Finding the Better Indexing units for Chinese Information Retrieval

Hongzhao He
Department of Computer Science and Engineering, Tianjin University
Weijin Road, Tianjin University
Tianjin, China, 300070
hehongzhao@hotmail.com

Pilian He
Department of Computer Science and Engineering, Tianjin University
Weijin Road, Tianjin University
Tianjin, China, 300070
plhe@tju.edu.cn

Jianfeng Gao
NLC, Microsoft Research Asia
No.49, Zhichun Road, Haidian District
Beijing, China, 100080
jfgao@microsoft.com

Changning Huang
NLC, Microsoft Research Asia
No.49, Zhichun Road, Haidian District
Beijing, China, 100080
cnhuang@microsoft.com

Abstract

In the processing of Chinese documents and queries in information retrieval (IR), one has to identify the units that are used as indexes. Words and n-grams had been used as indexes in several previous studies, which showed that both kinds of indexes lead to comparable IR performances. In this study, we carried out more experiments to find the better way to index Chinese texts. First, we investigated the inpacts on IR performance of the accuracy of word segmentation. Second, fifteen different groups of indexing units, which were the possible combination of words and character n-grams, were discussed detailedly. Experiments showed that better segmentation results in better IR performances, and a combination of words with uni-grams is the better choice to index Chinese texts for IR.

Introduction

It is well known that the major difference between Chinese information retrieval (IR) and IR in European languages lies in the absence of word boundaries in sentences. As Chinese sentences are written as continuous character strings, a preprocessing has to be done to segment sentences into shorter units that may be used as indexes. Those units may be of two kinds: words or character n-grams (simplifed as n-grams hereafter). In the previous studies, several experiments had been carried out using these two kinds of indexing units (Nie et al, 1996) (Kwok et al, 1996) (Chen et al, 1997) (Hsiao, 1997) (Gao et al, 2001a) (Shi et al, 2001). It turns out that these two kinds of units are effective in Chinese IR.

However, several problems have not been fully investigated: Does the accuracy of word segmentation have a significant impact on IR performance? Is it worthwhile to combine words with n-grams in Chinese IR? How should this be done? These are the questions we will examine in this study. A series of tests will be conducted on TREC and NTCIR-2 collection. This is a step forward to find a good way to index Chinese texts.

The remainder of this paper is organized as follows: In Section 1, we provide a brief survey of Chinese word segmentation. In Section 2, we discuss in detail of Chinese IR indexing units. In section 3, experimental results are presented. Finally, we present our conclusion.

1 Chinese Word Segmentation

1.1 Chinese Language

Chinese language is based on characters. There are 6763 frequently used Chinese characters.

Each Chinese word is a semantic concept that is about 1.6 characters on average. But there is no standard lexicon (or dictionary) of words -- linguists may agree on some tens of thousands of words, but they will dispute tens of thousands of others.

Furthermore, sentences are written without spaces between words. So a sequence of characters will have many possible segmentation in the word tokenization stage.

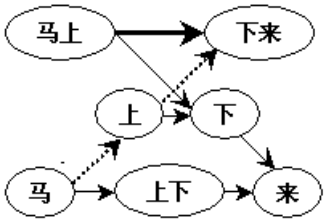

Figure 1: The word graph of Chinese sentence "马上下来"(Gao et al, 2001b)

Figure 1 shows the tokenization of a simple sentence with only four characters. Here, these four characters can be segmented in five ways into words. For example, the dotted path represents "*dismounted a horse*", and the bolded path represents "*immediately coming down*". This figure also shows seven possible "words", some of which (e.g., 上下) might be disputable on whether they should be considered "words".

Large amount of methods of Chinese word segmentation have been proposed to deal with these kinds of difficuties of Chinese.

1.2 Previous Methods for Chinese Word Segmentation

The segmentation of Chinese sentences into words requires linguistic knowledge. Several types of knowledge may be used: manually constructed dictionary which stores a set of known words, heuristic rules on word formation, or some statistical measures based on cooccurrences of characters. These three types of knowledge may be combined in different ways. For example, an approach based on a dictionary often uses also a set of heuristic rules. A statistical approach may also incorporate a set of heuristic rules.

Various experiments have been carried out on different segmentation approaches in the past 16 years. There is no one single approach shown to be clearly superior to the others. Most elaborated approaches can achieve a segmentation accuracy of over 90%. This performance has been believed to be sufficient for IR.

