Chinese Text Retrieval Without Using a Dictionary

Aitao Chen, Jianzhang He, Liangjie Xu
School of Information Management and Systems
{aitao,jzhe,jack}@sims.berkeley.edu

Fredric C. Gey, Jason Meggs
UC Data Archive & Technical Assistance (UC DATA)
{gey,jason}@ucdata.berkeley.edu
University of California at Berkeley, CA 94720

Abstract

It is generally believed that words, rather than characters, should be the smallest indexing unit for Chinese text retrieval systems, and that it is essential to have a comprehensive Chinese dictionary or lexicon for Chinese text retrieval systems to do well. Chinese text has no delimiters to mark word boundaries. As a result, any text retrieval systems that build word-based indexes need to segment text into words. We implemented several statistical and dictionary-based word segmentation methods to study the effect on retrieval effectiveness of different segmentation methods using the TREC-5 Chinese test collection and topics. The results show that, for all three sets of queries, the simple bigram indexing and the purely statistical word segmentation perform better than the popular dictionary-based maximum matching method with a dictionary of 138,955 entries.

1 Introduction

The written Chinese text has no delimiters to mark word boundaries, it consists of a string of characters and punctuation. The first step toward word-based indexing is to break a sequence of characters into words. The process of breaking a string of characters into words is called word segmentation. Word segmentation is known to be a difficult task because accurate segmentation of written Chinese text may require deep analysis of the sentences. Even Chinese speakers may disagree over how a sentence should be segmented because of the lack of a clear-cut definition on what constitutes a Chinese word. Some practical and popular word segmentation methods use dictionaries (lexicon), the simplest one being just a list of Chinese words. The dictionary coverage of words can have a significant impact on the accuracy of word segmentation. It is virtually impossible to list all the Chinese words in a dictionary because the set of words is open-ended. The construction of a comprehensive dictionary is itself a difficult task.

Another group of word segmentation methods uses the lexical statistics of the Chinese characters in corpora to mark the word boundaries. The lexical statistics may include the occurrence frequency of a character in text corpora, and the co-occurrence frequency of two or more characters in text corpora. What makes the statistical word segmentation approaches appealing is that they do not require a comprehensive dictionary to mark word boundaries.

It is generally believed that a comprehensive Chinese dictionary or lexicon is needed for a Chinese text retrieval system to perform well. We want to know if Chinese text retrieval systems without dictionaries can achieve comparable performance as those that do word segmentation using dictionaries. We implemented several purely statistical segmentation methods and dictionary-based methods to study the retrieval effectiveness of different word segmentation methods using the TREC-5 Chinese track data, which includes the test collection, 28 topics and relevance judgments.

The rest of the paper is organized as follows. Section 2 presents the characteristics of Chinese characters and words and the distribution of Chinese words; In section 3, we briefly review different approaches to Chinese word segmentation; section 4 discusses our implementation of several methods of word segmentation with an emphasis on the statistical method using mutual information; section 5 describes the experiment setup, including the test collection, the topics and the formula used in the experiments; section 6 presents and summarizes the evaluation results of using different word segmentation methods; in section 7 we discuss the results; and section 8 concludes the paper.

SIGIR 97 Philadelphia PA, USA

2 Chinese Characters and Words

A single Chinese character is called *hanzi* in Chinese. Some individual hanzi can function as words, in that case they are called free morphemes, whereas others cannot function as words by themselves, they must combine with other hanzi to form words. Those hanzi characters are called bound morphemes [12]. A Chinese word can be a single character, or two or more characters. However most Chinese words consist of two characters [2, p.194]. According to the *Frequency Dictionary of Modern Chinese* [7] (cited in [8]), among the top 9,000 most frequent words, 26.7% are unigrams, 69.8% are bigrams, and 2.7% are trigrams, and according to Liu's study [13], 5% words are unigrams, 75% are bigrams, 14% are trigrams, and 6% are words of four or more characters.

3 Previous works on word segmentation

Chinese word segmentation is the process of breaking a string of characters into words. The word segmentation problem has been an active research topic for over a decade. There is a large body of literature on Chinese word segmentation [1, 3, 6, 9, 11, 14, 16, 18, 17, 19, 21, 22] (see also references in [20]), which is the first step toward natural language processing, or text retrieval. The process of part-of-speech tagging of words, semantic tagging of words, syntactic analysis and semantic analysis of sentences all depend on the accurate segmentation of sentences into words. Over the years, many approaches have been developed to segment Chinese text into words. The methods can be broadly grouped into three categories: (1)statistical methods [6, 14, 17],(2) heuristic rule-based methods [3], and (3) combination of the statistical and rule-based methods [15, 18, 19, 22]. Wu and Tseng [21] provide a recent survey of Chinese word segmentation algorithms for text retrieval.

Sproat and Shih [17] describes a purely statistical approach to segmenting sentences into words. The approach utilizes no dictionary to decide the word boundaries, instead it groups two adjacent characters into a word based on the association strength between the two adjacent characters in the input sentence. The measure of association between two adjacent characters is the mutual information statistic [4]. This approach is simple, efficient and easy to implement, however it has the limitation of recognizing only words of one or two characters long. Luo [14] proposes an iterative process of segmenting Chinese sentences. The process of word segmentation and language modeling are alternatively carried out.

