation for IR

In CIR, segmentation is applied for both documents and user queries. Documents are first segmented to yield the index terms that are stored for the subsequent query matching process. User queries are also segmented prior to matching. Document segmentation is generally carried out automatically while query segmentation can be carried out differently in the IR experiments. These include automatic methods where queries are derived completely automatically from the supplied topics, and manual methods that include queries generated by all other methods that are non-automatic methods (Voorhees & Harman, 1996).

As such, the approach used for document and query segmentation does not necessarily need be the same (e.g. Allan et al., 1997; Huang & Robertson, 1997) although it is more common to use an identical segmentation approach for both document and query segmentation (e.g. Kwok & Grunfeld, 1996; Nie et al., 1997; Tong et al., 1996).

Among previous reported researches, only two groups of researchers conducted experiments to investigate the influence of using different types of segmentation for the query and retrieval process. In the experiment by Allan and his colleagues, documents were segmented using the single character approach, while queries were segmented based on single character and statistical approaches. Although they reported better results brought by single character approach on query segmentation, they did not provide a detailed analysis or offer an explanation on the results (Allan et al., 1997). In another experiment, Huang and Robertson implemented nine algorithms generated from six kinds of query segmentation approaches and two kinds of document segmentation approaches. However, they only reported two of those nine methods, where the query process is word-based and the document processes are both character-based and word-based (Huang & Robertson, 1997). Both methods were shown to yield almost identical results.

2.4. Evaluation and analysis of the segmentation approaches in IR experiments

In contrast to the lack of a standard test for reporting word segmentation performance, a "standard" IR test exists with the introduction of Chinese document retrieval in the fifth Text Retrieval Conference (TREC-5) in 1996 and subsequently in TREC-6 in 1997. The same document collection was used in both these conferences. They comprise a total of 164,811 documents totaling 170 MB in size from the People's Daily and Xinhua News Agency. The original set of 28 queries in TREC-5 was augmented with an additional 26 new queries in TREC-6. Participating research groups can freely devise their own experiments within the TREC tasks and compare their results with a set of relevant documents identified by TREC. This allowed participating researchers to implement various segmentation approaches in their experiments and evaluate the performance of segmentation by using the same queries and retrieval engine while varying the indexing methods in their systems (Leong & Zhou, 1997).

Table 1 shows a representative selection of researchers that participated in TREC-5 and TREC-6 CIR experiments. For the groups that contributed in both conferences, the latest information was used. In the table, the basic approaches used for document and query segmentation are identified whenever possible. Cases where additional linguistic knowledge has been used, and the approach reported to produce better results in cases where more than one segmentation approach was used in the experiment are also highlighted.

Table 1 shows that different research groups have found different preferable segmentation approaches for their systems. Some researchers obtained better results using character-based approaches while others obtained better results using word-based approaches. Although some researchers took into account of the method for query processing, it is inconclusive as to whether or not using the same segmentation methods in both the query and document process produces superior IR results. For example, the University of Massachusetts obtained better results by using the same method for query and document processing. In contrast, Royal Melbourne Institute of

Technology obtained better results by using a character-based approach in query segmentation and word-based approach in document segmentation, while City University obtained better results by using word-based approach in query segmentation and character-based approach in document segmentation.

In TREC-6, an attempt was made by some researchers to combine several retrieval lists generated from segmentation approaches on the assumption that such combination can harness the advantages of individual approaches and therefore improving the final retrieved results. While better results were reported using such a combination strategy by some (e.g. Fuller et al., 1997; Kwok, 1997a,b), others reported on the contrary (e.g. Leong & Zhou, 1997). However, it also became clear that such combinatory approach is too expensive to implement in practice since it requires documents to be indexed several times based on the number of approaches, retrieved several times and subsequently combining individual result sets to yield a final retrieval set.

Following on from these TREC experiments, TREC-9 was held in 2000 at Maryland. TREC-9 contained seven tracks in the form of web retrieval, cross-language retrieval (CLIR), spoken document retrieval, query analysis, question answering, interactive retrieval, and filtering. The TREC-9 CLIR task consisted of bilingual retrieval of Chinese newspaper articles from English queries. There was a total of 25 new topics that were written in English with Chinese translations and contain *title*, *description*, and *narrative* fields. The Chinese corpus used was different from the TREC-5/6, and consisted of about 210 MB of text (or 127, 938 documents) from three Hong Kong newspapers (Hong Kong Commercial Daily, Hong Kong Daily News, and Takungpao) that was encoded in BIG5 as opposed to GB code in the TREC-5/6 corpus. Thus, it was a case of Cantonese in TREC-9 and standard Chinese (Mandarin) in TREC 5/6 so that there were differences between vocabularies in these two corpuses. A monolingual Chinese–Chinese run was also permitted.

A total of 12 papers were submitted for this track. Most of these focused on the CLIR aspect of translation of English to Chinese queries using dictionary and other approaches. Overall, it was found that better cross lingual results were obtained when word-based indexing, since these indices are geared towards maximizing cross-lingual performance. A combination of word-based and character-based approaches were also attempted, but these provided mixed results. Some of these papers also reported results that were related to this work, and these are reviewed for completeness.

In BBN Technologies's (Xu & Weischedel, 2000) monolingual test, the corpus was indexed using bigrams and unigrams with stopwords discarded in the process. The results showed that using bigrams and unigrams resulted in a huge improvement in monolingual performance. These results were also better than the cross-lingual performance.

Fudan University (Wu, Huang, Guo, Liu, & Zhang, 2000) applied both bigram character-based and word-based approach to index the corpus for the monolingual run, and found that the word segmented index yielded better results than *n*-gram index. In contrast, IBM (Franz, McCarley, & Zhu, 2000) work which also involved both indexing based on character and word segmentation, found that the character-based results were better than word-based results across both types of Chinese–English translation and monolingually.

As part of Johns Hopkins University HAIRCUT experiments (McNamee, Mayfield, & Piatko, 2000), they assessed the use of 3-grams in a straight-up comparison with bigrams. They found that

3-grams performed appreciably worse than those using bigrams. They observed that this trend seemed to hold both in monolingual retrieval with natural language queries and in bilingual retrieval using word-based 'translations'.

The experiments at Microsoft Research, China (Gao et al., 2000) also reported results on Chinese monolingual retrieval where they created a bigram and unigram index that resulted in comparable IR results to the best performance using their word-based approach. They also combined an index comprising bigrams and words, but found that the results only yielded slight improvements of 2.6% over the uncombined case, but with the space and the time of indexing more than doubled.

In University of California at Berkeley (Chen, Jiang, & Gey, 2000) report, they presented a method that claimed to be efficient and effective as bigram indexing, but produces a much smaller index file than the overlapping bigram by using probability techniques to break up the sentence in bigrams and unigrams there were subsequently used for indexing. Their experiments showed that the Chinese cross-language retrieval performance using bilingual dictionaries is only about 57% of the monolingual retrieval performance.

2.5. Summary of related literature

Based on the review of related work, it is evident that much work and considerable progress has been made in the area of segmentation and CIR research. Nonetheless, many clear observations become apparent and unresolved problems remain:

- Different applications of segmentation approaches and different interpretation of them by various researchers have given rise to a varied set of results that are not comparable to each other. In the absence of a set of standard segmentation accuracy performance metrics, character-based approaches have been reported to outperform word-based approaches by some researchers, while is reverse is true for others. Due to the nature of the Chinese language, two main segmentation problems remains largely unresolved, namely, segmentation ambiguity and unknown words. Its impact of these on CIR is not largely been systematically evaluated.
- Different segmentation approaches have their own advantages and disadvantages. No general agreement is reached among different research groups as to which segmentation approach is better or more appropriate for the purpose of IR. However, many researchers have shown that the use of character-based approaches (especially the bigram approach) yields comparative results when compared with word-based approaches. Others advocate the use of word-based approach citing the disadvantage of indexing time and space.
- It is generally not reported how query segmentation was carried in relation to document segmentation. Thus, some researchers have used identical approaches for both query and document segmentation while others may have used different approaches simultaneously.
- Most previous researches evaluated the accuracy of segmentation and the performance of IR separately. So far, little has been done to co-relate these two sets of results. Furthermore, the majority of work does not go beyond the reported results in carrying out a more detailed and thorough analysis of identifying the factors in word segmentation that affects the IR results.

