On Chinese Text Retrieval

Jian-Yun Nie, Martin Brisebois
Département d'informatique et recherche opérationnelle,
University of Montreal,
BP.6128, succ. Centre-ville,
Montreal, Quebec, H3C 3J7 Canada

Xiaobo Ren
Center for Information Technologies Innovation
1575 Bd. Chomedey,
Laval, Quebec, H7V 2X2 Canada

e-mail: {nie, briseboi, ren}@iro.umontreal.ca

In previous studies, Chinese text retrieval has often been dealt with on the character basis. This approach is not suited to deal with complex queries. We suggest that Chinese text retrieval should work with words instead of characters. The crucial problem is to segment originally continuous Chinese texts into words. In this paper, we first propose a hybrid segmentation approach which unifies the commonly used approaches. The system SMART is then adapted to index the segmented Chinese texts. Finally, we suggest that Chinese text retrieval should move further to include a thesaurus in order to cope with the rich vocabulary of Chinese.

1. Introduction

Typically, an Information retrieval (RI) system determines the relevant documents according to the frequency of occurrences of the *words* of a query within the documents and the corpus (e.g. *tf*idf* weighting method). In Indo-European languages, the identification of words is a trivial task, but in Chinese, it is difficult because there is no separation between words: a sentence is written as a continuous character string such as "计算机已经用于各个领域" (Computers have been used in every area). Thus traditional approaches to RI cannot be directly applied to Chinese.

One solution is to proceed Chinese text retrieval on a character basis, i.e. queries are evaluated using character string matching against documents. This approach has been used in several experimental systems for both Chinese [6] and Japanese text retrieval [8, 17, 18]. However, character-based searching is only appropriate for text retrieval using concepts that may be expressed by a unique character string (e.g. proper names). It is not suited for Chinese text retrieval in general, due to the reasons described below.

SIGIR'96, Zurich, Switzerland©1996 ACM 0-89791-792-8/96/08.$3.50

1) A character-based approach would lead to a great deal of incorrect matching between queries and documents due to the quite free combination of characters in sentences. To take an example, if one wants to retrieve documents about 识别 (recognition), then it is possible to find a document containing the sentence 他认识别人 (he knows other people) by the character-based approach. This latter sentence should be segmented as 他 (he) 认识 (knows) 别人 (other people). We can see that 识别 (recognition) is not a word in this sentence.

2) One can argue that if the query string is long enough, this wrong-matching problem may be reduced. This is the hypothesis implicitly used in [18]. This is true, however, at the price that a complex concept should always be expressed by a fixed character string in both the documents and the queries. This is a too strong constraint to be met. For example, the relatively simple concept such as "Chinese character recognition" may not only be expressed as "汉字识别", but also implied in the expressions such as "汉字的识别" (recognition of Chinese characters), "识别汉字" (recognize Chinese characters), "识别手写汉字" (recognize hand-written Chinese characters), etc. These strings can hardly match the string "汉字识别".

3) In character-based approaches, every character is dealt with in the same way. This makes it difficult to distinguish function words such as prepositions (的,之 - of) from more meaningful words. If a query is expressed as "有关计算机的安装和维护的资料" (documents about the installation and maintenance of computers), then the substrings "有关" (about), "的" (of) and "资料" (documents) will be treated by a character-based searching as if they are as meaningful as "计算机" (computer), "安装" (installation) and "维护" (maintenance).

4) Character-based approaches do not allow us to easily incorporate linguistic knowledge (e.g. synonymy) into the searching process. Chinese language has a rich vocabulary. A concept may often be expressed in multiple ways. For example, "中文" (Chinese language) may equally be expressed as "汉语", "国语" or "华语". The concept "information" may be expressed as "信息", "资讯", "讯息", and so on. In order to achieve a high effectiveness, it is necessary for a system to incorporate a thesaurus which establishes relationships among all these similar words. This incorporation is difficult to achieve with character-based retrieval approaches.

The problems just mentioned can be solved only with a word-based retrieval approach. 1) If words are identified

correctly, word boundaries may prevent the system from matching a word against parts of different words as in the earlier example of "识别" (recognition). 2) For segmented texts and queries, it is easy to filter out non-meaningful words (or function words) by setting up a stop list. 3) Word-based approaches do not require a fixed morphology for concepts as much as for character-based approaches. 4) It is possible to incorporate a thesaurus into a word-based retrieval.

Word-based Chinese text searching has also been suggested by some earlier studies [23, 24]. However, the emphasis of these studies was on the segmentation of Chinese rather than their utilization in IR. No experiments have ever been provided to answer the questions of how segmented Chinese texts may be retrieved and what the influence of the segmentation quality is on the retrieval effectiveness.

In this paper, we will first propose a hybrid approach which unifies most of the approaches used in the past. It is our claim that the hybrid approach is the natural segmentation process used by human beings. The hybrid approach is compared with the two commonly used segmentation approaches using two corpora. Then the SMART system is adapted to index the segmented Chinese texts. Our experiments show that our hybrid approach results in better segmentation than the other approaches, and that the retrieval effectiveness is directly affected by the segmentation quality. Finally, we suggest that Chinese RI should move further to incorporate a thesaurus in order to cope with the rich vocabulary in general Chinese texts.