One popular dictionary-based word segmentation approach is the maximum matching method (also called the longest matching). The basic idea is that an input sentence should be segmented in such a way that the number of words produced should be minimum. It is not uncommon that an input sentence can be segmented in more than one way, though there is usually only one correct segmentation. A lexicon or dictionary is required to perform maximum matching. The coverage of a dictionary is essential to the quality of segmented text. If a dictionary contains only a small portion of the words in the corpus to be segmented, many words are treated as unknown. The handling of unknown words in the process of segmentation is a difficult task. The maximum matching works either from the beginning to the end of input phrase, or vice versa. The forward maximum matching groups the longest initial sequence of characters that matches a dictionary entry as a word, then starts at the next character after the most recently found word and repeats the process until the end of the input sentence. The backward maximum matching works from the end of a sentence toward the beginning. The entries in the dictionary include words, phrases, idioms, proper names, etc. A variation of the maximum matching method is the minimum matching, or shortest matching, which treats as a word the shortest initial string of characters that matches a dictionary entry.

Leung and Kan [11] propose a method of automatic learning rules that can be incorporated in a word segmentation system to achieve better accuracy. The rules are constructed based on the correlation between the incorrectly segmented strings and their contexts.

Instead of treating word segmentation as a preprocessing step, Gan et al. integrate the process of word segmentation into the task of sentence analysis; and Sun et al. tackle the tasks of word segmentation and part-of-speech tagging together utilizing general dictionaries as well as specialized dictionaries (e.g. proper name dictionaries), lexical statistical information and heuristic rules.

Numerous authors [20, 15] believe that words should be the smallest indexing unit. Thus word segmentation becomes the first step to Chinese text retrieval, and the accuracy of a word segmentation method is crucial to the effectiveness of Chinese text retrieval.

4 Indexing Techniques

We have carried out a series of experiments using the TREC-5 Chinese collection and 28 Chinese topics. We want to know if the purely statistical indexing and the n-grams indexing techniques work as well as the word-based indexing method. We indexed the collection using unigram, bigram, trigram and statistical segmentation proposed by Sproat and Shih [17].

4.1 Unigram Indexing

The unigram is probably the simplest and most efficient approach to indexing Chinese written text. It breaks a sequence of Chinese characters into individual ones. Each individual character is an indexing unit. For example, the index terms produced from input character string $c_1c_2c_3c_4c_5$ are: c_1, c_2, c_3, c_4 and c_5. The vocabulary size is limited by the number of characters in the coding standard for Chinese characters GB2312-80, which has 6,763 *hanzi* characters.

4.2 Bigram Indexing

In bigram indexing, all adjacent pairs of hanzi characters in text become indexing terms. For bigrams generated in this way, the last character of the previous bigram overlaps the first character of the next bigram. For example, the indexing terms representing character string $c_1c_2c_3c_4c_5$ are: c_1c_2, c_2c_3, c_3c_4, c_4c_5.

4.3 Trigram Indexing

A trigram is a consecutive sequence of three hanzi characters. All trigrams in text become indexing terms in trigram indexing scheme. Any two adjacent trigrams have two characters in common. For example, the indexing terms for character string $c_1c_2c_3c_4c_5$ are: $c_1c_2c_3$, $c_2c_3c_4$ and $c_3c_4c_5$. Table 1 shows the n-gram indexing terms produced from the same text string. Table 2

sentence	$c_1c_2c_3c_4c_5c_6$
unigrams	c_1, c_2, c_3, c_4, c_5, c_6
bigrams	c_1c_2, c_2c_3, c_3c_4, c_4c_5, c_5c_6
trigrams	$c_1c_2c_3$, $c_2c_3c_4$, $c_3c_4c_5$, $c_4c_5c_6$

Table 1: n-gram indexing methods

shows the number of unique and total number of unigrams, bigrams and trigrams in the TREC-5 Chinese test collection, about one third of possible Chinese bigrams occur at least once in the Chinese collection.

n-gram	no. distinct n-grams	no. n-grams
unigrams	6,236	64,611,662
bigrams	1,393,488	54,362,319
trigrams	8,119,574	49,886,331

Table 2: n-gram size of TREC-5 Chinese collection

4.4 Statistical Indexing

Before the text is indexed, the following operations were performed:

1. Collect occurrence frequency in the collection for all Chinese characters occurring at least once in the collection,
2. Collect occurrence frequency in the collection for all Chinese bigrams occurring at least once in the collection,
3. Compute the mutual information for all Chinese bigrams (see below), and
4. Apply the algorithm as described in [17] to segment the text into words.