3. Segmentation experiments

The following sections report on the research to systematically investigate the relationship between segmentation accuracy and its effect on CIR.

3.1. Test corpus

A whole month of March 1998 news from the economics section of the on-line Chinese People's Daily newspaper was used as the test corpus for this study. A total of 266 files that contain 182,292 Chinese characters and occupy 532 KB of computer space was used. Each file varied between 1 KB to 13 KB in size.

The character-based bigram approach was chosen as the basic automatic segmentation approach. During segmentation, both English words and other non-Chinese strings including Arabic numbers within the Chinese text are treated as segment units. Variations of the bigram approach that were tested included the

- Pure bigram method (abbreviated as PBI) that segments a typical sentence ABCDEF into AB, CD and EF.
- Overlapping bigram method (abbreviated as OVLAP) that segments a typical sentence ABCDEF into AB, BC, CD, DE and EF.
- Pure bigram method combined with 1-character word list (abbreviated as PSTOP) that segments a typical sentence ABCDEF into AB, C, DE and F if C is in 1-character word list.
- Overlapping bigram method combined with 1-character word list (abbreviated as OVSTOP) that segments a typical sentence ABCDEF into AB, C, DE and EF if C is in 1-character word list.

The 1-character word list used in segmentation experiments is constructed according to the statistics of 10 samples extracted from the whole corpus. It is based on the observation that many 1-character words can be used as the natural word boundaries of Chinese texts. A process to generate such a word list was proposed and verified (Foo & Li, 1998). This 1-character word list is also used as a stoplist in the subsequent CIR experiments.

In order to evaluate the accuracy of these segmentation methods, manual segmentation was first carried out on the test corpus. To keep the manual segmentation as correct and consistent as possible and ensure the comparability of results, a proposed set of rules and guidelines for modern Chinese segmentation for information processing was used (Liu, 1994). The manual segmentation was considered not only as a word segmentation technique but also represented an ideal segmentation result set. This was used to compute the accuracy measure for comparing the various segmentation methods.

3.2. Results of manual segmentation

After manual segmentation, the total number of words, and the total number of 1-character, 2-character and 3-character words in each document were computed and totaled as shown in Table 2. In the computation, word repetitions were also counted as distinct words. However, Arabic

Table 2
Statistic result of manual segmentation

	MN	MN1	MN2	MNm
Raw data	108 697	45 790	55 062	7845
Percentage (%)		42.13	50.66	7.21

MN: total number of words in test corpus, MN1: total number of 1-character words, MN2: total number of 2-character words, MNm: total number of words that are 3 characters or longer.

numbers, English strings and punctuation marks were treated as segmentation units and excluded from computation.

3.3. Accuracy measure and results of automatic segmentation performance

A segmentation accuracy measure was used to evaluate the performance of the four bigram-based automatic segmentation methods. Ambiguous words resulting from segmentation were identified by a dictionary, which simply incorporated several on-line lexicons without any modification, to match the incorrect segments against the dictionary terms. Since most single characters in Chinese words can be words, the procedure of dictionary look up was only applied to 2-character segments.

The segmentation accuracy was defined as the ratio of the total number of correct segments (identified with the aid of the manual segmentation results) to the total number of segments generated by the automatic segmentation method:

$$\text{Accuracy} = \frac{\text{CNw}}{\text{AN}} \tag{1}$$

Table 3 shows the statistics of the automatic segmentation methods and its associated segmentation accuracy values for all the documents in the test corpus. From the table, it can be seen that:

- The automatic methods were far less accurate than the manual result.
- Among the four automatic segmentation methods, PSTOP approach yielded the highest accuracy, followed by OVSTOP.

Table 3
Segment count of manual and various automatic approaches

Segmentation approach	CNw (%)	CNw1 (%)	CNw2 (%)	AMB (%)	AN	Accuracy (Eq. (1))
PBI	38 193 (39.5)	3931 (4.07)	34 262 (35.43)	10 238 (10.59)	96 700	0.3950
PSTOP	64 477 (57.31)	26 493 (23.55)	37 984 (33.76)	9646 (8.57)	112 520	0.5730
OVLAP	56 444 (34.81)	1382 (0.85)	55 062 (33.96)	17 369 (10.71)	162 160	0.3481
OVSTOP	73 508 (47.48)	22 901 (14.79)	50 607 (32.69)	15 872 (10.25)	154 808	0.4748

AN: total number of segments in automatic segmentation results, CNw: total number of correct segments, CNw1: total number of correct 1-character segments, CNw2: total number of correct 2-character segments, ABM: number of ambiguous segments.

- A large percentage of ambiguous words were evident from the results. As one would expect with the OVLAP and OVSTOP methods, even more ambiguous words were generated with the growth of incorrect segments.
- Although PSTOP is more accurate than OVLAP and OVSTOP, the number of correct 2-character words recognised by it is much less than those in OVLAP and OVSTOP.

4. Information retrieval experiments

4.1. Test corpus and query formulation

The same set of documents used for the segmentation experiments were used for the IR experiments. Based on overall economic subject contents of these documents, a set of queries was first proposed by four native Chinese speakers from the People Republic of China (PRC) who were also graduate research students in Nanyang Technological University (NTU). A total of 41 queries were initially elicited. Queries were expressed either as complete sentences or phrases. A final set of 20 queries was obtained by merging similar queries, removing redundant queries and those with few matching documents. The latter was possible since the researchers had made earlier relevance assessments of these documents. These initial relevance assessments were discarded for the next stage of the relevance judgement phase of the experiment. The average length of the queries is 8–9 Chinese characters that consist of one to four words. These 20 queries form the basis for the IR experiments.

Six volunteers took part in the experiment as relevance assessors. Of the six assessors, four of them were involved with providing the initial queries. All assessors were native Chinese speakers from PRC who were familiar with the domain subject area. Each assessor was given the list of queries from which they chose between two to five queries that were of interest to him or her. Since the size of the test corpus was relatively small with 266 documents, it was possible to make relevance judgements for all documents with respect to each query. Thus, each assessor was asked to read through each document in turn and evaluate it against the selected queries. When the document was deemed relevant to the query, the Document ID that is uniquely assigned to each document, was noted alongside the query. The queries that each assessor judged were different.

4.2. Procedures for query and document indexing

As noted in the review of related works, two methods of query processing were possible. One is to fix the same query segments and vary the indexing methods used in the document indexing process. The other is to use the same matching indexing method for both query and document process. In this research, both methods of query process were attempted to investigate how query segmentation affects IR results. Thus, the IR experiments were divided into two groups. In the first group, queries were segmented manually based on the same set of guidelines as those used in manual segmentation of documents. Stopwords recognised by the 1-character word list (Foo & Li, 1998) were removed from the resultant manual query segments. In the second group, queries are segmented using the same segmentation method as that used for document segmentation. When implementing the PSTOP, OVSTOP approaches, stopwords were also removed from the query segmentation results. Additionally, non-Chinese strings including Arabic numbers and English characters were excluded from index terms.

Using the result of the segmentation experiments, a total of five different indices were created for the IR experiments. This included the word-based manual segmentation method and four automatic character-based bigram segmentation methods outlined previously. When using the manual segmentation approach and PSTOP and OVSTOP to index documents, the 1-character word list was used as stopwords and excluded from the index terms. As in query processing, non-Chinese strings including Arabic numbers and English characters were excluded from index terms.

4.3. Information retrieval system

In order to carry out the IR experiments, the researchers used a modified an English-based IR systems known as mg (managing gigabytes) system that was originally developed by Lane and Witten (Lane, 1997; Witten et al., 1994). The modification was necessary to support CIR (Lim, 1999).

The modified *mg* system comprises two sub-systems, namely, *mgbuild* and *mgquery*. The *mgbuild* system is responsible for compressing and indexing the test corpus, while the *mgquery* system is used to process users' queries. The *mg* system uses a standard vector-space model to calculate the similarity between query and document using the cosine similarity measure and returns a ranked list of documents in decreasing order of their similarity with the query (Fuller et al., 1997; Lane, 1997).