2. Towards a unifying approach of segmentation

Let us first discuss the crucial problem of Chinese segmentation which aims to set word boundaries in originally continuous sentences. There have been two main groups of approaches to Chinese segmentation: dictionary-based approaches and statistical approaches.

Dictionary-based (also called rule-based) approaches [4, 9-12, 22, 25-28] operate according to a very simple concept: a correct segmentation result should consist of legitimate words (in a restrictive sense, those in a dictionary). In general, however, several legitimate word sequences may be obtained from a Chinese sentence. The maximum-matching (or longest matching) algorithm is often used then to select the word sequence which contains the longest (or equivalently, the fewest) words.

For example, the phrase 中国文学 (Chinese literature), may be segmented into the following five legitimate word sequences:

中国 文学	(China, literature = Chinese literature)
中国 文 学	(China, language, study)
中 国文 学	(middle, Chinese language, study)
中 国 文学	(middle, country, literature)
中 国 文 学	(middle, country, language, study)

The first (correct) sequence is chosen as the result because it is composed of the fewest words.

The above algorithm is often extended by a set of heuristic morphological rules: a character string which is not stored in the dictionary, but may be derived from the heuristic rules, is also a possible word candidate. Typically, heuristic rules are set for identifying words having some common structures such as affix structure (大众化 - popularize/popularization) or nominal pre-determiner structure (一百个人 - hundred people).

On the other hand, statistical approaches [3, 5, 7, 13, 19, 21] rely on statistical information such as word and character (co-)occurrence frequencies in the training data - often a set of manually segmented texts. A simple statistical approach is as follows:

Given a set of manually segmented training texts, the probability of a character string S to be a word is calculated as follows:

$$p(S) = \frac{\text{number of occurrences of } S \text{ being segmented as a word in the training set}}{\text{number of occurrences of } S \text{ in the training set}}$$

Given an input string to be segmented, the best solution is composed of a sequence of potential words S_i such that $\Pi_i p(S_i)$ is the highest.

Although many statistical approaches make use of more complex models (typically first-order Markov models), the principle remains the same, except that the probability for a string to be a word is dependent on a certain context (the preceding characters or words). That is, instead of using $p(S)$, one uses $p(S|w)$ or $p(S|c)$ where w and c are respectively the word or character just before S.

The above two groups of approaches have often been seen as competing one against another because of their different advantages and disadvantages that we summarize as follows:

Statistical approaches are more corpus-dependent than dictionary-based approaches. This is an advantage and a disadvantage at the same time: the corpus-dependency makes the approaches more sensitive to the particularity of the application area, but it also prevents the statistical information to be reusable in other applications.

To estimate probabilities in statistical approaches, the system should be presented with a great deal of training texts that have been segmented manually. The manual segmentation is a very costly operation. In addition, manual segmentation is often inconsistent, that is, the same expression describing a concept may be manually segmented sometimes as a single word, sometimes as two or more words. This inconsistency is due to the absence of a precise definition of 'word' in Chinese. It may greatly affect the performance of statistical approaches.

Aside from the practical problems noted above, a more serious problem concerns the models themselves. Most statistical approaches are limited to first-order Markov models. It has been documented [11] that such first-order models can hardly handle words containing more than two characters. If the first-order Markov model is to be extended to a higher order, however, two other problems may be introduced: 1). The prevalence of many "functional" characters with a particular grammatical function such as prepositions, interrogative/negative markers and conjunctions can cause the statistical data in terms of frequency of occurrence to be unfairly skewed when the model is extended to anything beyond a first-order scheme. 2). Collecting enough data to uniformly extend the model beyond the first-order level is difficult. Several methods have been proposed to address these problems [5, 19, 21]. The fact remains that these approaches cause an increase in

the model's complexity.

On the other hand, dictionary-based approaches have the advantage that the lexical knowledge used corresponds closely to our general knowledge about Chinese words and it is represented in a straightforward way such that human experts can easily verify its correctness. The generality of the knowledge used in these approaches also means that an approach may be reused in a different context without much modification. However, a prerequisite for high-quality results in dictionary-based segmentation is a dictionary which is *complete*. It is unrealistic to suppose that a truly complete Chinese dictionary will be available because of the enormous *size* such a potential dictionary would imply, its *domain dependency* (certain strings may be words in some domains while not in others), and the fact that new words are constantly being produced (the *creative* aspect of language).

The above summary shows that the two kinds of approach are indeed complementary: one relies on general word knowledge, the other on the domain-specific knowledge. Our approach aims to combine them in a single segmentation process such that it can benefit from the general and domain-specific knowledge at the same time.

The combination of the two approaches is also naturally suggested by our own (human) segmentation process in reading. When people segment Chinese texts, both types of knowledge are used. Usually, a correct segmentation may be determined unambiguously by cutting the sentence into usual legitimate words. In some circumstances, however, unusual words or new words may be used. In this case, people usually look into the context (or application area) in order to determine whether an unusual or new string may be a word. Although in human examination of context, syntactic, semantic and pragmatic analyses may be appealed, statistical information about the utilization of words (in the same area) still provides useful indication. Thus, a hybrid approach is a natural way to segment Chinese texts.