However, does the accuracy of word segmentation have a significant impact on IR performance? In this paper, we will use a special word segmentation system to investigate it.

1.3 Song's Segmentation System

Song's Segmentation System (System S) is a knowledge based system developed by Prof. Song from Beijing University of Language and Culture. The system is one dictionary-based segmentation complemented by a set of heuristic rules (Song, 2001). The basic algorithm is the longest-matching word segmentation algorithm. Several heuristic rules were used to identify such words as proper noun (e.g. 中国移动 – China Mobile), date expressions (e.g. 一九三四年 – year 1934), suffix structures (e.g. 使用者– user), etc. User can choose different rules they need to get different accuracy word segmentation results by selecting the appropriate switches on the user interface. Thus it is possible for us to investigate the impacts on Chinese IR of those different accuracy word segmentation results.

2 Chinese IR Indexing units

2.1 Different Indexing units for Chinese IR

Generally speaking, there are 2 chategories of Chiense IR indexing units: (1) n-grams and (2) words.

The advantage of n-grams is that it does not require any linguistic knowledge. This is the main reason for using n-grams in Chinese and other Asian languages (Lee and Ahn, 1996) (Ogawa, 1995). A string is simply cut down into units of fixed length. Usually, one uses uni-grams (or characters) and/or bi-grams. For example, a string ABCD (where each letter represents a Chinese character) can be segmented into bi-grams AB BC CD, or uni-grams A B C D.

Words have been the basic units of indexing in traditional IR, which requires linguistic knowledge as mentioned in Section 1.2.

2.2 Possible Indexing units for Chinese IR

There are several kinds of indexing units and their combinations for Chinese IR. In this paper, we divided possible indexing units into three groups of indexing units: (1) basic units; (2) combination of two basic units and (3) combination of more than two basic units.

Four kinds of basic units are investigated in this paper: (1) uni-grams; (2) words; (3) bi-grams and (4) tri-grams. It is also possible to use longer n-grams, e.g. four-grams. However, the cost for indexing in IR would be much higher as there will be a lot more possible units to be considered, thereby increasing the number of indexes. Our experiments show that this additional cost does not seem to be useful for Chinese IR because most meaningful Chinese words are composed of one or two characters (our statistics shows that the average length of words in usage is 1.59). That is to say, these four basic units can cover most of the words sucessfully. Therefore, we do not need to discuss about longer n-grams in this paper.

There are six possible combinatin of two basic units. E.g. the combination of uni-grams and words.

There are five possibe combination of more than two basic units. E.g. the combination of uni-grams, words and bi-grams.

In this paper, we investigate these 15 kinds of indexing units to find the better indexing units of Chinese IR.

3 Experiments and Results

3.1 Test collection

The tests are conducted on the TREC corpus and NTCIR-2 Chinese corpus.

The documents in TREC (Harman and Voorhees, 1996) are articles published in the People's Daily from 1991 to 1993, and a part of the news released by the Xinhua News Agency in 1994 and 1995. A set of 54 queries has been set up and evaluated by people in the NIST (US National Institute of Standards and Technology).

The documents in NTCIR-2 are called CIRB010 (Chinese Information Retrieval Benchmark Version 1). These documents are all news stories downloaded from web sites of Chinatimes, Chinatimes Commercial. Chinatimes Express, Central Daily News, and China Daily News during the period of May 1998 to May 1999 (Chen and Chen, 2001), (Chiang, 1999). A set of 50 queries has been set up and evaluated by people in the NTCIR.

Some characteristics of these corpuses are given in table 1 and table 2.

	Number of doc.	Total Size (megabyte)	Average length per doc.
TREC	164,789	167.4	507 characters
NTCIR-2	132,173	192.1	635 characters

Table 1: Characteristics of the document collection

	Number of queries	Average length per query
TREC	54	119 characters
NTCIR-2	50	167 characters

Table 2: Characteristics of the query collection

3.2 Okapi System

Once Chinese sentences have been segmented in separate items, traditional IR systems may be used to index them. These separate items are called "terms" in IR. For our experiments, We used the Okapi system Windows2000 version for our runs. The system was developed in October 2000. A detailed summary of the contributions to TREC1-9 by the Okapi system is presented in (Robertson and Walker, 1999). In this section, we give a very brief introduction to the system.