The returned list of results for the queries were subsequently processed along with the results of the relevance assessments and used as the basis to derive the standard data recall and data precision values to gauge the IR effectiveness. The proposed measures were consistent with the measures used in the TREC experiments (Harman, 1993).

4.4. Results of information retrieval

Corresponding to the two different methods of query indexing, two different groups of IR results were obtained. In each group, results were available for all the five methods, namely, PBI, PSTOP, OVLAP, OVSTOP and the manual method.

Table 4 shows a comparison of the detailed data recall and precision values obtained from the IR experiments using manual query segments. Fig. 2 shows the corresponding recall–precision graph for Table 4 to aid the visualisation of the differences in retrieval effectiveness among the different methods.

From Fig. 2, it can be seen that manual segmentation performs best among the methods. OVLAP and OVSTOP were not as good the manual result but better than PBI and PSTOP. Finally, the indexing methods using the 1 character-stop list, OVSTOP and PSTOP, outperformed OVLAP and PBI respectively although the results were very close to each other.

Similarly, Table 5 shows the comparison of the detailed recall and precision values for various retrieval results using the same segmentation approach for both query and document indexing. Fig. 3 shows the corresponding recall–precision graph for Table 5.

From Fig. 3, it can be seen that results of the automatic segmentation methods were comparable if not better than the manual method. The indexing methods using the 1 character-stop list,

Table 4
IR results of various indexing approaches using manual query segments

	Indexing methods				
	PBI	PSTOP	OVLAP	OVSTOP	Manual
No. of queries	20	20	20	20	20
Relevant	214	214	214	214	214
Retrieved	1878	1972	2127	2127	2066
Rel_ret	173	178	189	189	194
Recall level precision averages					
0.0	0.7903	0.7800	0.8518	0.8541	0.9076
0.1	0.7352	0.7256	0.7850	0.7817	0.8826
0.2	0.6370	0.6847	0.7244	0.7357	0.8039
0.3	0.5807	0.5831	0.6905	0.7006	0.7397
0.4	0.5168	0.5437	0.6055	0.6128	0.6989
0.5	0.4889	0.5109	0.5772	0.5770	0.6847
0.6	0.4702	0.4707	0.5244	0.5273	0.5420
0.7	0.4094	0.4138	0.4369	0.4447	0.4464
0.8	0.3236	0.3407	0.3571	0.3562	0.3780
0.9	0.2378	0.2862	0.2823	0.2846	0.3028
1.0	0.2117	0.2198	0.2432	0.2452	0.2458
Non-interpolated average precision					
	0.4451	0.4698	0.5200	0.5115	0.5658
Overall recall					
	0.8084	0. 8318	0. 8832	0.8832	0.9065

Rel_ret: relevant and retrieved, non-interpolated average precision represents the best IR performance that can be achieved (Rijsbergen, 1997; Salton & McGill, 1983), overall recall gives an overall picture of recall performance (Salton & McGill, 1983)—also used in He's TREC experiments (He et al., 1996).

OVSTOP and PSTOP, continued to outperform OVLAP and PBI respectively. PBI appears worst among all methods. The differences for the remaining methods were difficult to judge from the figure unlike Fig. 2.

4.5. Significant test on the IR effectiveness

In order to ascertain the accuracy of the previous observations, a significant test was carried out to provide statistical evidence to confirm if a given difference between two sets of IR results is in fact significant. In the IR community, the paired t-test is commonly used for this purpose (Beaulieu et al., 1996; Hull, 1993).

When the computed probability (i.e. p-value) is small enough (e.g. 0.05), one can conclude that the two sets of sample values are significantly different at the 5% level. The threshold p-value is often set at either 0.05, 0.01 or 0.001. Intuitively, one can think that $p = 0.01$ is more statistically significant than $p = 0.05$ (Motulsky, 1995; Rijsbergen, 1997; Salton & McGill, 1983). In this research, it is assumed when p-values are less than 0.01, the results are statistically significant at the 1% level.

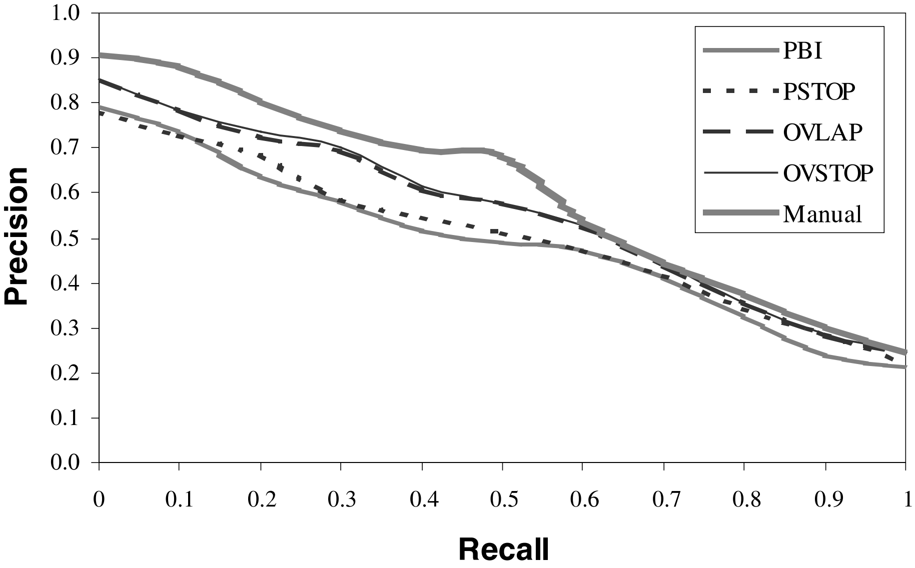


Fig. 2. Recall–precision graph comparing various indexing approaches using manual query segments.

Table 5
IR results of various indexing approaches using matching automatic query segments

	Indexing methods				
	PBI	PSTOP	OVLAP	OVSTOP	Manual
No. of queries	20	20	20	20	20
Relevant	214	214	214	214	214
Retrieved	1965	2147	2546	2339	2066
Rel_ret	175	189	198	198	194
Recall level precision averages					
0.0	0.8656	0.9250	0.8888	0.9000	0.9076
0.1	0.8323	0.9083	0.8788	0.8900	0.8826
0.2	0.8098	0.9000	0.8788	0.8817	0.8039
0.3	0.7716	0.8108	0.8349	0.8255	0.7397
0.4	0.6468	0.7634	0.7659	0.7739	0.6989
0.5	0.5951	0.6625	0.7260	0.7168	0.6847
0.6	0.5286	0.6342	0.6587	0.6698	0.5420
0.7	0.4574	0.5469	0.5937	0.5886	0.4464
0.8	0.3591	0.4518	0.5044	0.4827	0.3780
0.9	0.2726	0.3419	0.4086	0.4045	0.3028
1.0	0.2198	0.2455	0.3264	0.3153	0.2458
Non-interpolated average precision					
	0.5446	0.6311	0.6533	0.6506	0.5658
Overall recall					
	0.8178	0.8832	0.9252	0.9252	0.9065

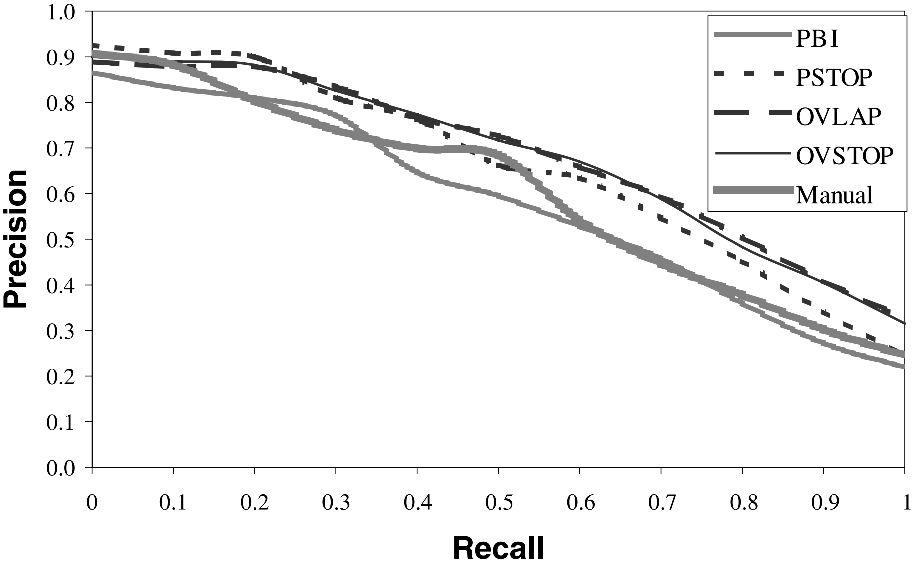


Fig. 3. Recall–precision graph comparing various indexing approaches using matching automatic query segments.