The combination of both kinds of knowledge is also feasible. In fact, a dictionary-based approach using longest-matching algorithm may also be seen as a special case of statistical segmentation: We can consider that each dictionary items identified in the input string has an equal probability p ($p<1$). Then the maximum-matching algorithm is equivalent to a statistical approach which chooses the segmentation result of the highest probability. For the earlier example of 中国文学 (Chinese literature), if we assign to each potential word an equal probability p (<1), then the first result 中国 文学 (Chinese literature) will have the highest probability p^2. So, the longest-matching algorithm may be easily seen as, or incorporated into, a statistical approach.

We suggest a hybrid approach based on the following principle: the dictionary is considered as the background knowledge and the statistical information as foreground knowledge. The background knowledge is taken into account by assigning it a *default probability* (p). For a word, if statistical information is available, it is used in priority; otherwise, if it is stored in the dictionary, then it is assigned the default probability.

This combination of the two kinds of knowledge is flexible. By varying the default probability value, we can change the relative importance of the statistical information and the dictionary. In particular, when the default probability value is set to 0, the hybrid approach will not take into account the words stored in the dictionary. Consequently, the hybrid approach becomes the statistical approach. On the other hand, when the default probability value is very high (near 1), the hybrid approach will consider almost exclusively the words stored in the dictionary. Thus we obtain the dictionary-based approach in this case. We see that the hybrid approach can cover a wide range of approaches from the statistical approach to the dictionary-based approach, as illustrated by the following figure:

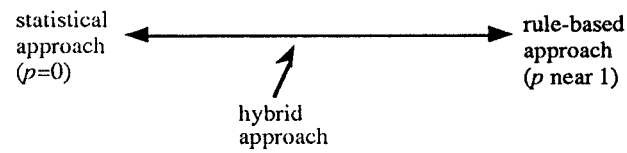

Figure 1. Comparison of the three approaches

3. Implementation

The dictionary used in our system contains over 87 000 entries. A set of heuristic rules is also incorporated to identify and segment words which follow some rules. In our implementation we deal with the following two groups of morphological rules.

Nominal pre-determiner structure:

This structure refer to the strings such as 每一周 (*every week*), 这一回 (*this time*), 每层 (*every layer*), 十一(*eleven*) 一九九一年(*in 1991*), 一百本(*one hundred books*).

We first define the following categories of characters:

- determiners: 这 (*this*), 那 (*that*), 此 (*this*), 该 (*this*), 其 (*its, his, her*), 每 (*each*), 各 (*every*), 某 (*some*), 首 (*first*), 哪 (*which*) ...

- ordinal-number markers: 第 (*number*)

- cardinal numbers: 零 (*zero*), 一 (*one*), 壹 (*one*), 二 (*two*), 贰(*two*), 十(*ten*), 百(*hundred*), 半(*half*) ...

- classifiers: 班(*class*), 帮(*band*), 包(*bag*), 杯(*cup*), 辈 (*generation*), 本(*book*), 遍(*time*), 间(*room*), 层 (*layer*), 年 (year), 月(month), 日(day)...

The rules concerning the formation of complex nominal pre-determiners (pre-det) from these characters are the following (where [...] indicates optional status and [...]* an optional arbitrary repetition):

ordinal cardinal [classifier] → pre-det
e.g. 第一周 (*first week*): ordinal cardinal classifier
第二 (*second*): ordinal cardinal

determiner [cardinal]* classifier → pre-det
e.g. 这一回 (*this time*): determiner cardinal classifier
每层 (*every layer*): determiner classifier

cardinal [cardinal]* [classifier] → pre-det
e.g. 廿一 (*twenty one*): cardinal cardinal
一百本 (*hundred books*): cardinal cardinal classifier

Affix structure:

By affix structure, we refer to the words derived from known words by adding a prefix or a suffix. For example, "小朋友" (little friend) is derived from "朋友" (friend) by adding the prefix "小" (little). Some other examples of prefix and suffix are given below:

prefix: 大(big) 总(general) 副(vice), ...
suffix: 人(person/people) 们(plural mark)
化(-ize/-ization)...

Semi-words

Most single characters in Chinese can be words. In dictionary-based approaches, if a character is not grouped with its neighboring characters, that individual character is usually considered to be a word. In fact, a single character has much less chance to be a word than a compound dictionary item, as noted by Bai [1]. Bai labels a single character as a "semi-word" in order to distinguish it from a compound word in the dictionary. He suggested that a compound word candidate should be preferred to a single-character word candidate. We incorporate this principle in our approach: a single character is assigned the probability $p/2$ if p is the default probability assigned to the items in the dictionary. By this means, the correct segmentation "日本 国民" for "日本国民" (Japanese people) is preferred to the incorrect one "日本国 民" (Japan, people).

The segmentation process

The segmentation process is similar to the statistical approach. Given an input string to be segmented, the following two steps are applied:

1. Each character in the input string is associated with all the candidate words starting from that character, together with their probability.
2. The candidate words are combined to cover the input string. The word sequence having the highest probability is chosen as the result.

Here is an example to show the process of the hybrid segmentation with the default probability set to 0.001 (cf. Figure 2).

