A UNIFYING APPROACH TO SEGMENTATION OF CHINESE AND ITS APPLICATION TO TEXT RETRIEVAL

Jian-Yun Nie [1]
Xiaobo Ren [2]
Martin Brisebois [1]

[1] University of Montreal,
BP.6128, succ. Centre-ville, Montreal, Quebec, H3C 3J7 Canada
[2] Center for Information Technology Innovation
1575 Bd. Chomedey, Laval, Quebec, H7V 2X2 Canada

In segmentation of Chinese, two competing approaches have been often used separately: the rule-based approach and the statistical approach. Each approach has its advantages and disadvantages. In this paper we describe a hybrid approach which unifies them in a single flexible segmentation process in which items stored in the dictionary or identified by heuristic rules are assigned a default probability. By varying the default probability value, the hybrid approach can cover a wide range of approaches from the purely statistical one to the purely rule-based one. Our experiments on two corpora show that by a proper setting of the default probability, the hybrid approach gives much better results than statistical or rule-based approaches alone. A text retrieval system is then adapted to the segmented Chinese texts. Preliminary results of the retrieval system are reported.

1. Introduction

Natural language processing is an important issue in many areas such as Information Retrieval (IR). IR systems for Indo-European languages are widely used in libraries, information centers and increasingly across the information web in computer networks. An IR system aims to select the texts from a corpus which are relevant to a given query [1]. Typically, a system determines the relevant documents according to the frequency of occurrences of the *words* of the query within the documents and the corpus. 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 in Chinese texts. Thus traditional approaches for IR cannot be directly applied to Chinese.

One might think that, as there is no available separation of words in Chinese, text retrieval can operate on a character string basis. This approach has been used in some experimental systems for Japanese text retrieval [2, 3] for which the same problem is encountered as for Chinese texts. However, this approach would lead to a great deal of

incorrect matching between queries and documents due to the almost 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 他认识别的人 (he knows other people) by the character-based approach. In addition, character-based retrieval would lead to an explosion of index file size due to the great number of character combinations as searching keys.

We believe that Chinese text retrieval should operate on segmented texts in order to gain efficiency and quality in the retrieval operation. Moreover, this approach can benefit much from the development of information retrieval for Indo-European languages.

The process of segmentation has been the subject of much intensive research in the area of computer-based analysis of Chinese for the past decade. These approaches may be classified into two main groups: the rule- and dictionary-based approach and the statistical approach. Approaches in the first group rely on knowledge defined by human experts (dictionary and heuristic rules) in segmentation. These approaches only make use of general knowledge on Chinese words: the words included in the dictionary are often the most usual ones, and the heuristic rules correspond to common word structures. On the other hand, approaches of the second group use specific statistical information about the corpus or application area. These two approaches have often been used separately in automatic segmentation processes, except in a few ones such as [4]. This does not correspond to the human segmentation process in which both general knowledge and specific information are used.

In this paper, we describe a hybrid approach for segmentation of Chinese which uses dictionary, heuristic morphological rules and statistical information in a single process. The basic idea is to consider general knowledge as background knowledge, and to place specific statistical information in front of it. This idea is achieved simply by assigning a default probability to items stored in the dictionary or identified by the heuristic rules.

This approach has a high flexibility: By varying the default probability value, the hybrid approach can cover a wide range of approaches from the purely statistical approach to the purely rule-based approach.

We tested our approach with two corpora. We have shown that for both corpora, the hybrid approach yields better results than the two competing approaches alone. We further adapted a general IR system, SMART, to our segmented Chinese texts. The performance of the IR system for Chinese is evaluated with respect to different segmentation approaches. It is shown that the segmentation quality has a great impact on the retrieval quality.

2. Statistical approach vs. Rule- and dictionary-based approach

Dictionary-based approaches [5-14] 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. This algorithm may be described as follows:

An input character string is compared with the contents of the dictionary so that all sequences of characters constituting recognized lexical items can be highlighted. Words are linked from beginning to end of the input string, with several candidate word chains being proposed. Among all possible word chains, the one with the fewest and thus the longest words is considered to be the best segmentation.