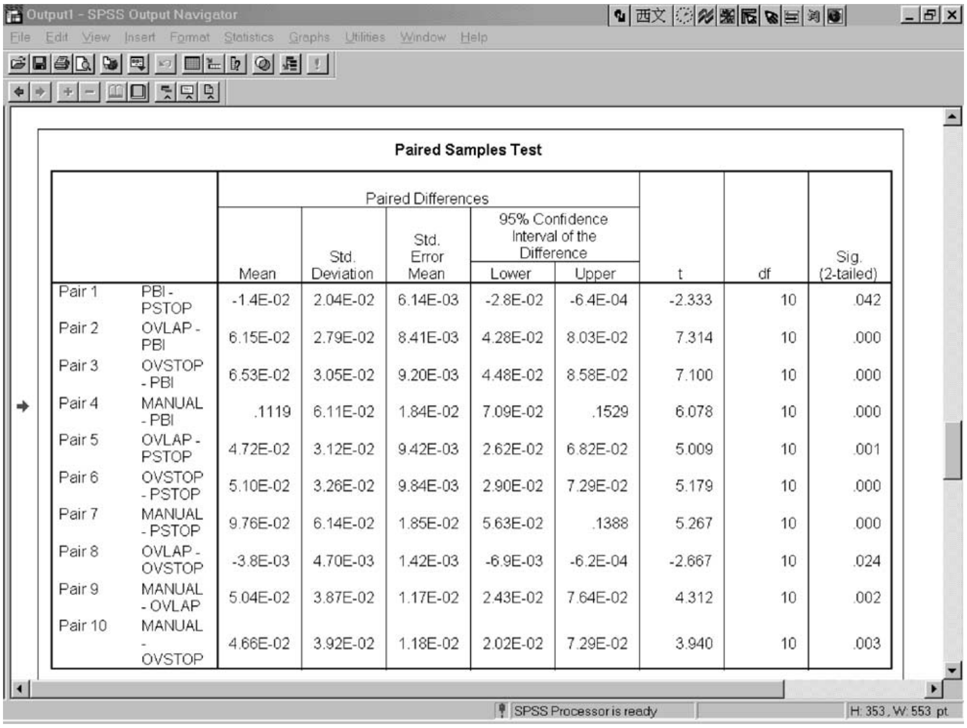
Figs. 4 and 5 show the comparison results of paired *t*-test obtained by using the statistic software SPSS 7.5 for Windows 95/NT (SPSS, 2002) corresponding to the different IR results for the different query process methods. In these figures, the values of *2-tailed Significance* (shown as Sig (2-tailed)) indicate the computed *p*-value.

The statistical results of Figs. 4 and 5 can be used to verify the earlier interpretation of the data recall and precision graphs. For example, in Fig. 4, the *p*-values between the pairs of PBI versus PSTOP (Pair 1), OVLAP versus OVSTOP (Pair 8) are greater than 0.01 but less than 0.05 implying the use of the 1 character word list is significant at the 5% level. All the other pairs, namely, manual pairs versus all automatic methods, OVLAP versus PBI, OVLAP versus PSTOP, OVSTOP versus PBI and OVSTOP versus PSTOP are all statistically significant at 1%. Similar conclusions may be inferred to confirm the observations for Fig. 5.

In terms of the IR effectiveness of using manual query segments, it can therefore be concluded that manual indexing is significantly better than all the automatic indexing methods, and that both the two overlapping bigram methods, OVLAP and OVSTOP, performed significantly better than the pure bigram methods, PBI and PSTOP, at the 1% level.

In terms of the IR effectiveness of using the same segmentation approach in query and document process, it can be concluded that manual indexing is only slightly better than PBI, however, the difference is not so great as expected. Compared with other three automatic approaches, manual segmentation method did not work better, and the differences were statistically significant. PSTOP, OVLAP and OVSTOP performed significantly better than PBI and manual indexing at the 1% level. However, these three methods are not significantly different at 1%.

When comparing the IR effectiveness between the two different query processes, the IR results for manual segmentation results are actually the same since they are common in both query processes. In the case of manual query segments, the same manual segmentation when applied to


Output1 - SPSS Output Navigator
File Edit View Insert Format Statistics Graphs Utilities Window Help
西文
SPSS Processor is ready
H: 353 , W: 553 pt

Paired Samples Test

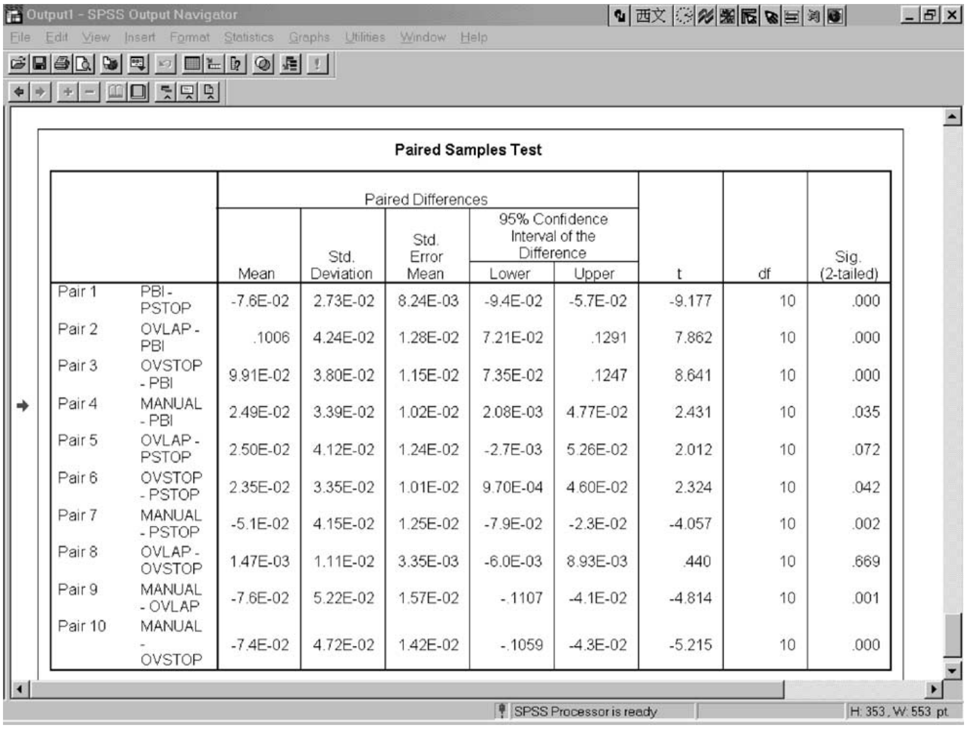
		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	PBI - PSTOP	-1.4E-02	2.04E-02	6.14E-03	-2.8E-02	-6.4E-04	-2.333	10	.042
Pair 2	OVLAP - PBI	6.15E-02	2.79E-02	8.41E-03	4.28E-02	8.03E-02	7.314	10	.000
Pair 3	OVSTOP - PBI	6.53E-02	3.05E-02	9.20E-03	4.48E-02	8.58E-02	7.100	10	.000
Pair 4	MANUAL - PBI	.1119	6.11E-02	1.84E-02	7.09E-02	.1529	6.078	10	.000
Pair 5	OVLAP - PSTOP	4.72E-02	3.12E-02	9.42E-03	2.62E-02	6.82E-02	5.009	10	.001
Pair 6	OVSTOP - PSTOP	5.10E-02	3.26E-02	9.84E-03	2.90E-02	7.29E-02	5.179	10	.000
Pair 7	MANUAL - PSTOP	9.76E-02	6.14E-02	1.85E-02	5.63E-02	.1388	5.267	10	.000
Pair 8	OVLAP - OVSTOP	-3.8E-03	4.70E-03	1.42E-03	-6.9E-03	-6.2E-04	-2.667	10	.024
Pair 9	MANUAL - OVLAP	5.04E-02	3.87E-02	1.17E-02	2.43E-02	7.64E-02	4.312	10	.002
Pair 10	MANUAL - OVSTOP	4.66E-02	3.92E-02	1.18E-02	2.02E-02	7.29E-02	3.940	10	.003

Fig. 4. SPSS output of paired samples *t*-test on IR results of using manual query segments.

documents yields the best performance over all other automatic approaches. In contrast, when the same segmentation method is used for both query and document process, three of the automatic segmentation methods (PSTOP, OVLAP, OVSTOP) yield better results than the manual one. Even the worst PBI is competitive to the manual result. From here, it becomes evident that for all automatic segmentation methods, using the same method for both the query and document process can lead to better IR effectiveness. As such, the highly laborious process of manual segmentation does not lead to superior IR performance, rather, an appropriate automatic segmentation approach (for both matching query and document processing) would be the appropriate choice for the development of CIR systems.