Example 大会决议和议程项目 (resolutions and procedure items of the congress)

1. After the dictionary look-up, the word candidates, together with their probability, are associated to each character in the string. For example, for the first character "大", two possible words are found: "大会" (congress) and "大" (big).
2. The combination procedure is applied to the input string to find the word chains which cover the input string. The following chain with the highest probability (= 1.0 * 0.956 * 0.945 * 1.0 * 0.936) is chosen:

大会 决议 和 议程 项目
(congress, resolution, and, procedure, item)

4. Experiments on segmentation

We tested our approach using two corpora, both from the United Nations (see the following figure for their characteristics). Both corpora are segmented manually by a Chinese speaker. Each corpus is split into a training set and a test set. The training set has been used to calculate the probability for potential words (as in the simple statistical approach described in section 2).

Corpora	Size (Kbyte)	training set	test set
Corpus 1	164	149	15
Corpus 2	1 270	1 247	272

Table 1. Characteristics of the corpora

Different default probability values have been used in the hybrid segmentation in order to examine their impact on the global segmentation performance. Table 2 shows the amount of errors using the hybrid approach to segment the training set and the test set of corpus 1 (similar observations have been obtained on Corpus 2).

We see in this table that the dictionary-based and statistical approaches alone do not yield satisfactory results, either for the training set or for the test set. In the case of the statistical approach (p=0), the training set is segmented with a very high accuracy (98.5% of accuracy), but a lot of errors are produced for the test set (38.6% of error). On the other hand, in the case of the dictionary-based approach (p≈1), the error ratio is almost the same for the training data and test data. The segmentation accuracy is around 91%.

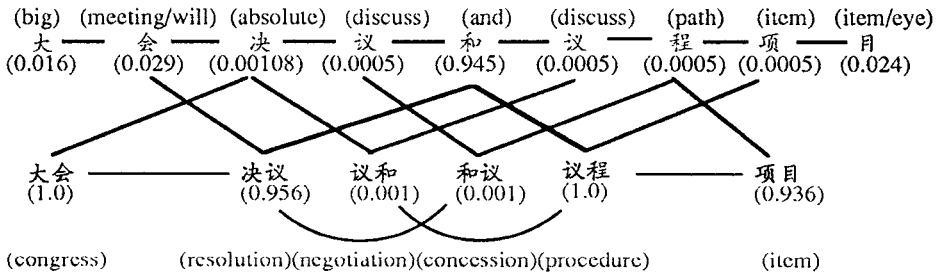

Figure 2. An example of segmentation process

default probability (p)	Nb. (%) of errors for training set (34433 words)	Nb. (%) of errors for test set (3487 words)
0	52 (0.15%)	1346 (38.60%)
0.00001	50 (0.14%)	272 (7.80%)
0.0005	50 (0.14%)	105 (3.01%)
0.001	50 (0.14%)	104 (2.98%)
0.005	62 (0.18%)	103 (2.95%)
0.01	73 (0.21%)	101 (2.90%)
0.02	106 (0.31%)	109 (3.13%)
0.05	152 (0.44%)	105 (3.01%)
0.1	196 (0.57%)	103 (2.95%)
0.2	292 (0.85%)	99 (2.84%)
0.3	381 (1.11%)	112 (3.21%)
0.4	479 (1.39%)	133 (3.81%)
0.5	552 (1.60%)	142 (4.07%)
0.9999	3405 (9.89%)	324 (9.29%)

Table 2. Impact of the default probability

In the case of the hybrid approach (when the default probability is between 0.00001 and 0.9999), much better results are obtained. The following graph shows the variation of segmentation accuracy of the hybrid approach on the test data in both corpora.

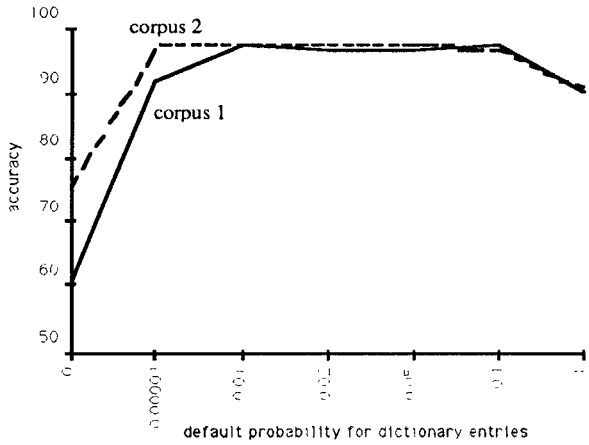

Figure 3. Segmentation accuracy for different approaches

Although the accuracy ratios, even in the best cases (around 97%), seem to be still lower than some previous reports (over 99%), we should notice that an important part of these errors are due to the inconsistency in manual segmentation which is used to evaluate the automatic segmentation processes. For example, some strings (especially some usual locutions) such as "起作用" (influence/make effect), "有效率" (efficient), "特别是" (in particular), are sometimes segmented as single words, sometimes as two separate words by the same Chinese speaker. This inconsistency is unavoidable as long as there is no clear definition of word in Chinese and the corpus is of considerable size. For a correct accuracy evaluation, any manual segmentation for the above cases should be considered as a correct one. However, in our automatic evaluation process of the accuracy, only one possible solution is considered to be correct. This makes the accuracy measures lower than their actual level.