Mutual information measures the association of two events. The mutual information $I(x,y)$ between two events x and y is defined as

$$I(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

where $p(x)$ is the probability of observing event x, $p(y)$ the probability of observing event y, and $p(x,y)$ the probability of observing both events. If two events occur independently, the joint probability $p(x,y)$ would be close to the product of $p(x)$ and $p(y)$, thus the mutual information would be close to zero. On the other hand, if two events are strongly related, the joint probability $p(x,y)$ would have a much larger value than the product of $p(x)$ and $p(y)$, so $I(x,y)$ would be much bigger than zero; if two events occur complementarily, the mutual information value would be negative. Table 3 shows the occurrence frequency values and mutual information values for six Chinese bigrams found in the TREC-5 Chinese collection. In table 3, column $f(c1)$ is the occurrence frequency value of the first Chinese character of a bigram; column $f(c2)$ the occurrence frequency value of the second

bigrams	f(c1)	f(c2)	f(c1c2)	I(c1,c2)
淘汰(eliminate)	1,549	1,632	1,343	15.06
苹果(apple)	1,208	50,416	1,021	10.08
漂亮(beautiful)	1,445	6,301	859	12.57
非常(unusually)	37,579	50,257	7,157	7.93
如果(if)	57,975	50,416	10,884	7.91
不水(not water)	311,474	90,495	1	-8.76

Table 3: Mutual information of six Chinese bigrams

bigrams	f(c1)	f(c2)	f(c1c2)	I(c1,c2)
中国	615,222	925,353	228,090	4.69
国大	925,353	417,826	6,791	0.18
大陆	417,826	15,331	6,946	6.13
陆新	15,331	256,559	22	-1.46
新发	256,559	328,500	1,058	-0.30
发现	328,500	139,630	11,946	4.07
现的	139,630	2,017,405	4,340	-0.00
的油	2,017,405	26,690	676	-0.30
油田	26,690	24,869	2,412	7.87

Table 4: Mutual information values of bigrams.

Chinese character; column $f(c_1c_2)$ the occurrence frequency value of a bigram; and the last column, $I(c_1, c_2)$, is the mutual information. The first five bigrams — 淘汰(eliminate), 苹果(apple), 漂亮(beautiful), 非常(unusually), and 如果(if) — are genuine Chinese words; whereas the last one, 不水(not water), is not a Chinese word. As expected, the genuine words have mutual information values of being much bigger than zero, whereas the bigrams that are not words have mutual information values of being much smaller than zero.

The $p(c_1)$ in the mutual information definition is estimated by $f(c_1)/N$, $p(c_2)$ is estimated by $f(c_2)/N$, and the probability of observing two characters c_1 and c_2 occurring in the collection together in fixed order c_1c_2 is estimated by

$$p(c_1c_2) = p(c_1)p(c_2|c_1) = \frac{f(c_1)}{N}\frac{f(c_1c_2)}{f(c_1)} = \frac{f(c_1c_2)}{N}$$

where N is the collection size, which is 64,611,662 characters for the TREC-5 Chinese collection, and $f(c_1c_2)$ is the occurrence frequency value of bigram c_1c_2 in the collection. The example below illustrates the computation of mutual information for bigram 苹果(apple, or *ping2 guo3* in Chinese *pinyin* notation).

$$\begin{aligned} I(ping2guo3) &= \log_2 \frac{f(ping2guo3) * N}{f(ping2)f(guo3)} \\ &= \log_2 \frac{1021 * 64611662}{1208 * 50416} \\ &= 10.08 \end{aligned}$$

One of the titles of the TREC-5 Chinese topics is "中国大陆新发现的油田 (The newly discovered oil fields in China)." We will use this title as an example to show how the written Chinese text is segmented using the purely statistical method. First, the written text is broken into phrases — a consecutive sequence of Chinese characters is considered as a phrase. All the non-Chinese characters in the text are ignored. Second, each phrase is segmented as follows:

step	phrases	action
1	中国大陆新发现的油田	remove 油田(oil fields)
2	中国大陆新发现的	remove 大陆(mainland)
3	中国 新发现的	remove 发现(discover)
4	中国 新 的	stop

Table 5: Word segmentation process using mutual information.

1. Compute the mutual information values for all adjacent bigrams in a phrase,
2. Treat the bigram of the largest mutual information value as a word and then remove it from the phrase. The removal of the bigram may result in one or two shorter phrases,
3. Perform step 2 on each of the shorter phrases until all phrases consist of one or two characters.

Table 4 shows the mutual information values for all the bigrams in the phrase 中国大陆新发现的油田(The newly discovered oil fields in China). The first column shows all the bigrams generated from the input phrase; the second column, $f(c1)$, is the number of occurrences in the TREC-5 Chinese collection of the first character of a bigram; the third column, $f(c2)$, is the number of occurrences of the second character of a bigram in the same collection; the fourth column, $f(c1c2)$, is the number of occurrences of the bigram in the same collection; and the last column, $I(c1, c2)$, is the mutual information value of a bigram.

From table 4, we see the bigram 油田 (oil fields) has the largest mutual information value (7.87), so the bigram 油田(oil fields) is taken as a word, and removed from the phrase. The removal of the bigram 油田(oil fields) from the phrase results in one shorter phrase 中国大陆新发现的. Now the bigram 大陆(mainland) has the largest mutual information value (6.13) in the new phrase, so the two characters are grouped into a word, and is removed from the shorter phrase. The removal of the bigram 大陆(mainland) produces two smaller phrases 中国 (China) and 新发现的. The first phrase 中国(China) is not segmented further because it is two-character long, whereas the second phrase, however, is segmented further into three smaller new phrases: 新(new), 发现(discover), and 的(adjective marker). The segmented sentence becomes 中国 大陆 新 发现 的 油田, which is the correct segmentation for the original phrase. The process of segmenting the phrase 中国大陆新发现的油田 (The newly discovered oil fields in China) using mutual information values of all the bigrams in the phrase is illustrated in table 5.