5. Analysis and discussion of results

In this research, an attempt was made to carry out an in-depth analysis of a number of results with the aim to investigate how segmentation results account for differences in IR performance. The analysis was restricted to a number of cases where the results demonstrate significant difference according to the paired *t*-test result obtained from average IR performance. At the same time, a new set of paired *t*-test is conducted on individual queries to identify queries that exhibited significant different IR results. The analysis is divided into two parts following the division of the experiments, namely, using manual segmented queries and automatic segmented queries.


Output1 - SPSS Output Navigator
File Edit View Insert Format Statistics Graphs Utilities Window Help
SPSS Processor is ready
H: 353 , W: 553 pt

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	PBI - PSTOP	-7.6E-02	2.73E-02	8.24E-03	-9.4E-02	-5.7E-02	-9.177	10	.000
Pair 2	OVLAP - PBI	.1006	4.24E-02	1.28E-02	7.21E-02	.1291	7.862	10	.000
Pair 3	OVSTOP - PBI	9.91E-02	3.80E-02	1.15E-02	7.35E-02	.1247	8.641	10	.000
Pair 4	MANUAL - PBI	2.49E-02	3.39E-02	1.02E-02	2.08E-03	4.77E-02	2.431	10	.035
Pair 5	OVLAP - PSTOP	2.50E-02	4.12E-02	1.24E-02	-2.7E-03	5.26E-02	2.012	10	.072
Pair 6	OVSTOP - PSTOP	2.35E-02	3.35E-02	1.01E-02	9.70E-04	4.60E-02	2.324	10	.042
Pair 7	MANUAL - PSTOP	-5.1E-02	4.15E-02	1.25E-02	-7.9E-02	-2.3E-02	-4.057	10	.002
Pair 8	OVLAP - OVSTOP	1.47E-03	1.11E-02	3.35E-03	-6.0E-03	8.93E-03	.440	10	.669
Pair 9	MANUAL - OVLAP	-7.6E-02	5.22E-02	1.57E-02	-.1107	-4.1E-02	-4.814	10	.001
Pair 10	MANUAL - OVSTOP	-7.4E-02	4.72E-02	1.42E-02	-.1059	-4.3E-02	-5.215	10	.000

Fig. 5. SPSS output of paired samples *t*-test on IR results of using the same segmentation method in query and document process.

5.1. Using manual query segments

Generally speaking, when using the manual query segments, the similarity of query and document becomes largely dependent on whether the multi-character words in the query can be correctly recognised in the document. This is because the manual query segments for retrieval are basically multi-character words (including 2-character and longer words) with minimum 1-character words. The segmentation result recognising more multi-character words does not only increase the similarity between query and document thereby improving precision, but also increase the chance of being retrieved thereby improving recall.

In the first group of in-depth analysis experiments, comparisons were conducted among the results of the manual, PSTOP and OVSTOP approaches that represent the better runs of the different distribution level as shown in Fig. 2. In Fig. 2, the results of five approaches roughly fall into three levels of distribution. The manual approach represents the best result. The results of the OVSTOP and OVLAP approaches were in the middle, while the results of PBI and PSTOP were in the bottom. In the two lower levels, OVSTOP and PSTOP represent the better set of results. Based on a set of paired *t*-test on individual queries, two queries were selected for analysis: Query 7 (Q7) where the manual result was significantly better than PSTOP and OVSTOP, and Query 13 (Q13) where OVSTOP was significantly better than PSTOP.

The comparisons among these approaches are shown in Tables 6 and 7 respectively. For each approach, the first column indicates the Doc_ID of the documents that are retrieved and relevant, the second column indicates the total number of documents retrieved, the third column indicates the recall value when each relevant document is retrieved, the fourth column denotes the corresponding precision value and the non-interpolated average precision over all the relevant documents. The last few columns indicate the number of query term matches in the document.

5.1.1. Comparison between manual and automatic approaches

Q7, with a total of 17 relevant documents, was selected for further analysis and comparison between manual and PSTOP, and manual and OVSTOP. From Table 6, it can be seen that DOC 47, DOC 122, DOC 157 and DOC 180 were not retrieved by both PSTOP and OVSTOP since these two approaches cannot recognise the 3-character query term "消费者". However, they can be retrieved by the manual approach thereby yielding a higher recall value. This is been assisted by the fact that the manual approach can correctly identify this phrase with the intended meaning as reflected in the document, so that the relevant documents were identified and ranked higher.

However, for PSTOP and OVSTOP, documents were also retrieved because of the more general terms "保护" and "权益". In this situation, many irrelevant documents were retrieved along the way and the relevant documents might not be ranked before such irrelevant documents, thereby adversely affecting the precision. Table 6 shows that the manual approach found all the 17 relevant documents after retrieving 22 documents, while PSTOP only found 12 relevant documents after retrieving 33 documents and OVSTOP found 13 relevant documents after retrieving 27 documents. Thus, the manual approach has the higher precision than PSTOP and OVSTOP since with more multi-character words recognised, the possibility of the query matching against documents is correspondingly increased, leading to improvement in the ranking of retrieved relevant documents.

5.1.2. Comparison between automatic approaches

For Q13, the OVSTOP result with a total of 10 relevant documents was significantly superior to the PSTOP result. Since stopwords and non-Chinese strings including Arabic numbers and English characters were not indexed in the query, the manual segmentation result of the original query "3月份的深圳证券交易所股票指数" became "月份/深圳/证券交易所/股票/指数". Furthermore, words longer than two characters cannot be retrieved by PSTOP and OVSTOP approaches. Thus, the comparison is only focused on the resulting 2-character query terms as shown in Table 7.

As shown in Table 7, OVSTOP retrieved 5 of the 10 relevant documents for Q13 while PSTOP only retrieved 1 relevant document. On a closer examination of the segmentation results obtained from these two approaches, it was found that DOC 39, DOC 84, DOC 117 and DOC 155 cannot be retrieved by PSTOP because the query term "指数" was not correctly segmented. In this situation, the recall of PSTOP for Q13 was much lower than OVSTOP. This is so since the nature of the overlapping bigram in OVSTOP ensured that the query term "指数" is detected. Additionally, PSTOP yielded a lower precision value than OVSTOP since it only retrieved 1 relevant document after retrieving 9 documents while OVSTOP retrieved 5 relevant documents after retrieving 13 documents.