Another important factor is related to the segmentation of proper names. In our corpora, there are quite a number of non-Chinese proper names such as "斯蒂芬. 施韦贝尔" (Stephen Shwebel). Without a special processing, these strings cannot be segmented correctly. In [16] we described a statistical unknown word detection process which succeeded in identifying frequent proper names. In our current corpora, however, the same approach do not apply because these proper names occur only occasionally. For example, the above proper name "斯蒂芬. 施韦贝尔" only occurs once. For the recognition of proper names of non-Chinese people, some heuristic rules may be very helpful. Usually, these proper names are translated into Chinese using a relatively small set of characters. A string of such characters that is not a common word has a high chance to be the translation of a non-Chinese name. This approach has been used in, among others, [20].

For the statistical approach and the dictionary-based approach, they have some more problems.

About the dictionary-based segmentation

The criterion of longest matching is not sufficient in a number of cases. For example, for the string "替人类" (for humanity), the following two possible segmentation results are equally plausible for this approach: 替人 类 (for other people, kind) and 替 人类 (for, humanity). An arbitrary choice is then used. It is expected that only 50% of these cases are segmented correctly. Another problem is related to the recognition of affix structures. The conditions for affix structures are sometimes not strict enough. This makes strings to be incorrectly identified as words. For example, In our process, "人" (people/person) is defined as a possible suffix (as in "中国人" (Chinese people)). So it can also be attached to the known word "许多" to form a wrong word "许多人" (many people) in "许多人家" (many families). Again, the longest-matching algorithm alone cannot choose the correct one (the second) among the two possibilities: "许多人 家" (many people, family) and "许多 人家" (many families).

About the statistical segmentation

This approach strongly relies on the good coverage of the training data for the test data. The problem of incomplete coverage makes the approach unable to segment correctly some common words. For our test sets, the words "引渡" (extradition), "深海" (high sea), "惩办" (prosecute), "交响乐队" (orchestra), and so on, are segmented into single characters because these words never occurred in our training data.

Most of the above problems specific to the dictionary-based approach and the statistical approach can be solved in the hybrid approach by an appropriate combination of the two kinds of knowledge on words. For the problems found in the dictionary-based approach, most of them can be solved by taking into account the statistical information on the words. This allows us to choose the most frequent words in priority. In comparison with the statistical approach, the incomplete coverage of statistical information may be

compensated by the incorporation of the dictionary. This is the reason why the hybrid approach results in higher accuracy than the two other approaches.

5. Application to Text Retrieval

In this section we describe our adaptation of SMART system to Chinese texts and our experimental results.

Indexing

When Chinese texts have been segmented, traditional RI approaches may be adapted for their retrieval. This is the approach we took: we adapted the SMART system [2] to index our segmented Chinese texts after a slight modification in order to tokenize Chinese texts correctly. The tf*idf weighting scheme is used in our experiments.

In order to ignore the non-meaningful words for document indexing and searching, we set up a stop-list of over 300 Chinese words which are the most used functional words in our corpora. These words are usually prepositions, adverbs and non meaningful nouns and verbs. Here are some items included in the stop-list:
的 (of), 按照 (according), 把 (make), 被 (by), 比 (than), 比较 (relatively), 并 (and), 并且 (also), 不论 (either), 不能 (cannot), 才 (only), 常 (often), 除非 (unless), 此外 (in addition), 问题 (problem), 注意到 (notice)

Test corpus

The adapted retrieval system has been tested using the Corpus 2 which is composed of a number of sections, each in turn consisting of a number of paragraphs. As our purpose here is to evaluate the effectiveness of text retrieval in non-structured Chinese texts, we do not consider the structural information. We simply consider each paragraph as a retrieval unit (document), and all are put together to form the test corpus which amounts to 3307 documents. The average length of the documents is about 90 words or about 160 Chinese characters.

Queries

We make use of the table of content of the original Corpus 2 to set the test queries. (Note that this table of content is not included in our corpus for retrieval test). 13 important themes are determined as our test queries, and the corresponding paragraphs according to the table of content are considered to be their relevant documents. In so doing, we expect the relevance judgments to be as objective as possible. The average length of the queries is 9 Chinese characters, and the average number of relevant documents is 16.

Preliminary test results

We run the system with the segmentation results of three different approaches: the dictionary-based approach, the statistical approach and the hybrid approach with the default probability set at 0.001. As the queries are Chinese phrases, they are also segmented respectively using the three segmentation processes, and then transformed into Boolean queries (disjunction). The correspondence R(d,q) between a document d and a query q is evaluated using the following fuzzy Boolean model:

$R(d,t) = w(t,d)$ where w(t,d) is the weight of the word t in the document d;

$R(d,A \wedge B) = R(d,A) * R(d,B)$

$R(d,A \vee B) = R(d,A) + R(d,B) - R(d,A) * R(d,B)$

$R(d,\neg A) = 1 - R(d,A)$

Figure 3 shows the variation of the precision ratio over the recall ratio for the three segmentation approaches.

We see that the retrieval performances using the three segmentation approaches are consistent with their segmentation accuracy. At this point, we can say that the better the segmentation, the better the retrieval.