The above approach is often extended by a set of heuristic morphological rules [7]: a character string which is not stored in the dictionary, but may be derived from the 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) or nominal pre-determiner structure (一百个人 - hundred people).

Rule- and dictionary-based approaches have the advantage of being simple, general and often efficient: The heuristic knowledge built into the system corresponds closely to knowledge about linguistic phenomena occurring in Chinese words and this knowledge is represented in a straightforward way, allowing human experts to verify its correctness. It has been shown that a simple rule-based approach may often achieve a performance comparable to that of a sophisticated statistical approach.

However, a prerequisite for high-quality results in rule- and 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).

Although the maximum-matching algorithm may solve the major part of segmentation ambiguity, several possible segmentation results may still remain because they have equal lengths. To solve the remaining ambiguity, it has often been suggested that syntactic, semantic, or even pragmatic analysis should be used [6]. In practice, however, we do not have enough knowledge for the last two analyses to be feasible at the present time. Even for the syntactic analysis, although one succeeded in analyzing the core part of Chinese syntax [15], it still seems to lack of syntactic rules in Chinese that have a good coverage and are as rigorous as in Indo-European languages. In IR context, especially, as texts may be written in different styles and concern various areas, this solution is difficult to materialize now. Instead of using sophisticated linguistic analyses, we suggest to use statistical information as an alternative solution.

Statistical approaches [16-21] do not need pre-established dictionary and rules. They rely on statistical information such as word and character (co-)occurrence frequencies in the text which may be obtained automatically from training data set manually. One of the advantages of statistical approaches is their capacity to cope with the particularities of

application areas through the statistical information. The simplest statistical approach is as follows:

Given a manually segmented training document set, 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.

Many statistical approaches make use of more complex, typically first-order Markov, models. Although statistical approaches avoid the tedious task of establishing a dictionary and heuristic rules, they require a great deal of manually segmented texts to train the model. The training data are also difficult to set up (often not much easier than setting up a dictionary). Moreover, inconsistency is often unavoidable and difficult to check in manual segmentation, affecting the reliability of the statistical information obtained. In addition, the acquisition of statistical information is not cumulative. Probabilities need to be revised constantly. From this point of view, there is no clear advantage for statistical approaches on data preparation.

Through the above analysis, we can see that rule- and dictionary based approaches and statistical approaches have quite complementary properties: The former is general but application-insensitive; the latter is specific but the statistical information cannot be generalized. It is natural then to suggest a hybrid approach which combines them in a single approach in order to compensate the drawbacks of each approach with the advantages of the other.

Hybrid approaches have been used by a few researchers. Fan and Tsai [4], for example, describe a statistical approach which incorporates a dictionary. The probability of a dictionary entry is first assumed to be 1, then revised by a relaxation process using statistical information. However, the relaxation process will not apply to the words on which there is no statistical information, that is, the relaxation process may have a poor coverage. If uncovered words appear, the segmentation accuracy may be seriously affected. In our hybrid approach, a default probability much lower than 1 is assigned manually to all the lexical items in the dictionary. We do not have the problem of poor coverage. If statistical information is also available, it can be integrated readily with human established dictionary.

3. A hybrid segmentation approach

From a cognitive point of view, statistical data provide a sort of short term knowledge about the application context, whereas the vocabulary stored in a dictionary may be seen as long term knowledge generally accepted by people. 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) also provide useful indication. This latter information can be incorporated into a computer-based analysis. Our hybrid segmentation process works in a similar way:

A dictionary is used as a repository of background knowledge. Each entry in the dictionary is assigned a default probability. If statistical data are available, we can also establish a statistical dictionary which consists of a set of potential words together with their probability to be valid words in the given corpus. The two dictionaries can then be merged together in a statistical segmentation process such that both kinds of information are used.

Merging dictionary with statistic information

Although statistical approaches and rule-based approaches have often been seen as two competing ones, they are indeed compatible. In fact, a rule-based approach using longest-matching algorithm may also be seen as a special case of statistical segmentation: each potential word in the input string which is stored in the dictionary or derived from a heuristic rule, is assigned an equal probability (less than 1). Then the maximum-matching algorithm is equivalent to a statistical approach which chooses the segmentation result of the highest probability. For example, for the phrase 中国文学 (Chinese literature), there may be the following segmentation possibilities according to the dictionary:

中国 文学
中国 文 学
中 国文 学
中 国 文学
中 国 文 学

If each potential word is assigned equal probability values p (<1), then the first segmentation which contains the fewest words will have the highest probability p^2. The other possible results will have lower probabilities (p^3 or p^4). This result is the same as with the maximum-matching algorithm.