Table 6
Comparison of IR results of manual and automatic approaches (manual vs. PSTOP, manual vs. OVSTOP)

Manual query segment	Doc ID	#Doc retrieved	Recall	Precision	# Query term1 (保护)	# Query term2 (消费者)	# Query term3 (权益)
Q7	*Manual*						
保护/	46	1	0.0588	1.0000	8	9	3
消费者/	234	2	0.1176	1.0000	3	11	4
权益	119	3	0.1765	1.0000	4	10	3
	104	4	0.2353	1.0000	1	8	4
	118	5	0.2941	1.0000	9	17	3
	9	6	0.3529	1.0000	3	6	1
	182	7	0.4118	1.0000	2	3	2
	122[a,b]	8	0.4706	1.0000	0	9	0
	7[a]	9	0.5294	1.0000	1	6	0
	120	10	0.5882	1.0000	2	8	0
	44	11	0.6471	1.0000	1	4	0
	185	14	0.7059	0.8571	1	8	0
	204	15	0.7647	0.8667	0	5	1
	47[a,b]	16	0.8235	0.8750	0	2	0
	157[a,b]	17	0.8824	0.8824	0	4	0
	180[a,b]	21	0.9412	0.7619	0	2	0
	85	22	1.0000	0.7727	2	2	0
				AVE precision			
				0.9421			
	PSTOP						
	46	1	0.0588	1.0000	8	0	0
	234	3	0.1176	0.6667	3	0	3
	118	4	0.1765	0.7500	7	0	2
	119	5	0.2353	0.8000	3	0	1
	182	7	0.2941	0.7143	4	0	0
	9	8	0.3529	0.7500	3	0	0
	204	11	0.4118	0.6364	0	0	1
	104	15	0.4706	0.5333	1	0	0
	120	17	0.5294	0.5294	2	0	0
	44	23	0.5882	0.4348	1	0	0
	85	31	0.6471	0.3548	1	0	0
	185	33	0.7059	0.3636	1	0	0
				AVE precision			
				0.4431			
	OVSTOP						
	46	1	0.0588	1.0000	8	0	3
	234	3	0.1176	0.6667	3	0	4
	119	4	0.1765	0.7500	4	0	3
	104	5	0.2353	0.8000	1	0	4
	182	6	0.2941	0.8333	5	0	2
	118	7	0.3529	0.8571	9	0	3
	9	9	0.4118	0.7778	3	0	1

(*continued on next page*)

Table 6 (*continued*)

Manual query segment	Doc ID	#Doc retrieved	Recall	Precision	# Query term1 (保护)	# Query term2 (消费者)	# Query term3 (权益)
	204	17	0.4706	0.4706	0	0	1
	85	20	0.5294	0.4500	2	0	0
	120	21	0.5882	0.4762	2	0	0
	185	23	0.6471	0.4783	1	0	0
	44	25	0.7059	0.4800	1	0	0
	7	27	0.7647	0.4815	1	0	0
				AVE precision 0.5013			

[a] Not retrieved by PSTOP.
[b] Not retrieved by OVSTOP.

Table 7
Comparison of IR results between automatic approaches (OVSTOP vs. PSTOP)

Manual query segment	Doc ID	#Doc retrieved	Recall	Precision	# Query term1 (月份)	# Query term2 (深圳)	# Query term3 (股票)	# Query term4 (指数)
Q13 月份/ 深圳/ 证券交易所/ 股票/ 指数	*OVSTOP*							
	266	9	0.1000	0.1111	0	0	0	2
	84[a]	10	0.2000	0.2000	0	0	0	2
	39[a]	11	0.3000	0.2727	0	0	0	2
	155[a]	12	0.4000	0.3333	0	0	0	2
	117[a]	13	0.5000	0.3846	0	0	0	2
				AVE precision 0.1302				
	PSTOP							
	266	9	0.1000	0.1111	0	0	0	2
				AVE precision 0.0111				

[a] Not retrieved by PSTOP.

As multi-character words are more meaningful to express linguistic concepts in Chinese documents, the correct recognition of them would have more significant effects on the results. Thus, it is reasonable to conclude that the segmentation approach that recognises the higher amount of correct 2-character words and longer words with more than two characters will yield better IR results when using manual query segments.

5.1.3. Impact of ambiguous words

In evaluating the influence of word segmentation on IR, the role of ambiguous words is also considered. With reference to Table 4 (segment count), it can be seen that in the automatic approaches, OVLAP produces the largest number of ambiguous words while PSTOP has the fewest number of ambiguous words. However, from the experiments, OVLAP was shown to yield better recall and precision values than PSTOP. This indicates that the adverse influence of ambiguous words is not significant enough to eliminate the positive influence of correct recognition of 2-character words even if it is commonly thought that ambiguous words would have a negative impact on IR results.

5.2. Using the same segmentation approach in query and document

When using the same method in the query and document process, IR results are more dependent on the possibility of query-document match than the accurate choice of index terms. The approaches facilitating the match between query and document will improve the IR results. In the experiments, the pure bigram approach and its variants obtained quite good results compared with manual approach. This implies that although the automatic approaches are not as accurate as the manual approach in segmentation, the IR results can be very good due to the same incorrect segments that exist in both document and query segmentation, thereby leading to a perfect match between them.

In the second group of in-depth analysis experiments, the comparison was also selectively conducted among the results that are significantly different. Based on a set of paired *t*-test on individual queries, three queries were selected for analysis. The first set of two queries represents a comparison between the manual and automatic approaches: Query 12 (Q12) where the manual result is significantly worse than PSTOP, and Query 4 (Q4) where the manual result is significantly worse than OVSTOP. The second set of a single query, Query 6 (Q6), was selected to compare the results between automatic approaches, where it was found that the PBI result is significant worse than PSTOP and OVSTOP results.

5.2.1. Comparison between manual and automatic approaches

The comparisons among approaches for Q12 and Q4 are shown in Tables 8 and 9 respectively. For the OVLAP approach in Table 9, only the first five matched query terms are presented in cases where more than five terms are found in the documents.

As shown in Table 8, there was no difference in the recall values of these two approaches. Both two approaches were able to find all the 10 relevant documents. However, the PSTOP approach found 9 relevant documents after retrieving 28 documents while the manual approach needed to retrieve 88 documents to yield the same result. This made the average performance of the manual approach worse than the PSTOP approach. To explain the reason, a close look was taken on DOC 77 and DOC 7, which were ranked in the lowest position in manual result but were retrieved by PSTOP after only retrieving 10 documents. It was found that DOC 77 was retrieved by the manual approach because of the term “市场” that was a general word in Chinese economy news. In the PSTOP result, DOC 77 was retrieved because of the term “房”, which was a more specific word to express the user’s query. DOC 7 was retrieved by the manual only because of the term “中国” which was also a general word while it is retrieved by PSTOP because of both the terms “中国” and “房”. Thus, DOC 77 and DOC 7 yield superior ranking position in the PSTOP result.

Table 8
Comparison of IR results between manual and automatic approaches (manual vs. PSTOP)

Original query	Doc ID	#Doc retrieved	Recall	Precision	Matched query term1	Matched query term2	Matched query term3	
Q12	Manual 中国/房地产/市场							
中 国 的	137	1	0.1000	1.0000	0	3	1	
房 地 产	140	2	0.2000	1.0000	3	6	9	
市场	9	8	0.3000	0.3750	1	1	2	
	247	11	0.4000	0.3636	1	1	1	
	66	12	0.5000	0.4167	20	0	4	
	125	20	0.6000	0.3000	0	1	2	
	175	48	0.7000	0.1458	0	1	1	
	85	86	0.8000	0.0930	3	0	0	
	77	88	0.9000	0.1023	0	0	2	
	7	105	1.0000	0.0952	1	0	0	
				AVE precision 0.3892				
	Doc ID	#Doc retrieved	Recall	Precision	Matched query term1	Matched query term2	Matched query term3	Matched query term4
	PSTOP 中国/房/产市/场							
	137	1	0.1000	1.0000	0	4	0	0
	247	2	0.2000	1.0000	1	4	0	0
	77	4	0.3000	0.7500	0	3	0	0
	140	5	0.4000	0.8000	2	6	0	0
	66	7	0.5000	0.7143	21	3	0	1
	7	10	0.6000	0.6000	1	2	0	0
	9	11	0.7000	0.6364	1	1	0	1
	175	12	0.8000	0.6667	0	1	1	0
	125	28	0.9000	0.3214	0	1	0	0
	85	90	1.0000	0.1111	2	0	0	0
				AVE precision 0.6600				

Examining the manual results of DOC 7 and DOC 77, there was no exact term that matched the query term "房地产" indeed. However, there are many words related to this subject, such as "商品房", "房屋" and "住房". Although the manual approach can produce correct results in query and document segmentation, the query-document match was lost in the retrieval and IR effectiveness was reduced. In contrast, although the segmentation result of PSTOP is not as accurate as the manual approach, the incorrect segments in the query and document happened to form a match in the retrieval.