4.5 Maximum Matching and Minimum Matching

We implemented the maximum matching method as well as the minimum matching method for Chinese word segmentation. Both methods require a dictionary (lexicon) to segment text. The matching direction can be forward or backward. The forward matching starts from the beginning of a phrase, and works toward the end, whereas the backward matching starts from the end of a phrase, then works toward the beginning of the phrase. While the maximum matching method groups the longest initial sequence of characters that matches a dictionary entry as a word, the minimum matching method treats the shortest initial sequence of characters that matches a dictionary entry as a word. The minimum matching method degenerates to the unigram segmentation when the dictionary contains an entry for each single character. The dictionary used for segmentation has 138,955 entries, including words, compounds, phrases, idioms, proper names etc. We made no attempt to identify and resolve the ambiguity of

text	中国大陆新发现的油田
unigrams	中 国 大 陆 新 发 现 的 油 田
bigrams	中国 国大 大陆 陆新 新发 发现 现的 的油 油田
trigrams	中国大 国大陆 大陆新 陆新发 新发现 发现的 现的油 的油田
statistical (mutual information)	中国 大陆 新 发现 的 油田
maximum matching (forward)	中国 大陆 新 发现 的 油田
correct segmentation	中国 大陆 新 发现 的 油田

Table 6: Indexing terms produced from the same text by different segmentation methods.

step	phrases	action
1	中国大陆新发现的油田	remove 中国(China)
2	大陆新发现的油田	remove 大陆(mainland)
3	新发现的油田	remove 新(newly)
4	发现的油田	remove 发现(discover)
5	的油田	remove 的 (marker)
6	油田	remove 油田(oil fields)
7		stop

Table 7: Maximum matching (forward) segmentation process.

word segmentation, and to handle unknown words specially. The dictionary is stored in memory as a trie structure. The matching process starts from the root of the trie structure, and traverses down the trie one character at a time if the next character in the input string matches a character on the next level of the trie structure. The longest initial sequence of characters that does not match any entry in the dictionary is treated as an unknown word. Table 6 shows the index terms produced from phrase 中国大陆新发现的油田(The newly discovered oil fields in China) with various indexing techniques.

Table 7 shows the process of segmenting phrase 中国大陆新发现的油田(The newly discovered oil fields in China) using the dictionary-based maximum matching method.

5 The Experiments

5.1 The Test Collection

The test collection used in all the runs of experiments reported in this paper is the TREC-5 Chinese collection, it has two subcollections: *People's Daily* and *Xinhua News Agency*. The People's Daily subcollection contains newspaper articles published between 1991 and 1993 in China. Table 8 summarizes the statistics about the two subcollections. Column two is the number of

subcollections	number of documents	avg length (characters)	length std dev
People's Daily	139,801	369.68	684.98
Xinhua News Agency	24,988	517.59	326.62

Table 8: TREC-5 Chinese Test Collection Statistics.

documents in each subcollection, column three the average document length in characters and column four the standard deviation of document length. Document length is given in number of Chinese characters rather than words because of the difficulties of accurately determining words boundaries in Chinese text.

5.2 The Topics

We used all 28 topics of TREC-5 Chinese track. All the original topics have three fields: *title*, *description*, and *narrative*. The *description* field provides a list of key concept terms for a topic. All the Chinese topics are also translated into English, however the English translation is not utilized in all the experiments reported in this paper. We will call the original set of topics the long queries, and the set of topics consisting of the *title* field only the short queries. The set of manual queries were derived from the original topics and the Chinese test collection when we were working on TREC-5 Chinese track.

Topic number 14 of TREC-5 Chinese track is included below.

Title: 中国的爱滋病例 (Cases of AIDS in China)

Description: 中国，云南，爱滋病，HIV，高危险群患者，注射器，病毒。
(China, Yunnan, AIDS, HIV, high risk group, syringe, virus)

Narrative: 相关文件应当包括中国那些地区的爱滋病例最多，爱滋病毒在中国是如何传播的，以及中国政府如何监测爱滋病并控制它的传染。
(A relevant document should contain information on the areas in China that have the highest AIDS cases, how the AIDS virus was transmitted, and how the Chinese government combats AIDS problem).

The derived short query from the above long query is

Title: 中国的爱滋病例 (Cases of AIDS in China).