In our experiments, the information retrieval evaluation system is called the Okapi Basic Search System (BSS). It is a set-oriented ranked

output system designed primarily for probabilistic type retrieval of textual material using inverted indexes. There is a family of built-in weighting scheme functions, known as BM25 and its variants. In this paper, we use BM2500 as the weighting scheme function. In addition to weighting and ranking facilities it has the usual Boolean and quasi-boolean (positional) operations and a number of non-standard set operations. Indexes are of a fairly conventional inverted type. The detailed description can be found in (Gao et al, 2001c)

3.3 Different Accuracy Word Segmentation vs. Chinese IR Performance

As mentioned in Section 1.3, we can use System S to investigate the Chinese IR performance with different accuracy word segmentation results. User can choose different rules they need to get different accuracy word segmentation by selecting the appropriate switches on the user interface.

In order to evaluate the IR performance based on word indexing with different segmentation results, we use a collection (431KB), which have been segmented manually. We call this collection as TT.

Here, we define Accuracy as following:

$$Accuracy = \frac{\#\,\text{Total Char} - \#\,\text{Error Char}}{\#\,\text{Total Char}} \times 100\,\%$$

where Total Char is the total Chinese characters in the collection and Error Char is the total characters which are segmented incorrectly.

We use TT and TREC collections to do a series of experiments as following:

1. Baseline: we do not choose any switch on the user interface. That is to say, only the longest-matching word segmentation algorithm is used.
2. We only choose the "disambiguity switch" on the user interface. On this condition, several errors of Baseline may be corrected. E.g. 中共/与其/他国/家 (China/rather than/other countries/home) will be corrected as 中 共 / 与 / 其 他 / 国 家 (China/and/other/countries).
3. We only choose the "number switch" on the user interface. On this condition, number expressions will be recgonized. (E.g. 一九三四– 1934).
4. We choose the "propernoun switch" and "number switch" on the user interface. On this condition, propernoun or number experessions will be recgonized. (E.g. 东方红三号– Dong Fang Hong III).
5. We choose the "suffix switch", "propernoun switch" and "number switch" on the user interface. On this condition, not only propernoun or number experessions but also suffix structures (e.g. 使用者– user) will be recgonized.
6. We choose all the switches above. On this condition, we can get the best acurracy segmentation.

Number	Accuracy of TT (%)	Average Precision of TREC (11pt Avg)
1	91.10	0.3729
2	92.48	0.3769
3	92.19	0.3774
4	93.09	0.3719
5	93.09	0.3747
6	94.43	0.3822

Table 3: Different segmentation results vs. Chinese IR performance

In Table 3, the second column showes the accuracy of segmentation and the third column showes the average of 11-points' precision value, which is the average of precision value on 11 points evenly chosen from 0.0 to 1.0 in the precision-recall curve. Table 3 shows that better segmentation results in better IR performances, but the difference is not significant. The reason is that we can not reach 100% accuracy in word segmentation by now. The segmentation errors may cause noise in some degree, so the improvement of better segmentation may be not significant.

3.4 Different Indexes vs. Chinese IR performance

As we have mentioned in section 2.2, we use NTCIR-2 collection to do a series of experiments[1]. We segment the queries and documents with different indexes as following respectively: (1) uni-grams(U); (2) words(W); (3) bi-grams(B); (4) tri-grams(T); (5) (U+W); (6) (U+B); (7) (U+T); (8) (W+B); (9) (W+T); (10) (BT); (11) (U+W+B); (12) (U+W+T); (13) (U+B+T); (14) (W+B+T); (15) (U+W+B+T). Then we use Okapi system to retrieval 15 groups of relevant documents. We use the two kinds of relevance assessment (Relaxed Relevance and Rigid Relevance) to evaluate the results. The standard of Rigid Relevance assessment is stricter than Relaxed Relevance assessment (Chiang, 1999).

No	Units	Relaxed	Rigid
1	U	0.6644	0.6002
2	W	0.7022	0.6485
3	B	0.6837	0.6104
4	T	0.5892	0.4979
5	U+W	0.7296	0.6654
6	U+B	0.7126	0.6369
7	U+T	0.6625	0.5619
8	W+B	0.7064	0.6447
9	W+T	0.6677	0.5837
10	B+T	0.6506	0.5525
11	U+W+B	0.7237	0.6566
12	U+W+T	0.6883	0.5998
13	U+B+T	0.6719	0.5825
14	W+B+T	0.6824	0.5899
15	U+W+B+T	0.6959	0.6099

Table 4: the results of different indexing units

Table 4 shows the average of 11-points' precision value. We can see that almost every kind of indexing units can get satisfied retrieval results in some degree.