We can further observe an important difference between the retrieval using the dictionary-based segmentation and those using the two other segmentation approaches. However, there is only a marginal difference between the statistical approach and the hybrid approach. We indicated earlier that one major difference between the two approaches is due to the incomplete coverage of the statistical

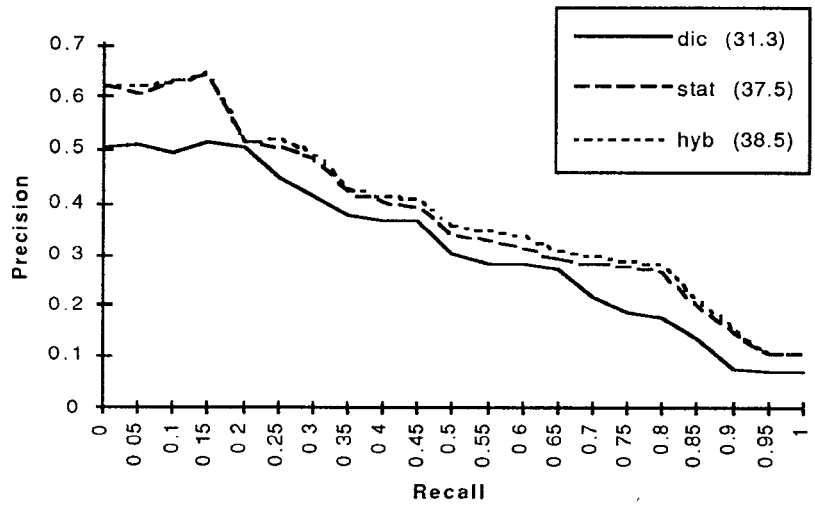

Figure 4. Comparison of the retrieval performance

information. However, our queries do not use these common words. So the difference between them is not reproduced in the retrieval.

Now, let us give some more analysis on the query evaluations.

Unknown words

In the following query:

q1. 国际移徙和难民 (international migration and refugees)

the word "移徙" (migration) is not a common word, and is not stored in our dictionary. However, it is frequently used in our corpus. So, by the dictionary-based approach, the word is segmented into two separated characters but recognized as a word by the two other approaches. Thus, for this query, the retrieval performance using dictionary-based segmentation (average precision of 27.7%) is much lower than using the two other segmentation approaches (44.7% for the statistical approach and 59.2% for the hybrid approach).

Incorrect segmentation

Some words may be composed of two or more simpler words. In this case, it is difficult to determine whether a string should be segmented into a single word or into several words. The correct segmentation of such strings has a great impact on the retrieval performance, as shown by the following examples.

q5. 五十周年纪念的筹备 (the preparation of the celebration of the fiftieth anniversary)

In this query, the string "周年纪念" (anniversary/ anniversary celebration) is considered as a single word by the dictionary-based approach because it is stored in our dictionary. This word may be separated to two common words: "周年" (anniversary) and "纪念" (celebration). In the case of this query, it should be separated. In some relevant documents, the concept "anniversary celebration" is expressed in different ways than "周年纪念", for example, "纪念五十周年" (celebrate the fiftieth anniversary), "五十周年 ... 纪念" (the fiftieth anniversary ... celebration), and so on. These documents do not match the word "周年纪念". It may be expected that, if, during document-query comparison, the complex word is allowed to break down into simpler words, then some of the unmatched relevant documents may be identified. However, we may face with the reverse problem as shown by in the following example.

q8. 世界银行和区域开发银行的贷款 (credits of the World Bank and the Regional Development Bank)

The string "世界银行" (the World Bank) is stored as a word (a proper name) in our dictionary, so segmented into a single word by the dictionary-based approach. This is the correct segmentation in this case. The same string is segmented into two words by the two other approaches: "世界" (world) and "银行" (bank). As a consequence, the retrieval performance using the dictionary-based segmentation is significantly better than that using the two other approaches.

Considering the above two examples together, how to deal with the problem of complex word vs. simple words is still an open question, because a solution of one problem may engrave the other problem.

6. An attempt of integrating a Chinese thesaurus

The retrieval process we used is a keyword-based process, thus it inherits all the problems of this latter, in particular, for a document to be considered "relevant", it has to contain the same keywords as those in the query. If a concept is expressed in a different way in a document, then it is possible that the document cannot be retrieved, although it is highly relevant. This problem of *silence* has been noted for a long time. Query expansion, which is aimed to find more potentially relevant documents, has often been suggested as the solution to it. In our previous study [15], a flexible method has been developed for the use of manually established thesauri in query expansion. In this approach relations stored in a manual thesaurus are viewed as fuzzy relevance relations between terms (or words):

$$\{... A \rightarrow_c A_1, ...\}$$

where $A \rightarrow_c A_1$ means that the term A_1 is relevant to the term A to the extent c ($\in [0,1]$). The extent c is determined according to the nature (i.e. synonymy, hypernymy, etc.) of the thesaurus relation between A and A_1 as well as to the relevance feedback information. These fuzzy relevance relations are used to expand the initial Boolean query as follows:

For each term A in the initial Boolean query, if $A \rightarrow_c A_1$ is a relevance relation, then A_1^c is put in disjunction with A in the query (i.e. A is replaced by $A \vee A_1^c$). The expression A_1^c means: if a document corresponds to A_1 to the extent w, then it corresponds to the original term A to the extent c*w only.