And the manually reformulated query for the long query is

term (translation)	qf	term (translation)	qf
艾滋病 (AIDS)	15	我国 (our country)*	1
中国 (China)	11	感染 (infect)*	6
监测 (monitor)	6	注射器 (syringe)	1
病毒 (virus)	6	预防 (prevent)*	5
HIV (HIV)	4	传染 (infect)	1
患者 (patient)	4	防治 (prevent and cure)*	4
病人 (patient)*	2	病例 (medical cases)	3
云南 (Yunnan)	3	预防为主 (prevention first)*	3
性病 (STD)*	1	吸毒 (drug use)*	1
发现 (found)*	6		

The second and fourth columns are the within-query term frequencies. The terms marked with '*' are new query terms that were manually selected from the the test collection. The important concept terms are emphasized by assigning higher weights (i.e. larger frequencies). The set of manual queries were manually segmented into words. The stop words and the unimportant words were excluded from the queries. The construction of manual queries can be characterized by an iterative process that

consists of three steps: (1) do a trial run using the current query; (2) examine the top-ranked documents and manually select the terms that seem to be promising from the top documents; (3) add the chosen terms from the previous step to the current query and assign weights manually to the new terms to form a new query. The inclusion of new terms may cause the weights of some existing terms to be readjusted. The process started with the original topics and was repeated a few times for each topic. On the average, we spend about 2.8 hours on each topic.

5.3 The formula

For all the runs of experiments reported in this paper, we used the Berkeley adhoc formula developed by Cooper [5] at TREC-2. The adhoc retrieval results produced at Berkeley have shown that the formula is robust for long queries and manually reformulated queries, and the results of applying the same formula to the TREC-5 Chinese collection further demonstrated the robustness of the formula [10]. The formula was fitted on training data using logistic regression. The logodds of relevance of document D to query Q is given by

$$\log O(R|D,Q) = -3.51 + \frac{1}{\sqrt{N}+1}\Phi + 0.0929 * N \qquad (1)$$

$$\Phi = 37.4 \sum_{i=1}^{N} \frac{qtf_i}{ql+35} + 0.330 \sum_{i=1}^{N} \log \frac{dtf_i}{dl+80} - 0.1937 \sum_{i=1}^{N} \log \frac{ctf_i}{cf} \qquad (2)$$

where

N is the number of terms common to both query and document,
qtf_i is the occurrence frequency within a query of the ith match term,
dtf_i is the occurrence frequency within a document of the ith match term,
ctf_i is the occurrence frequency in a collection of the ith match term,
ql is query length (number of terms in a query),
dl is document length (number of terms in a document), and
cf is collection length, i.e. the number of occurrences of all terms in a test collection.

The summation in equation (2) is carried out over all the terms common to query and document. Although the formula was derived from English collection, it performed well on Chinese collection as the results in [10] had shown.

5.4 Dictionary and stop list

The stop list has 825 entries. It includes pronouns, determiners, prepositions, adverbs, conjunctions, etc. The dictionary has 138,955 entries, which include words, phrases, compounds, idioms, proper names, etc. About 60,000 entries were manually selected from the TREC-5 Chinese collection and were added to a dictionary of about 80,000 entries obtained from a website. Table 9 presents the distribution of the dictionary entries by length (in characters).

5.5 Evaluation

In each run of experiment, for each query, the top 1,000 documents were retrieved from the Chinese collection of 164,789 documents, the retrieved documents were ranked in decreasing order by the probability of relevance estimated using equation (1). There are 2,182 relevant documents in total for all 28 topics. For all runs, we calculate the 11-point recall-precision values using the TREC evaluation program written by Chris Buckley. Our base run uses the index file created from the text segmented using the forward maximum matching method.

length (chars)	no. entries	percent (%)
1	1,193	0.8585
2	77,837	56.0160
3	31,461	22.6411
4	25,672	18.4750
5	1,793	1.2903
6+	999	0.7189
total	138,955	100

Table 9: The distribution of dictionary entries by length

6 Results for different query sets

To evaluate the retrieval effectiveness of different indexing techniques, we created eight index files. The unigram, bigram and trigram indexing involves no dictionary, and no stop words[1]. Table 10 summarizes the methods used in creating the index files. The topics are indexed the same way as the collection. We used three sets of queries: short queries, long queries and manual queries. Eight runs of experiments are produced for each set of queries, one run against each index file. The formula used in all runs is the Berkeley adhoc formula developed in TREC-2.

6.1 The long queries

Each query in the set of long queries has the title, description and narrative fields. The queries in this set are the original queries used in TREC-5 Chinese track. Table 11 summarizes the results of using the set of long topics. In comparison to the base run, the 11-point average of precision over all 28 topics for bigram indexing is 3.34% better than the base run result, while the average of precision for statistical word segmentation using mutual information is 5.23% better than the base run result. The results for unigram and trigram indexing are much worse than the base run result. Manual queries were constructed by adding new words and changing the weights (frequency) of words after examining the top documents from an initial search using the automatic long query. For example in the query 14: *Cases of AIDS in China*, the TREC-5 topic description uses the word 愛滋病 for AIDS. This word is the common (familiar) term for AIDS used in Hong Kong and Taiwan. Used in this form as a query word, it only indexes 5 documents of the TREC-5 Chinese collections. In other documents which discuss AIDS, the official term 艾滋病 (phonetically similar to the English pronunciation) is used.

Table 12 shows the results of running the manually reformulated queries against all eight index files. The manually reformulated queries perform around 27 percent better than the automatically constructed long queries. All methods except trigrams and minimum matching achieve about 45 percent overall average precision. It is interesting to note that the bigram method retrieves more overall relevant documents (although only another 16 documents, 0.8% more than the baseline run). The poor performance of trigram indexing was mainly due to the fact that most of query terms in the set of manual queries are two-character words, while the indexing terms in the trigram index file are three-character strings.