As shown in Table 9, the manual approach found 9 of the 11 relevant documents while OVLAP found 10 relevant documents thereby yielding a higher recall value for OVLAP. Additionally, the

Table 9
Comparison of IR results between manual and automatic approaches (manual vs. OVLAP)

Original query	Doc ID	#Doc retrieved	Recall	Precision	Matched query term1	Matched query term2	Matched query term3		
Q4	Manual 中国/石油工业/现状/发展								
中国石油工业的现状与发展	130	1	0.0909	1.0000	中国 1	石油工业 4	发展 12		
	136	4	0.1818	0.5000	现状 2	发展 1			
	128	35	0.2727	0.0857	发展 5				
	17	37	0.3636	0.1081	中国 8	发展 23			
	129	66	0.4545	0.0758	中国 1	发展 4			
	134	99	0.5455	0.0606	发展 3				
	132	101	0.6364	0.0693	发展 2				
	135	107	0.7273	0.0748	发展 2				
	131	128	0.8182	0.0703	发展 1				
				AVE precision 0.1859					
	Doc ID	#Doc retrieved	Recall	Precision	Matched query term1	Matched query term2	Matched query term3	Matched query term4	Matched query term5
	OVLAP 中国/国石/石油/油工/工业/业的/的现/现状/状与/与发/发展								
	130	1	0.0909	1.0000	中国 1	国石 5	石油 16	油工 4	工业 5
	128	3	0.1818	0.6667	石油 5	发展 5			
	135	4	0.2727	0.7500	国石 1	石油 5	的现 1	发展 3	
	129	5	0.3636	0.8000	中国 1	国石 2	石油 4	发展 4	
	132	9	0.4545	0.5556	石油 1	油工 1	发展 2		
	136	10	0.5455	0.6000	石油 1	现状 2	发展 1		
	17	11	0.6364	0.6364	中国 10	国石 2	石油 5	工业 5	业的 6
	134	29	0.7273	0.2759	石油 2	发展 3			
	230[a]	41	0.8182	0.2195	石油 1				
	131	129	0.9091	0.0775	业的 1	发展 1			
				AVE precision 0.5074					

[a] Not retrieved by manual.

manual approach retrieved 2 relevant documents after retrieving 30 documents, while the OVLAP approach found 8 relevant documents after retrieving 30 documents. Therefore, the OVLAP approach yielded a higher precision value.

The document found by OVLAP but missed by the manual approach for Q4 is DOC 230. It is retrieved by OVLAP as a result of the query term "石油" that matched the document segmentation result. Although DOC 230 is related to Q4, no term in the document matched the query term "石油工业". This decreased the recall value of the manual approach.

On the other hand, it was found that although the manual approach found 9 relevant documents, only DOC 130 matched the more important query term “石油工业” that indicated the specific information need of the user. For the other 8 relevant documents, they were retrieved mainly because of the terms “中国”, “现状” and “发展”, which were general terms used in Chinese economic news. Although the term “石油” occur in these 8 relevant documents, they cannot match the query term “石油工业”. In this situation, many irrelevant documents were retrieved and ranked before the relevant documents. Therefore, the precision value of the manual approach is reduced. In contrast, all the first 9 of the 10 documents (DOC 130–DOC 230) retrieved by OVLAP approach match with the more specific query term “石油”. This imply that the longest match cannot ensure the flexible match between query and document since queries and documents often describe the same concept with phrases that are not exactly the same. This detailed analysis suggests that OVLAP and PSTOP could be more appropriate to obtain better IR results although some of the matched query terms might not be legal words in context. For example, in DOC 17 there were a total of four matching “石油” terms in the manual result, but five matching “石油” were found in the OVLAP result. The extra term “石油” is actually part of the longer word “石油公司”.

5.2.2. Comparison between automatic approaches

The comparisons among automatic approaches for Q6 are shown in Table 10.

As shown in Table 10, DOC 16, DOC 34 and DOC 144 can be retrieved by both PSTOP and OVLAP but missed by PBI. Since the query terms “中国” and “开展” can be correctly recognised in the query segmentation by all the three approaches the missed documents are mainly due to the different segmentation results of the string “保险业务”. However, by examining the non-retrieved documents, it was found that the majority of words “保险” and “业务” in the documents were segmented correctly by PBI. The number of “保险” and “业务” terms recognised by PBI and PSTOP had no major difference. DOC 16, DOC 34 and DOC 144 were not retrieved by PBI because no term in these documents can be matched with the incorrect query segmentation of “的保”,“险业” and “务的”.

On the other hand, although PBI can recognise “保险” and “业务” in DOC 120, these two more specific terms cannot match with the incorrect segmentation of the query either. DOC 120 was retrieved by PBI only because of the matched term “中国”, which was generally used in Chinese economic news. In this situation, many irrelevant documents were retrieved and could be ranked before the relevant documents, thereby reducing the precision. This observation implies that the PBI approach with a weak ability to correctly segment 2-character words damages the consistency of query terms and document terms, thereby contributing to the inferior IR performance of the PBI approach.

5.2.3. Impact of ambiguous words

When considering the impact of ambiguous words on IR, there is no evidence to show that ambiguous words had obvious adverse impact. Although OVLAP produces the largest number of ambiguous words and manual segmentation is assumed as an ideal approach without any ambiguous words, the experiments show that OVLAP yields significantly better results than the manual approach.

Table 10
Comparison of IR results between automatic approaches (PBI and PSTOP, PBI and OVLAP)

Original query	Doc ID	#Doc retrieved	Recall	Precision	Matched query term1	Matched query term2		
Q6	*PBI* 中国/的/保/险业/务的/开展							
中国 的 保险 业务 的 开展	7	2	0.2000	0.5000	中国 1	险业 1		
	120	92	0.4000	0.0217	中国 1			
				AVE precision 0.1043				
	Doc ID	#Doc retrieved	Recall	Precision	Matched query term1	Matched query term2	Matched query term3	
	PSTOP 中国/保险/业务/开展							
	16[a]	1	0.2000	1.0000	保险 8			
	144[a]	2	0.4000	1.0000	保险 8	业务 3		
	120	4	0.6000	0.7500	中国 1	保险 9	业务 1	
	34[a]	5	0.8000	0.8000	中国 1	保险 4	业务 1	
	7	9	1.000	0.5556	中国 1	保险 4		
				AVE precision 0.8211				
	Doc ID	#Doc retrieved	Recall	Precision	Matched query term1	Matched query term2	Matched query term3	Matched query term4
	OVLAP 中国/国的/的保/保险/险业/业务/务的/的开/开展							
	144[a]	1	0.2000	1.0000	保险 9	险业 1	业务 3	务的 1
	16[a]	2	0.4000	1.0000	保险 8			
	120	3	0.6000	1.0000	中国 1	保险 14	险业 1	业务 2
	7	4	0.8000	1.0000	中国 1	保险 6	险业 2	
	34[a]	5	1.0000	1.0000	中国 1	保险 6	业务 1	
				AVE precision 1.0000				

[a] Not retrieved by PBI.

6. Conclusions and future work

Based on the work on this research, it can be concluded that the segmentation approaches used for document and query processing do have an influence on the IR results although there is no direct relationship between segmentation accuracy and IR results. The following important observations can be concluded from the research:

- The IR performance is affected by both document segmentation and query segmentation. If one of them is kept constant and the other is varied using different approaches, different IR results are obtained.

- The consistency of indexing in both query and document processing seems more important than the higher rate of accuracy in recognising correct word segments. Thus, simple bigram approaches can yield competitive results compared with manual segmentation as long as the same segmentation approach is used for both query and document processing.
- When using manual query segments, the possibility of query-document match is largely determined by how many query terms of multiple characters can be recognised in document segmentation. As a result, the approaches recognising a larger amount of words with 2 and more characters can obtain higher values of precision and recall so that manual approach performs better than all the automatic approaches. Although OVLAP and OVSTOP results contain much more incorrect segments than PBI and PSTOP, they perform better in terms of IR effectiveness since they can recognise more multi-character words in the manual results.
- If the segmentation approach is kept the same for both query and document process, the correct segmentation does not appear essential to improve IR. The experimental results show that the manual approach cannot work better than the automatic approaches. The possible reason is that the segments isolated in manual results are more specific and precise in meaning but are weak in the flexibility of query-document comparison when the same concepts are expressed in different ways. It implies that these same problems would be encountered in other word-based especially dictionary-based approaches. On the other hand, the experiment still indicates that the automatic approaches recognising a higher number of 2-character words have the stronger ability to make query and document segments matched.
- Using simple bigram and its variants indeed produces a large number of ambiguous words. However, there is no evidence to show the ambiguous words have a visible impact on values of precision or recall in our experiments.