The rule- and dictionary-based approach being seen as a special statistical approach, it is then possible to combine them in a single hybrid segmentation process. In such a hybrid approach, if statistical information about a dictionary item is available, it is used in priority; otherwise, the default probability is assigned to that item. By varying the default probability value, we can change the relative importance of the statistical information and the dictionary. 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 a purely statistical approach. On the other hand, when the default probability value is very

high (near 1, but <1), the hybrid approach will consider almost exclusively the words stored in the dictionary. Thus we obtain the rule-based approach in this case. We see that the hybrid approach can cover a wide range of approaches from the purely statistical approach to the purely rule-based approach, as illustrated by the following figure:

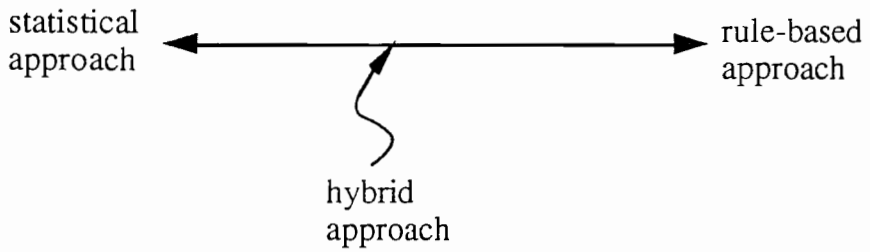

Figure 1. Comparison of the three approaches

In rule-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 string included in a dictionary, as noted by Bai [22]. Bai labels a single character not in the dictionary as a "semi-word" in order to distinguish it from a word in the dictionary. In his approach, the latter is used in preference to the former. In our approach, we apply the same principle: a single character is assigned the probability $p/2$ where p is the default probability assigned to dictionary items.

Heuristic rules

Apart from the dictionary, a set of heuristic rules is also incorporated into our segmentation process in order to identify and segment words which follow some rules (for example, numbers and dates). In this paper we only deal with the following two groups of morphological rules. More discussion about heuristic rules may be found in [7].

Nominal pre-determiner structure

Words corresponding to this structure frequently occur in Chinese, for example, 每一周 (*every week*), 这一回 (*this time*). In order to establish a set of heuristic rules for this structure, we first define the following categories of single characters:

- determiners: 这 (*this*), 那 (*that*) 此 (*this*) 该 (*this*) 其 (*its, his, her*)
  每 (*each*) 各 (*every*) 某 (*some*) 首 (*first*) ...
- 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*)
  组(*group*) 次(*time*) 层 (*layer*) 年 (*year*) 月(*month*) 日(*day*)...

The following rules cover a major part of the words in this structure (where [...] indicates optional status and [...]* an optional arbitrary repetition):

ordinal cardinal [classifier] → pre-det	第一周 (*first week*) 第二 (*second*)
determiner [cardinal]* classifier → pre-det	这一回 (*this time*) 每层 (*every layer*)
cardinal [classifier] → pre-det	十一(*eleven*) 一九九一年(*in 1991*)
	一百本(*hundred books*)

Apart from these general rules, some special cases are also considered. For example, some determiners (各, 首) cannot be followed by an ordinal as in 首一次, 各一组, but can be followed by a classifier suck as 首次 (first time), 各组 (each group).

Affix structure

In our segmentation, for a word to be considered as having an internal affix structure, both of the following conditions should be true:

1. The first (last) character should be a possible prefix (suffix). For example:
   prefix: 大(big) 小(small) 总(general) 副(vice) ...
   suffix: 人(person) 们(plural mark) 权(right) 会(association) 化(-ize/-ization), ...
2. The remaining characters should form a known word.