6.2 The short queries

Short topic runs have been a feature of recent TREC evaluations, as they more accurately mirror actual user behavior in commercial information retrieval services. In our experiments, the short queries were all automatically constructed from the single-sentence *title* field of the TREC-5 Chinese topics. These queries average 12.5 characters long versus 107.0 characters for the long

[1]The bigram and trigram methods are not truly word segmentation methods because of the overlap between two adjacent bigrams or trigrams. In this paper, we loosely consider them as approaches to word segmentation.

	index file	indexing terms	segmentation/indexing method	dictionary and stop-list used
1	unigram	unigrams	unigram	none
2	bigram	bigrams	bigram	stop-list only
3	trigram	trigrams	trigram	stop-list only
4	mi	bigrams, unigrams	statistical with mutual information	stop-list only
5	max (f)	words, phrases	maximum matching (forward)	both
6	max (b)	words, phrases	maximum matching (backward)	both
7	min (f)	words, phrases	minimum matching (forward)	both
8	min (b)	words, phrases	minimum matching (backward)	both

Table 10: Index files

Recall	unigram	bigram	trigram	mi	max (f)	max (b)	min (f)	min (b)
0.00	0.7751	0.7504	0.6962	0.7696	**0.8000**	0.7966	0.7404	0.7265
0.10	0.5609	0.6241	0.5006	0.6500	**0.6465**	0.6414	0.5543	0.5611
0.20	0.4076	0.5243	0.3600	0.5355	**0.5283**	0.5028	0.4336	0.4432
0.30	0.3400	0.4778	0.2932	0.4705	**0.4308**	0.4518	0.3595	0.3734
0.40	0.2904	0.4375	0.2546	0.4324	**0.3841**	0.4085	0.3049	0.3245
0.50	0.2486	0.3864	0.2153	0.3872	**0.3455**	0.3671	0.2569	0.2903
0.60	0.2050	0.3295	0.1815	0.3346	**0.2947**	0.3131	0.2216	0.2351
0.70	0.1576	0.2749	0.1586	0.2843	**0.2439**	0.2678	0.1657	0.1912
0.80	0.0982	0.2173	0.1142	0.2353	**0.1891**	0.2017	0.1221	0.1217
0.90	0.0300	0.1241	0.0581	0.1378	**0.1051**	0.1105	0.0819	0.0778
1.00	0.0031	0.0108	0.0091	0.0208	**0.0282**	0.0341	0.0197	0.0118
average precision	0.2609	0.3677	0.2405	0.3744	**0.3558**	0.3465	0.2738	0.2862
	-26.67%	3.34%	-32.40%	5.23%	**baseline**	-2.61%	-23.04%	-19.56%
relevant retrieved	1614	2017	1735	1948	**1910**	1825	1731	1693

Table 11: Average precision of long queries using different segmentation methods

topic queries and 249.7 characters for the manually reformulated queries.

Short queries have significantly poorer performance than their long query counterparts. In this case, however, automatic dictionaryless segmentation (except for trigrams) outperforms all flavors of our dictionary-based segmentation – mutual information statistical segmentation overall precision of .2850 is 21 percent better than the base run. Table 13 shows the evaluation results of using the short queries.

7 Discussion

In unigram indexing, the indexing terms are the single characters. As mentioned before, some characters not only can function as words by themselves, but also can combine with other characters to form words that can have different meanings. For example, when character 毒(poison) combines with 气(gas), it forms the word 毒气(poison gas); when it combines with the character 病(disease), it forms the word 病毒, which mean virus; whereas when it combines with the character 品, it forms the word 毒品, which means drugs. Single characters are much more ambigous than words. We believe that the ambiguity of single characters (i.e. unigrams) are the main factor contributing to the poor performance of unigram for the the set of long queries. Another factor that might have degraded the performance of unigrams for the set of long queries is that the single-character function words were not removed from the queries during indexing. Some single-character function words are 将(will), 的(of), 与(and), and 在(in). As Nie et al argue in [15], there are a number of reasons why unigram indexing is not appropriate for text retrieval.

Despite the problems of unigram indexing, the average precisions for the set of short queries and the set of manually reformulated queries are surprisingly good in comparison with the results of the dictionary-based maximum matching. In the manually reformulated queries, all the function words are removed and in the short queries, the number of single-characters is small because of the short length of the topics.

Since most Chinese words consist of two characters, we expect that the bigram indexing and the mutual information segmentation would produce better indexing terms, and thus achieve better retrieval performance in comparison to unigram indexing and trigram indexing. Most of the indexing terms produced by mutual information segmentation probably are actual Chinese words.

For all three sets of queries, trigram indexing performed considerably worse than the baseline. The major reason is that many of the words representing the key concepts in the topics are two-character words, hence they are excluded in the queries since only the trigrams are used. It is clear that the constraint that all indexing terms are trigrams is too stringent and thus detrimental to good performance. In general, the number of unique trigrams in a corpus is much larger that that of unique bigrams. As shown in table 2, the TREC-5 Chinese collection has 8,119,574 trigrams and 1,393,488 bigrams. Due to the large number of unique trigrams in a collection, trigram indexing probably introduces considerably more noise since many of the trigrams are not meaningful.