Although the research has produced a number of useful results, there are some inherent limitations on the issues of corpus and research methods. In particular, we acknowledge that the small document set has diminished the value of research although we have produced useful results and additional insights into the problems of CIR. As part of our next stage of research, we expect to expand the corpus size and carry out tests using the TREC-5/6 and TREC 9 corpuses to enable benchmarking against published results to be achieved, and to incorporate new hybrid segmentation approaches that are deemed appropriate for CIR.

References

Allan, J., et al. (1996). INQUERY at TREC-5. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Allan, J., et al. (1997). INQUERY does battle with TREC-6. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

Beaulieu, M. et al. (1996). Okapi at TREC-5. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Chen, A., Jiang, H., & Gey, F. (2000). English–Chinese cross-language IR using bilingual dictionaries. Available: http://trec.nist.gov/pubs/trec9/t9_proceedings.html, Maryland.

City University. (1996). Chinese results. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Claritech Corporation. (1996). Chinese results. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Cornell University. (1996). Chinese results. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Cuny (1996). Chinese results. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Dai, Y.B. (1997). Developing a new statistical method for Chinese text segmentation. First year report of Master of Science, School of Applied Science, Nanyang Technological University, Singapore.

Foo, S. B., & Li, H. (1998). An integrated bigram approach with single-character word list for Chinese word segmentation. *TEXT Technology, 8*(4), 17–28.

Franz, M., McCarley, J. S., & Zhu, W.-J. (2000) English–Chinese information retrieval at IBM. Available: http://trec.nist.gov/pubs/trec9/papers/ibm_clir.pdf, Maryland.

Fuller, M., et al. (1997). MDS TREC6 report. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

Gao, J., Xun, E., Zhou, M., Huang, H., Nie, J.-Y., Zhang, J.-Y., & Su, Y. (2000). TREC-9 CLIR experiments at MSRCN. Available: http://trec.nist.gov/pubs/trec9/papers/trec-9.pdf, Maryland.

George Manson University. (1996). Chinese results. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Harman, D. (1993). Overview of the second text retrieval conference. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

He, J., Xu, J., Chen, A., Meggs, J., & Gey, F. C. (1996). Berkeley Chinese information retrieval at TREC-5: Technical report. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Huang, X. J., & Robertson, S. E. (1997). Okapi Chinese text retrieval experiments at TREC-6. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of SIGIR'93* (pp. 329–338), Zurich, USA.

Information Technology Institute. (1997). Chinese results. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

Jie, C. Y., Liu, Y., & Liang, N. Y. (1991). The design and realization Chinese automatic segmenting system CASS. *Journal of Chinese Information Processing, 5*(4), 27–34 (in Chinese).

Kwok, K. L. (1997a). Comparing representations in Chinese information retrieval. Available: http://ir.cs.qc.edu/#publi_.

Kwok, K. L. (1997b). Lexicon effects on Chinese information retrieval. Available: http://ir.cs.qc.edu/#publi_.

Kwok, K. L., & Grunfeld, L. (1996). TREC-5 English and Chinese retrieval experiments using PIRCS. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Lane, D. (1997). MG pages. Available: http://www.mds.rmit.edu.au/mg.

Leong, M. K., & Zhou, H. (1997). Preliminary qualitative analysis of segmented vs bigram indexing in Chinese. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

Leung, C. H., & Kan, W. K. (1996). Parallel Chinese word segmentation algorithm based on maximum matching. *Neural, Parallel and Science Computations, 4*(3), 291–303.

Lim, H. K. (1999). Chinese text retrieval system. *Thesis of Master of Applied Science*, School of Applied Science, Nanyang Technological University, Singapore.

Liu, Y. (1994). *The rules of modern Chinese segmentation for the purpose of information processing and approaches of automatic Chinese segmentation* (pp. 36–63). Beijing: Tsinghua University Press (in Chinese).

Mateev, B., Munteanu, E., Sheridan, P., Wechsler, M., & Schuble, P. (1997). ETH TREC-6: Routing, Chinese, cross-language and spoken document retrieval. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

McNamee, P., Mayfield, J., & Piatko, C. (2000). The HAIRCUT system at TREC-9. Available: http://trec.nist.gov/pubs/trec9/papers/jhuapl.pdf, Maryland.

Motulsky, H. (1995). Intuitive biostatistics. Available: http://www.graphpad.com/www/pvalue.htm.

Ngo, C. W., & Lai, K. F. (1996). Experiments on routing, filtering and Chinese text retrieval in TREC-5. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Nie, J. Y., Brisebois, M., & Ren, X. B. (1996). On Chinese text retrieval. In *Proceedings of SIGIR'96* (pp. 225–233), Zurich, Switzerland.

Nie, J. Y., Chevallet, J. P., & Bruandet, M. F. (1997). Between terms and words for European language IR and between words and bigrams for Chinese IR. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

Nie, J.-Y., Gao, J., Zhang, J., & Zhou, M. (2000). On the use of words and *N*-grams for Chinese information retrieval. In *Proceedings of IRAL2000, Fifth international workshop on information retrieval with Asian languages*, Hong Kong.

Rijsbergen, C. J. (1997). *Information retrieval*. London: Butterworths.

Royal Melbourne Institute of Technology. (1996). Chinese results. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Royal Melbourne Institute of Technology. (1997). Chinese results. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill Book Company.

Smeaton, A., & Wilkinson, R. (1996). Spanish and Chinese document retrieval in TREC-5. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

Sproat, R., & Shih, C. (1990). A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages, 4*(4), 336–351.

Sproat, R., & Shih, C. (1996). A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics, 22*(3), 377–404.

SPSS. (2002). SPSS 11.0. Available: http://www.spss.com/spssbi/spss/.

Tong, X., Zai, C., Milic-Frayling, C., & Evans, D. A. (1996). Experiments on Chinese text indexing—CLARIT TREC-5 Chinese track report. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

University of California, Berkeley (1996). Chinese results. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

University of Massachusetts (1996). Chinese results. Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland.

University of Massachusetts (1997). Chinese results. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

Voorhees, E. M., & Harman, D. (1996). *Overview of the fifth text retrieval conference* (TREC-5). Available: http://trec.nist.gov/pubs/trec5/t5_proceedings.html, Maryland, 1996.

Wang, Y. C., Su, H. J., & Mo, Y. (1990). Automatic processing of Chinese words. *Journal of Chinese Information Processing, 4*(4), 1–11 (in Chinese).

Wilkinson, R. (1997). Chinese document retrieval at TREC-6. Available: http://trec.nist.gov/pubs/trec6/t6_proceedings.html, Maryland.

Witten, I. H., Moffat, A., & Bell, T. C. (1994). *Managing gigabytes: compressing and indexing documents and images*. New York: Van Nostrand Reinhold.

Wu, L., Huang, X.-j., Guo, Y., Liu, B., & Zhang, Y. (2000). FDU at TREC-9: CLIR, Filtering and QA Tasks. Available: http://trec.nist.gov/pubs/trec9/papers/FduT9Report.pdf, Maryland.

Wu, Z. M., & Tseng, G. (1993). Chinese text segmentation for text retrieval: achievements and problems. *Journal of the American Society for Information Science, 44*(9), 532–542.

Wu, Z. M., & Tseng, G. (1995). ACTS: An automatic Chinese text segmentation system for full text retrieval. *Journal of the American Society for Information Science, 46*(2), 83–96.

Xu, J., & Weischedel, R. (2000). TREC-9 cross lingual retrieval at BBN. Available: http://trec.nist.gov/pubs/trec9/papers/bbn-trec9.pdf, Maryland.