Most internal affix structures fit these conditions. However, the second condition is not always true. For example the string 副总经理 (*vice general manager*) cannot be identified to be a single word having a duplicated prefix structure 副 + 总 + 经理 due to the second condition. The setting of this condition is to prevent classifying some strings incorrectly as a word such as for the string 保护人权 (*protect human rights*). Without the second condition, this string may be identified as a single word formed from the word 保护 by adding two successive suffixes 人 and 权, which may mean "the rights of protectors". This latter case occurs more frequently in our corpora than the former. Thus we keep the condition for a practical reason. However, we are aware that this condition should be replaced later by other more refined conditions.

4. Implementation

In order to give a more thorough view of our system, we describe several implementation details in this section.

Dictionary organization

Both manual dictionary and statistical information are stored in a run-time dictionary. In order to increase efficiency in dictionary look-up, this dictionary is organized as an open

hash table. The first Chinese character C_1 (2 bytes) of a word is used to calculate a unique location Hash(C_1) in the hash table. Each location in the hash table points to a list of words starting by the character. The following figure shows a fragment of the run-time dictionary (where i is the hash address for 人 and i+1 for 忍, i.e. Hash(人) = i and Hash(忍) = i+1; and the real number are probabilities of the words):

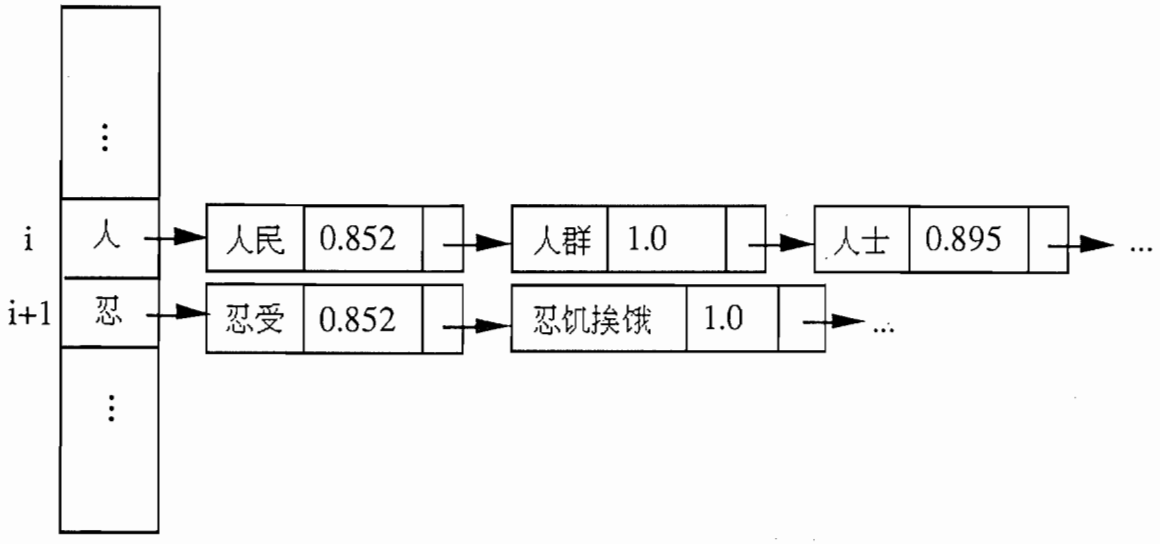

Figure 2. Organization of the run-time dictionary

Our manual dictionary contains over 91 000 entries. A few thousand new words are identified in the statistical information. These new words are mainly names of non-Chinese people and countries, or words that can be identified by heuristic rules and are not included in the manual dictionary.

The organization of the segmentation process

The segmentation process is similar to a purely statistical approach. Given an input string to be segmented, the following two main sub-processes are applied to it:

1. Dictionary look-up:
   This sub-process associates to each character in the input string a list of the candidate words, together with their probability, which are substrings of the input string starting from this position.
2. Find the best combination of the candidate words:
   This sub-process combines the word candidates to cover the entire input string and chooses those combinations that have the highest probability.