This method has been tested on CACM collection with the thesaurus Wordnet [14]. Significant improvement has been obtained by making query expansion in this way. We believe that the same query expansion is extremely useful to Chinese IR. However, to our knowledge, there is no Chinese thesaurus available for RI systems. In an attempt to establish such a thesaurus, we made use of an electronic bilingual dictionary (EDICT). In this dictionary, an English word is interpreted into Chinese words. For example:

ability n.才干;能力;(通常pl.)本领,技能 a manifold abilities 多才多艺
abolish vt.废止,废除,撤废 ~.able a.可废除的 ~.ment n.废止,撤废
administration n.管理;管理部门;统治;行政机关, 政府;内阁.给与,配药; 遗产财政管理.
advantage n.1.利益;便利.2.优势,优越;有..的长处;.胜过, 较..有利.vt.利于,对..有益
association n.1.联合,结合.2.公会,社团,协会.3.联想. 4.交际,结交.5.[运]A式足球.
budget n.预算,营运费;家计;囊中之物;一束 vt.vi.编入预算,编制预算;预定[时间等]
commerce n.1.商业;通商,贸易,交易.2.交际,交涉. 3.一种牌戏.
congress n.1.会议,大会.2.(C-)议会,国会[美国的]; 国会的开会期.3.协会.4.性交.

It is possible to transform the dictionary into a simple thesaurus: All the Chinese words that translate the same English word (or one of its senses) are considered to relate to each other. For example, the words which translate "ability" (才干, 能力, 本领, 技能) are considered to relate to each other (they indeed denote the same sense in this case). The words which translate the three different senses of "commerce" may be considered as three groups of inter-related words: {商业, 通商, 贸易, 交易} (commerce), {交际, 交涉} (social exchange) and {一种牌戏} (a card game).

The above examples show the cases where the English word senses are well distinguished in this dictionary. This is not the case for many other English words for which all the senses are merged in their translation. For example, for "administration", the sense of "drug administration" (配药) is merged with that of "government administration" (行政机关, 政府), "cabinet" (内阁), "administration department" (管理部门), and so on. So the above simple creation of relations among Chinese words also creates a lot of noise in addition of desired relations.

In the thesaurus created from the dictionary, each word is connected to about 20 other words. This thesaurus is applied in the query expansion as described earlier. As there is no information about the type of relation created, we assign the same value c for each related word, i.e. each pair of related words A and B is seen as the fuzzy relevance relation $A \rightarrow_c B$. The following table shows the retrieval performance using query expansion on the result of dictionary-based segmentation with three different values for c: 0.05, 0.1 and 0.5.

recall	c=0	c=0.05	c=0.1	c=0.5
0	0.5005	0.5023	0.5045	0.4853
0.05	0.5082	0.5100	0.5122	0.4930
0.1	0.4967	0.5009	0.5046	0.5211
0.15	0.5163	0.5234	0.5239	0.5369
0.2	0.5052	0.5057	0.5068	0.4754
0.25	0.4481	0.4474	0.4478	0.4164
0.3	0.4143	0.4193	0.4193	0.3775
0.35	0.3798	0.3845	0.3889	0.3801
0.4	0.3693	0.3717	0.3730	0.3560
0.45	0.3690	0.3685	0.3688	0.3405
0.5	0.3026	0.3019	0.3017	0.2860
0.55	0.2851	0.2864	0.2870	0.2642
0.6	0.2815	0.2818	0.2818	0.2614
0.65	0.2734	0.2717	0.2691	0.2512
0.7	0.2154	0.2182	0.2204	0.2385
0.75	0.1872	0.1898	0.1923	0.2095
0.8	0.1764	0.1822	0.1847	0.2077
0.85	0.1343	0.1376	0.1374	0.1596
0.9	0.0724	0.0757	0.0754	0.1010
0.95	0.0697	0.0701	0.0698	0.0980
1	0.0697	0.0701	0.0698	0.0980
average	0.3131	0.3152	0.3162	0.3123

Table 3. Impact of the thesaurus on text retrieval

(For the query evaluations using the two other segmentation approaches, similar results are obtained.)

We can see that the retrieval performance is only marginally affected by the simple thesaurus. This result may be explained by the following facts:

1). A lot of "noise" words are introduced in the query. The possible positive impact brought by the truly related words is neutralized by the noise.

2). The documents from the United Nations are written in a formal language. The vocabulary is restricted. In such an application, the utility of thesaurus is marginal.

Despite the problems, our experiments are interesting because they show that query expansion is feasible in Chinese IR. Although the experimental results do not show a significant impact of the thesaurus on the retrieval effectiveness, we believe that, with a Chinese thesaurus of higher quality, query expansion will be extremely useful to Chinese RI, especially for news retrieval. This is one of our future research subjects.

7. Conclusions and Future work

In this paper, we described a hybrid segmentation approach which makes use of both human-defined word knowledge and statistical information. In comparison with other segmentation approaches, this approach is highly flexible: it can cover a wide range of segmentation approaches from the statistical approach to the dictionary-based approach. In terms of performance, the hybrid approach proves to be clearly better than the two others.

We also adapted the SMART system to retrieve the segmented Chinese texts. Our adaptation shows the feasibility of using RI systems designed for Indo-European languages to Chinese. Our preliminary experiments showed that the retrieval effectiveness is directly affected by the quality of segmentation. So a good Chinese RI system should include a good segmentation process.

We also tested a query expansion approach using a simple thesaurus established from a bilingual dictionary. The query expansion had little impact on the retrieval performance of the system because of the poor quality of the thesaurus established and the strict vocabulary used in our test corpus. Our future work aims to build a better Chinese thesaurus for RI purposes, and to test the query expansion approach for Chinese news searching.