All dictionary-based word segmentation algorithms face two problems: (1) unknown words; (2) identification and subsequent resolution of segmentation ambiguities. In our implementations of the dictionary-based segmentation algorithms, we took the simplest approach to those problems by ignoring them. The dictionary-based segmentation methods we implemented always produce one result for each input sentence. The unmatched string of characters are grouped together as one word, which may or may not be a true Chinese word.

It seems that the direction of matching, forward or backward, does not much affect the retrieval effectiveness. For the set of short queries, the minimum matching works better than maximum matching, whereas for the set of long and manual queries, the maximum matching outperforms substantially the minimum matching method.

We believe that many words in the TREC-5 Chinese collection are not included in our dictionary, especially the proper names. We extracted 287 university and college names, and 627 company names from the Chinese collection using simple pattern matching techniques. Only 21 out of the 287 names are included in our dictionary, and only 14 out of the 627 company names are in our dictionary. The common unknown words include personal names,

Recall	unigram	bigram	trigram	mi	max (f)	max (b)	min (f)	min (b)
0.00	0.8624	0.8309	0.7008	0.8372	**0.8551**	0.8433	0.8154	0.7961
0.10	0.6880	0.6938	0.4720	0.6831	**0.7304**	0.7059	0.6590	0.6372
0.20	0.5757	0.6242	0.3464	0.6298	**0.6429**	0.6378	0.5679	0.5279
0.30	0.5286	0.5684	0.3005	0.5824	**0.5787**	0.5716	0.5093	0.4841
0.40	0.4756	0.5119	0.2652	0.5292	**0.5105**	0.5074	0.4570	0.4448
0.50	0.4263	0.4598	0.2349	0.4560	**0.4583**	0.4575	0.4060	0.4036
0.60	0.3829	0.4041	0.2082	0.4054	**0.4146**	0.4111	0.3544	0.3660
0.70	0.3404	0.3551	0.1528	0.3631	**0.3514**	0.3487	0.2873	0.3098
0.80	0.2809	0.3064	0.1326	0.3116	**0.2894**	0.2833	0.2253	0.2482
0.90	0.1859	0.2261	0.0868	0.2285	**0.2314**	0.2327	0.1544	0.1494
1.00	0.0356	0.0823	0.0122	0.0625	**0.0499**	0.0674	0.0340	0.0344
average precision	0.4203	0.4522	0.2397	0.4533	**0.4519**	0.4481	0.3937	0.3904
	-6.99%	0.06%	-46.95%	0.31%	**baseline**	-0.84%	-12.87%	-13.61%
relevant retrieved	2036	2088	1711	2064	**2033**	2020	2022	2008

Table 12: Average precision of manually expanded queries using different segmentation methods

Recall	unigram	bigram	trigram	MI	Max (f)	Max (b)	Min (f)	Min (b)
0.00	0.7325	0.6783	0.5937	0.7215	**0.6779**	0.6511	0.6401	0.6473
0.10	0.5149	0.4594	0.3433	0.5157	**0.4701**	0.4543	0.4754	0.4567
0.20	0.4443	0.4098	0.2296	0.4444	**0.3869**	0.3725	0.3971	0.3705
0.30	0.3835	0.3715	0.1962	0.3948	**0.3142**	0.3246	0.3469	0.3242
0.40	0.3206	0.3430	0.1709	0.3598	**0.2652**	0.2652	0.3033	0.2949
0.50	0.2690	0.2811	0.1477	0.3107	**0.2150**	0.2158	0.2561	0.2574
0.60	0.2156	0.2444	0.1252	0.2492	**0.1689**	0.1577	0.2065	0.2149
0.70	0.1802	0.1885	0.1064	0.1748	**0.1294**	0.1136	0.1663	0.1647
0.80	0.1331	0.1189	0.0889	0.0859	**0.0914**	0.0819	0.0996	0.1112
0.90	0.0529	0.0557	0.0337	0.0437	**0.0433**	0.0266	0.0494	0.0549
1.00	0.0040	0.0051	0.0013	0.0085	**0.0055**	0.0050	0.0133	0.0128
average precision	0.2770	0.2687	0.1636	0.2849	**0.2346**	0.2250	0.2531	0.2465
	18.07%	14.53%	-30.26%	21.44%	**baseline**	-4.09%	7.88%	5.07%
relevant retrieved	1647	1787	1432	1641	**1588**	1529	1555	1536

Table 13: Average precision of short queries using different segmentation methods

place names, company names, transliterated names, names of new products, abbreviation of full names, etc. Further research is undertaken to see if a dictionary of about 300,000 entries will alter our conclusions.

The results in tables 11, 12 and 13 show that the bigram indexing and mutual information-based segmentation outperform the popular dictionary-based maximum matching. It would be interesting to see if the conclusion still holds if the collection were perfectly segmented.

8 Conclusions and Future Work

The focus of this experimental research has been on the performance of different Chinese language segmentation methods, with particular reference to comparison of dictionaryless versus dictionary-based methods. The TREC-5 collection is the first large-scale Chinese language test collection for which relevance judgments have been developed for a general query set. Its 28 queries against 164,789 documents provides an ample background against which to compare multiple segmentation algorithms.