The first sub-process is quite straightforward. The complexity of the algorithm is mainly determined by the combination procedure. The following recursive algorithm is used:

Procedure best-combine($C_1...C_iC_{i+1}...C_n$);

/* $C_1...C_iC_{i+1}...C_n$ is the input string */

1. For each word candidate $C_1...C_i$ at the beginning:

   a) find the set of the best combinations for the remaining string $C_{i+1}...C_n$:

      R := best-combine($C_{i+1}...C_n$),

   b) for each S_k in R:

      - combine the word candidate $C_1...C_i$ with S_k,
      - assign the probability $p(C_1...C_i)*p(S_k)$ to the segmentation starting by the word $C_1...C_i$ followed by S_k;

2. Return the set of combinations covering the string that have the highest probability together with that probability.

We give some examples to illustrate the segmentation process. These examples show the actual process of the hybrid segmentation with the default probability set to 0.001.

Example 1: 大会决议和议程项目

1. After the dictionary look-up, the following word candidates, together with their probability, are associated to each character in the string:

大: 大会 1.000000, 大 0.016073
会: 会 0.029028
决: 决议 0.955782, 决 0.001081
议: 议和 0.001000, 议 0.000500
和: 和议 0.001000, 和 0.944933
议; 议程 1.000000, 议 0.000500
程: 程 0.001000
项: 项目 0.936073, 项 0.023973
目: 目 0.000500

2. The combination procedure is applied recursively to the input string such that word sequences are built from end to beginning. For the substring 目, there is only one possibility with probability = 0.005. For the substring 项目, two combinations are possible:

项目 0.936073
项 目 0.000013

Only the best one (项目) is used for further combination with characters before it. So for the substring 程项目, the only retained combination is 程 项目. For the substring 议程项目, we have again two possibilities, but only 议程 项目 will be retained. This combination process is to be applied until the first character of the input string has been combined. Finally, the following correct segmentation is chosen as the best result:

大会 决议 和 议程 项目

which is of the highest probability.

We notice that although there are several combinations for the substring 决议和议程 (决议 和 议程, 决 议和 议程, 决议 和议 程) that would all remain as possible solutions in a rule-based approach, our hybrid segmentation is able to determine the correct one using the statistical information: 决议 和 议程. This example shows the contribution of statistical information.

Example 2: 1993年8月17日第47/233号决议

In our implementation, special attention has been paid to the determination of complex pre-determiner strings that contain Chinese and ASCII characters as in this example. A string of ASCII numbers (and some other kinds of special strings) is considered as an inseparable token.

After the dictionary look-up, the following word candidates are associated:

1993: 1993年 0.001000, 1993 0.001000
年: 年 0.683775
8: 8月 0.001000, 8 0.001000
月: 月 0.935073
17: 17日 0.001000, 17 0.001000
日: 日 0.767800
第: 第47/233号 0.001000, 第 0.533917
47/233: 47/233号 0.001000, 47/233 0.001000
号: 号 0.869823
决: 决议 0.955782, 决 0.001081
议: 议 0.000500

The candidate words 1993年, 8月, 17日, 第47/233号, 47/233号 are all identified as pre-determiner structures. They are assigned the default probability.

Finally, the selected result is the following:

1993年 8月 17日 第47/233号 决议

5. Experiments

We tested our hybrid approach on two corpora, both from the United Nations. We segmented both corpora manually. Automatic segmentation results are compared with the manual one to evaluate their accuracy. 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 (see section 2 for the calculation). The characteristics of the corpora are highlighted in the following table:

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. The following table shows the number 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):

default probability p for items from manual dictionary	No. of errors in segmenting training set (34433 words)	No. of errors in segmenting test set (3487 words)
0	52	1346
0.00001	50	272
0.0005	50	105
0.001	50	104
0.005	62	103
0.01	73	101
0.02	106	109
0.05	152	105
0.1	196	103
0.2	292	99
0.3	381	112
0.4	479	133
0.5	552	142
0.9999	3405	324

Table 2. Influence of the default probability in the hybrid segmentation

We can see in this table that both competing approaches alone do not yield satisfactory results, either for the training set or for the test set. In the case of the pure statistical segmentation (p=0), the segmentation of the training set is very good. This is reasonable because the approach is trained by the same data. When the approach is applied to the test data, however, we observed a high ratio of error (38.6%). This is mainly because the training data do not completely cover the test set. This observation is consistent with the remark we made earlier that for a statistical approach to yield good results, it is essential that the training data has a good coverage of the application area.