For the four basic units: uni-grams, words, bi-grams and tri-grams, it is clear that the best performance can be obtained when using words as indexing units. Words are the natural indexing units in Chinese IR. Actually, people often use key words to describe the main idea of documents. That is why words is the best one of the four basic indexes.

However, we also noticed that bi-grams may result in a performance comparable to words. It may seem surprising because many bi-grams are meaningless. Notice, however, although many bi-grams are meaningless, if the segmentation results are the same between documents and quries, good retrieval results can still be got. For example, if the word 厄瓜多尔(Ecuador) is in a query, the segmentation result should be 厄瓜/瓜多/多尔 for bi-grams. When the same word appears in a document, the segmentation result should also be 厄瓜/瓜多/多尔, which is the same as the result in the query. Although neither of the bi-grams (e.g. 厄瓜) has actually meaning, the IR system may still retrieval the document successfully for both the document and the query have the same indexing units, which results in the similarity between the document and the query is higher than others.

On the other hand, when use tri-grams as indexing units, the result is not good, which is only 88.68% (Relaxed) and 82.96% (Rigid) of using words. As we mentioned in Section 2.2, the average length of words in usage is 1.59. That is to say, tri-grams do not match the actually words. So its result is not good. Now, we can conclude, that for Chinese IR, it is no worth using tri-grams as indexing units.

However, for uni-grams, the results are not very bad: 94.62% (Relaxed) and 93.52% (Rigid) of using words. It is clear that words are made of characters (or uni-grams). That is to say, we also can use uni-grams to get the needed information in documents. However there are only about 3,000 commonly used Chinese characters. If we only use uni-grams as IR indexes, the search space is quite limited, so the precision may become lower in practical use.

[1] Similar experiments were carried out with the TREC collections and got the similar results.

For the combinations of two basic units in column 5 to 10 in table 4, the best indexing units is the combinnation of uni-grams and words.

When we only use the longest-matching word segmentation algorithm to segment Chinese documents and queries for IR, the advantage is that longer words usually describe more precise meanings than shorter words. It could be expected that the retrieval precision (the proportion of relevant documents among those retrieved) may be high. However, as we notice, if a long word contains several short words, then only the long word will be identified as an index. The short words included are ignored. For example, if 操作系统 (operating system) is identified as a word, 操作 (operating) and 系统 (system) will not. In practice, very often, we can also refer to an "operating system" by just "system". Although the word "system" is included in "operating system", it will be considered as a completely independent index from "operating system" by IR systems. The effect of this is the loss in recall, or the phenomenon of silence. That is, some relevant documents will not be retrieved.

When the indexing units are the combination of uni-grams and words, in fact, uni-grams (single characters) may ensure a certain level of recall. Therefore, the combination of uni-grams and words may be a reasonable compromise between precision and recall.

We also noticed that the combination of uni-grams and bi-grams can also get comparable results, which is 97.67% (Relaxed) and 95.72% (Rigid) of the combination of uni-grams and words. As we have mentioned before, bi-grams may result in a performance comparable to words. It is naturally that the combination of uni-grams and bi-grams should result in a performance comparable to the combination of uni-grams and words.

We also investigate other possible combinations of these four basic units. From table 4, we can see that all the results are worse than the combination of uni-grams and words.

We can see clearly that as long as different kinds of indexes are combined, the IR performance increases. Some combinations do not increase much the cost in time and space. This is the case for the combinations of uni-grams and words. There are only about 6 000 Chinese characters in the GB codes. The addition of these characters does not increase much the search vector space, and the cost of indexing and retrieval.

On the other hand, any indexing scheme that involves bi-grams is very costly in both time and space. Virtually, there are 6 000 * 6 000 possible bi-grams in Chinese. Although many of these bi-grams actually do not appear, the number is still much higher than the possible words and characters in Chinese. This will result in a very large vector space, leading to excessive indexing time and space. Compared with the combination of uni-grams and words, there is no advantage for bi-grams, except that it does not require a dictionary. However, it is no longer a problem to acquire a high quality Chinese dictionary nowadays, so the use of bi-grams is not justified.

On the other contrary, user will use query expansion to expand the query to retrieval more related documents. On this condition, words is a convenient choice.