Results of three different query sets (short single-sentence queries, lengthy topic descriptions, and manually reformulated queries) uniformly show that dictionaryless bigram and mutual information measure statistical segmentation perform at least as well, and in some cases outperform dictionary-based methods. For short queries, unigrams perform well, but for longer queries the unigram approach deteriorates markedly. The conventional wisdom is that segmentation based upon an authenticated dictionary will perform better than purely mechanical segmentation. Our results show otherwise for a large but incomplete dictionary of 138,955 Chinese words.

9 Acknowledgments

A portion of this work was supported by grant NSF IRI-9630765 from the Database and Expert Systems program of the Computer and Information Science and Engineering Directorate of the National Science Foundation.

We would like to thank the anonymous referees for their constructive comments.

The current address for Liangjie Xu is: CoreTech Group, Excite Inc. 555 Broadway Ave., Redwood City, CA 94063, USA

References

[1] Chao-Huang Chang and Cheng-Der Chen. A Study on Integrating Chinese Word Segmentation and part-of-speech tagging. *Communications of the Chinese and Oriental Languages Information Processing Society*, 3:69–77, 1993.

[2] Yuen Ren Chao. *A Grammar of Spoken Chinese*. University of California Press, Berkeley, 1968.

[3] Keh-Jiann Chen and Shing-Huan Liu. Word Identification for Mandarin Chinese Sentences. In *Proceeding of COLING*, pages 23–28, August 1992.

[4] K. Church and P. Hanks. Word Association Norms, Mutual Information, and Lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83. Association for Computational Linguistics, 1989.

[5] W. S. Cooper, A. Chen, and F. C. Gey. Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 57–66, March 1994.

[6] Charng-Kang Fan and Wen-Hsiang Tsai. Automatic Word Identification in Chinese Sentences by the Relaxation Technique. *Computer Processing of Chinese and Oriental Languages*, pages 33–56, November 1988.

[7] FDMC. *Xiandai hanyu pinlu cidian (Frequency dictionary of modern Chinese)*. Beijing Language Institute Press, 1986.

[8] Pascale Fung and Dekai Wu. Statistical augmentation of a Chinese machine-readable dictionary. In *Second Annual Workshop on Very Large Corpora*, pages 33–56, 1994.

[9] Kok-Wee Gan, Martha Palmer, and Kim-Teng Lua. A Statistically Emergent Approach for Language Processing: Application to Modeling Context Effects in Ambiguous Chinese Word Boundary Perception. *Computational Linguistics*, pages 531–553, 1996.

[10] F. C. Gey, A. Chen, J. He, L. Xu, and J. Meggs. Term Importance, Boolean Conjunct Training, Negative Terms, and Foreign Language Retrieval: Probabilistic Algorithms at TREC-5. In D. K. Harman, editor, *Text Retrieval Conference (TREC-5)*, 1996.

[11] Chi-Hong Leung and Wing-Kay Kan. A Statistical Learning Approach to Improving the Accuracy of Chinese Word Segmentation. *Literary and Linguistic Computing*, pages 87–92, 1996.

[12] Charles N. Li and Sandra A. Thompson. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press, Berkeley, California, 1981.

[13] Y. Liu. New Advances in Computers and Natural Language Processing in China. *Information Science (In Chinese)*, 8:64–70, 1987.

[14] Xiaoqiang Luo and Salim Roukos. An Iterative Algorithm to Build Chinese Language Models. In *Proceedings of ACL'96*, pages 139–143, 1996.

[15] Jian-Yun Nie and Martin Brisebois. On Chinese Text Retrieval. In *SIGIR*, pages 225–233, 1996.

[16] Jian-Yun Nie, Marie-Louise Hannan, and Wanying Jin. Combining Dictionary, Rules and Statistical Information in Segmentation of Chinese. *Computer Processing of Chinese and Oriental Languages*, pages 125–143, 1995.

[17] Richard Sproat and Chilin Shih. A Statistical Method for Finding Word Boundaries in Chinese Text. *Computer Processing of Chinese and Oriental Languages*, 4:336–351, March 1990.

[18] Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. *Computational Linguistics*, 22:377–404, September 1996.

[19] Maosong Sun, Dayang Shen, and Changning Huang. CSeg&Tag1.0: A Practical Word Segmenter and POS Tagger for Chinese Texts. In *Proceedings of the Fifth Applied Natural Language Processing Conference*, pages 119–126, 1997.

[20] Zimin Wu and Gwyneth Tseng. Chinese Text Segmentation for Text Retrieval: Achievements and Problems. *Journal of the American Society for Information Science*, 44:532–542, October 1993.

[21] Zimin Wu and Gwyneth Tseng. ACTS: An Automatic Chinese Text Segmentation System for Full Text Retrieval. *Journal of the American Society for Information Science*, 46:83–96, March 1995.

[22] Ching-Long Yeh and Hsi-Jian Lee. Rule-Based Word Identification for Mandarin Chinese Sentences - A Unification Approach. *Computer Processing of Chinese and Oriental Languages*, 5:97–118, March 1991.