On the other hand, in the case of the purely dictionary- and rule-based approach (p≈1), the error ratio is almost the same for the training data and test data. The segmentation accuracy is around 90%. In comparison with the other reports of near 99% of accuracy using such an approach, we note that in our corpora, there are quite a number of names of non-Chinese people or countries. Our current segmentation does not incorporate rules for the detection of such names. This subject has been investigated in some other studies, for example [23].

In the case of truly hybrid approach (when the default probability is between 0 and 1 exclusively), better results are obtained. The best results correspond to the setting of the default probability between 0.001 and 0.1. In some cases (between 0.00001 and 0.001), we even observed a performance on the training data better than the purely statistical approach.

The following graph shows the variation of segmentation accuracy of the hybrid approach on the test data in both corpora. The default probability value varies from 0 to near 1. We can draw the conclusion that the hybrid approach is significantly better than the two competing approaches alone. When the default probability is set between 0.001 and 0.1, we obtain the best results for both corpora about 97% accurate).

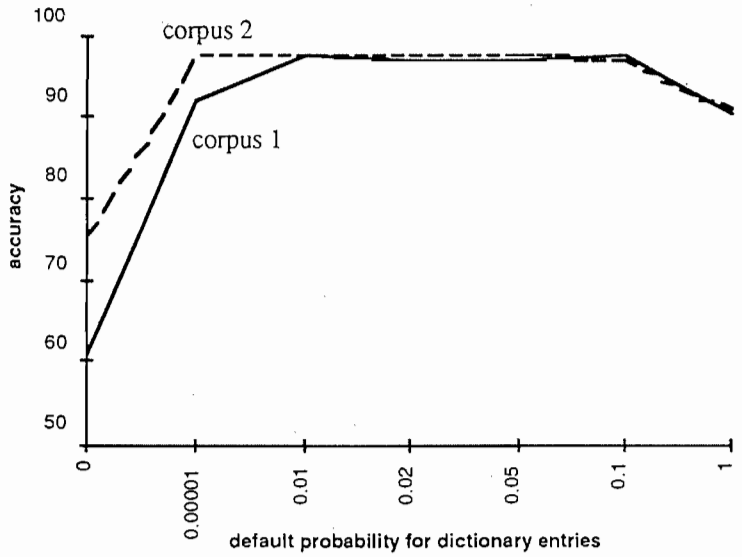

Figure 3. Segmentation accuracy for different approaches

6. Application to Text Retrieval

The problem of Chinese text retrieval has been investigated in [24, 25]. However, These studies mainly concerns the segmentation of Chinese texts rather than their retrieval. To our knowledge, there is no general IR system built for Chinese texts until now.

In this study, we try to build a complete IR system for Chinese texts. Note that when Chinese texts have been segmented, traditional IR approaches may be adapted to their retrieval. This is the approach we took: we adapted the SMART [26] system in our implementation. SMART is a text retrieval system developed in Cornell University. This system compasses a variety of tools for text tokenizing, word statistic measuring, and query evaluation.

Implementation

The application of SMART to index and retrieve segmented Chinese texts may seem to be easy and direct. However, as SMART is designed for English texts, it does not deal

with non-ASCII characters such as Chinese characters. To adapt it to Chinese texts, two solutions are possible:

1. extend the character set considered by SMART to cover non-ASCII characters;
2. encode Chinese texts by ASCII characters.

In our current implementation, the second solution is used. Chinese characters are encoded in HZ format in which each Chinese character is encoded by two ASCII characters. A Chinese character string is delimited within ~{ and ~} in order to make difference from ordinary (non quoted) ASCII characters. For example, the following string

SMART 信息检索系统

is encoded in HZ as the ASCII string: SMART ~{PEO"<lKwO5M3~}.

The problem with the encoded HZ texts is that Chinese characters are often encoded by symbols such as punctuation markers (?, !, . %, ...). As SMART checks for tokens according to English writing, Chinese characters are often incorrectly cut in the direct application. To solve this problem, we modified the SMART tokenizing program in order to deactivate the original tokenizing process and replace it with a new one which keeps the delimited Chinese codes together.