In conclusion, the better indexing units for Chinese IR are combination of uni-grams and words[2].

Conclusion

Many experiments have been done on Chinese IR. However, several problems have not been fully investigated: Does the accuracy of word segmentation have a significant impact on IR performance? Is it worthwhile to combine words with n-grams in Chinese IR? How should this be done?

In this study, we investigated the performance of different accuracy word segmentation results by using System S. The results shows that better segmentation results in better IR performances, but the difference is not significant. The reason is that we can not reach 100% accuracy in word segmentation by now. The segmentation errors may cause noise in some degree, so the

[2] Similar results can be got with the TREC collections.

improvement of better segmentation may be not significant.

In this study, we also made a series of experiments to find the better indexing units for Chinese IR. Our experiments show that words, uni-grams and bi-grams can achieve comparable performances. It is no worth to use tri-grams as indexes in Chinese IR. However, if we consider the time and space factors, then it is preferable to use words (and uni-grams) as indexes.

The previous experiments have tested several indexing methods that turn out to be reasonable for Chinese IR. In this paper, we tested several additional approaches. It turns out that a combination of the longest matching word segmentation algorithm with uni-grams is a good method for Chinese IR. The indexing and retrieval speed is much faster than that with bi-grams.

This series of tests is only the first step of our ongoing research program. In a later stage, Chinese IR will be used as a step in English Chinese cross language IR.

Acknowledgements

Our thanks go to Prof. Song who provide us System S. We also would like to thank Stephen Walker and Stephen Robertson who provide Okapi system to us.

References

Aitao Chen and Jianzhang He (1997) *Chinese Text Retrieval Without Using a Dictionary*. SIGIR 1997.

Chen K.H. and Chen H.H. (2001) *The Chinese Text Retrieval Task of NTCIRII Workshop*. NTCIR Workshop 2 Meeting, pp. 4-15.

Chiang Y.T. (1999) *A study on Design and Implementation for Chinese Information Retrieval Benchmark*. Master Thesis, Department of Library and Information Science, National Taiwan University.

China Times. http://news.chinatimes.com .tw/

Gao J.F., Nie J.Y., Zhang J., Zhou M. and Su Y. (2001a) *TREC-9 CLIR Experiments at MSRCN*. NIST Special Publication.

Gao J.F., Joshua Goodman, Mingjing Li and KaiFu Lee. (2001b) *Toward a unified approach to statistical language modeling for Chinese*. ACM Transactions on Asia Language Information Processing.

Gao J.F., Cao G.H., He Hongzhao, Zhang Min, Nie J.Y., Stephen Walker, Stepen Roberson (2001c) *TREC-10 Web Track Experiment at MSRA*. In TREC10.

Harman, D.K. and Voorhees, E. M. (1996) Eds. *Information Technology: The Fifth Text REtrieval Conference (TREC5)*, NIST SP 500-538. Gaithersburg, National Institute of Standards and Technology, 1996.

Kwok, K. L. and Grunfeld, L (1996) TREC-5 English and Chinese Retrieval Experiments using PIRCS, *The Fifth Text REtrieval Conference (TREC5)*, 1996.

Lee, J. H. and Ahn, J. S. (1996) *Using n-grams for Korean text retrieval.* Conference on Research and Development in Information Retrieval, ACM SIGIR, Zurich, pp. 216-224.

Nie J.Y., Martin Briseboies and Xiaobo Ren (1996) *On Chinese Text Retrieval.* SIGIR 1996, pp. 225-233.

Ogawa, Y. (1995) *A new characterbased indexing organization using frequency data for Japanese documents.* Conference on Research and Development in Information Retrieval, ACMSIGI*R*, Seattle, pp. 121-129.

Reuy-Lung, Hsiao (1997) *Surveys of Some Critical Issues in Chinese Indexing – Chinese Doument Indexing and Word Segmentation.* SIGIR 1997.

Robertson, S. E., and Walker, S. (1999) *Okapi/Keenbow at TREC8.* In TREC9.

Shi. S.C. et al, (2001) *The Development of Chinese Text Retrieval Technology – from whole Text Retrieval to Knowledge Retrieval Based on Natural Language Processing* (in Chinese), 20th Annual Conference of Chinese Information Association Society, pp. 79-88.

Song Rou, (2001) *Word Segmentation and Its Applications*, Chinese-Japanese Natural Language Processing Joint Research Promotion Conference, Kyoto, 2001