Acknowledgment: We would like to thank Chris Buckley who gave us useful hint for the adaptation of SMART to Chinese.

References

1. S. Bai, "Semi-word" method for Chinese word segmentation. *International Conference on Chinese Computing*, Singapore, 304-309 (1994).
2. C. Buckley, Implementation of the SMART information retrieval system. Cornell University, Technical report 85-686, (1985).
3. J.-S. Chang and e. al., Chinese word segmentation through constraint satisfaction and statistical optimization. *ROCLING-IV*, Taiwan, 147-165 (1991).
4. K.-J. Chen and S.-H. Kiu, Word identification for

Mandarin Chinese sentences *5th International Conference on Computational Linguistics*, 101-107 (1992).
5. T.-H. Chiang and e. al., Statistical models for segmentation and unknown word resolution. *5th R.O.C. Computational Linguistics Conference*, 123-146 (1992).
6. L.-F. Chien, Fast and quasi-natural language search for gigabits of Chinese texts. *Research and Development in Information Retrieval, ACM-SIGIR*, Seattle, 112-120 (1995).
7. T. Dunning, Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, vol. 19, pp. 61-74 (1993).
8. H. Fujii and W. B. Croft, A comparison of indexing techniques for Japanese text retrieval. *Research and Development in Information Retrieval, ACM-SIGIR*, 237-246 (1993).
9. K.-K. He, H. Xu, and B. Sun, The Design Principle for a Written Chinese Automatic Segmentation Expert System. *Journal of Chinese Information Processing*, vol. 5, pp. 1-14 (1991).
10. W. Jin, A Case Study: Chinese segmentation and its disambiguation. Computing Research Laboratory, New Mexico State University, Las Cruces, Technical report MCCS-92-227, (1992).
11. B.-I. Li and e. al., A maximal matching automatic Chinese word segmentation algorithm using corpus tagging for ambiguity resolution. *R.O.C. Computational Linguistics Conference*, Taiwan, 135-146 (1991).
12. N. Y. Liang and Y.-B. Zhen, A Chinese word segmentation model and a Chinese word segmentation system PC-CWSS. *COLIPS*, vol. 1, pp. 51-55 (1991).
13. M.-Y. Lin, T.-H. Chiang, and K.-Y. Su, A preliminary study on unknown word problem in Chinese word segmentation. *ROCLING V*, 147-176 (1992).
14. G. Miller, Wordnet: an on-line lexical database. in *International Journal of Lexicography*, vol. 3, 1990
15. J.-Y. Nie and M. Brisebois, An inferential approach to inforlation retrieval and its implementation using a manual thesaurus. *Artificial Intelligence Review* ((to appear) 1996).
16. J.-Y. Nie, W. Jin, and M.-L. Hannan, A hybrid approach to unknown word detection and segmentation of Chinese. *International Conference on Chinese Computing*, Singapore, 326-335 (1994).
17. Y. Ogawa, A new character-based indexing organization using frequency data for Japanese documents. *Research and Development in Information Retrieval, ACM-SIGIR*, Seattle, 121-129 (1995).
18. Y. Ogawa, A. Bessho, and M. Hirose, Simple word strings as compound keywords: An indexing and ranking method for Japanese texts. *Research and Development in Information Retrieval, ACM-SIGIR*, 227-236 (1993).
19. R. Sproat and C. Shih, A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, vol. 4, pp. 336-351 (1991).
20. R. Sproat, C. Shih, W. Gale, and N. Chang, A stochastic finite-state word-segmentation algorithm for Chinese. *ACL conference*(1994).
21. M.-S. Sun and e. al., Some Issues on the statistical approach to Chinese Word Identification. *3rd International Conference on Chinese Information Processing*, 246-253 (1992).
22. L.-J. Wang, T. Pei, W.-C. Li, and L.-C. Huang, A Parsing method for identifying words in Mandarin Chinese sentences. *12th International Joint Conference on Artificial Intelligence*, Sydney, Australia, 1018-1023 (1991).
23. Z. Wu and G. Tseng, Chinese text segmentation for text retrieval: Achievements and problems. *Journal of the American Society for Information Science*, vol. 44, pp. 532-542 (1993).
24. Z. Wu and G. Tseng, ACTS: An automatic Chinese text segmentation system for full text retrieval. *Journal of the American Society for Information Science*, vol. 46, pp. 83-96 (1995).
25. H. Xu, K.-K. He, and B. Sun, The implementation of a written Chinese automatic segmentation expert system. *Journal of Chinese Information Processing*, vol. 5, pp. 38-47 (1991).
26. T.-S. Yao, G.-P. Zhang, and Y.-M. Wu, A rule-based Chinese automatic segmentation system. *Journal of Chinese Information Processing*, vol. 4, pp. 37-43 (1990).
27. C.-L. Yeh and e. al, Rule-based word identification for Mandarin Chinese sentences - A unification approach. *Computer processing of Chinese and Oriental Languages*, vol. 5 (1991).
28. Y.-X. Zhou and W.-T. Wu, A Practical Method of Segmentation of Chinese -- A Method Based upon Chain Table. *Journal of Chinese Information Processing*, vol. 4, pp. 34-41 (1989).