In indexing, SMART ignores the words which are considered as common-words. A list, called stop-list, of such words is set up for English. We enhanced the English stop-list by about 300 common Chinese words. These words are often adverbs and prepositions that are not important for IR purposes. We also included in the stop-list the Chinese symbols such as punctuation markers. Here are some items included in the stop-list:

按照, 把, 被, 比, 比较, 并, 并且, 不论, 不能, 才, 常, 除非, 此外.

The indexing process of SMART may now be applied to the Chinese texts (documents and queries) in order to extract important keywords from them.

Experiments

The adapted retrieval system has been verified by using the test set of Corpus 2. The test data are composed of 797 relatively independent paragraphs. We consider each paragraph as an independent document in our experiments. A set of 10 queries in Chinese in the domain of the these documents has been set up and manually evaluated by examining through the documents. The query evaluation of the system is compared with the manual evaluation in order to evaluate the system's performance in terms of precision and recall defined as follows:

$$\text{recall} = \frac{\text{the number of relevant document retrieved}}{\text{the number of relevant documents in the corpus}}$$

$$\text{precision} = \frac{\text{the number of relevant document retrieved}}{\text{the number of document retrieved}}$$

We applied the modified SMART system to the results of three different segmentation process: the purely statistical approach, the purely rule-based approach and the hybrid approach with default probability = 0.001. For document indexing, we used *tf*idf* scheme for keyword weighting [1]. Queries are evaluated using a simple Boolean retrieval method.

The following figure shows the variation of the precision ratio over the recall ratio for the three segmentation approaches.

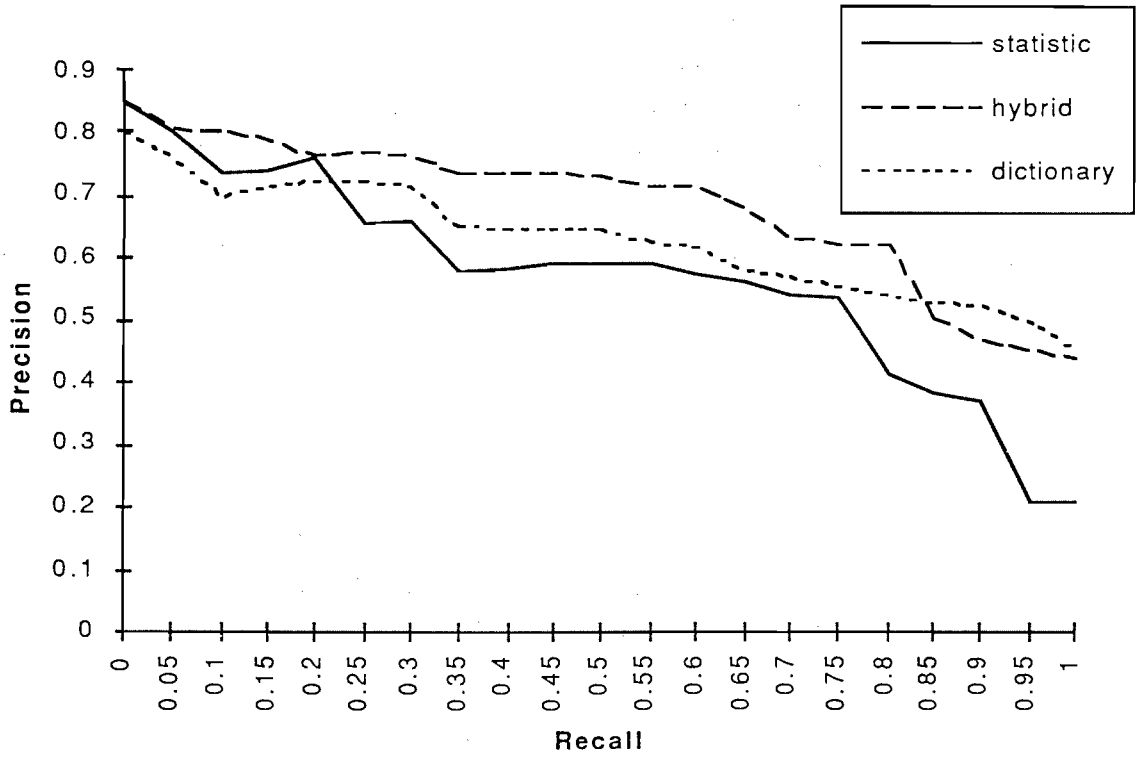

Figure 4. Evaluation of the retrieval performance

It can be seen that the hybrid segmentation leads to the best retrieval performance. This may be seen more clearly in the following table in which we give the *average precision* of the retrieval with respect to the three segmentation processes. The average precision is the common measure used for IR systems which is the average of precision ratios when the recall ratio = 0%, 5%, 10%, 15%, 20%, ..., 100% respectively. The following table shows the comparison of the system's performance with respect to the segmentation processes.

Segmentation approach	Average precision
statistical	56.79
hybrid (p = 0.001)	68.24
rule/dictionary-based	62.86

Table 3. Retrieval performance

We can compare this table with Figure 3 and see that the retrieval performance is strongly consistent with that of the segmentation. The same ranking is maintained for both segmentation and retrieval: the hybrid approach, the rule- and dictionary-based approach, and finally the statistic approach. This leads to the conclusion that Chinese texts should be segmented with a high quality segmentation process if one expects a high retrieval performance.

7. Future work

In this paper, we described a hybrid segmentation approach which makes use of both human-defined knowledge and statistical information. In comparison with other segmentation approaches, this approach is marked by its high flexibility: it can cover both the statistical approach and the rule-based approach by varying the default probability assigned to manually established lexical items. The hybrid framework allows us to see that statistical information and man-defined lexical knowledge represent two extreme cases in segmentation, but they are not incompatible, thus can be combine in a single process.

We also tried to adapt a general information retrieval system, SMART, to retrieve segmented Chinese texts. Our adaptation shows the feasibility of using IR systems designed for Indo-European languages to Chinese.

As one of the subjects for our future work, we plan to enhance our segmentation process by incorporating more heuristic rules, in particular, for dealing with proper names. In a previous work, we investigated this subject [27] but it has not been integrated into the present implementation.

On Chinese text retrieval, there is a lot to be done. In an attempt to obtain better recall we will investigate the application of word stemming to Chinese words in such a way that comparison becomes possible between 大众 and 大众性, 现代 and 现代的. This goal may be achieved by considering heuristic morphological rules as for segmentation.

References

[1] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*: McGraw-Hill, 1983.

[2] 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.

[3] 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.

[4] C.-K. Fan and W.-H. Tsai, "Automatic word identification in Chinese sentences by the relaxation technique," *Computer Processing of Chinese and Oriental Languages*, vol. 4, pp. 33-56, 1988.

[5] 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.

[6] 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.

[7] K.-J. Chen and S.-H. Kiu, "Word identification for Mandarin Chinese sentences," *5th International Conference on Computational Linguistics*, 101-107, 1992.
[8] 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.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] Z.-D. Dong, "Chinese platform project of Chinese information processing and Chinese language research," *Communications of COLIPS*, vol. 3, pp. 79-88, 1993.
[16] J.-S. Chang and e. al., "Chinese word segmentation through constraint satisfaction and statistical optimization," *ROCLING-IV*, Taiwan, 147-165, 1991.
[17] 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.
[18] 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.
[19] T.-H. Chiang and e. al., "Statistical models for segmentation and unknown word resolution," *5th R.O.C. Computational Linguistics Conference*, 123-146, 1992.
[20] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, vol. 19, pp. 61-74, 1993.
[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] S. Bai, ""Semi-word" method for Chinese word segmentation," *International Conference on Chinese Computing*, Singapore, 304-309, 1994.
[23] R. Sproat, C. Shih, W. Gale, and N. Chang, "A stochastic finite-state word-segmentation algorithm for Chinese," *ACL'94*1994.
[24] Z. Wu and G. Tseng, "ACTS: An automatic Chinese text segmentation system for full text retrieval," *Journal of the American Society for Information Sciencé*, vol. 46, pp. 83-96, 1995.
[25] 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.
[26] C. Buckley, "Implementation of the SMART information retrieval system," Cornell University, Technical report 85-686, 1985.
[27] 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.