A Hybrid Chinese Information Retrieval Model

Zhihan Li, Yue Xu, Shlomo Geva

Discipline of Computer Science
Faculty of Science and Technology
Queensland University of Technology
Brisbane, Australia
zhihanlee@gmail, yue.xu@qut.edu.au, s.geva@qut.edu.au

Abstract. A distinctive feature of Chinese test is that a Chinese document is a sequence of Chinese with no space or boundary between Chinese words. This feature makes Chinese information retrieval more difficult since a retrieved document which contains the query term as a sequence of Chinese characters may not be really relevant to the query since the query term (as a sequence Chinese characters) may not be a valid Chinese word in that documents. On the other hand, a document that is actually relevant may not be retrieved because it does not contain the query sequence but contains other relevant words. In this research, we propose a hybrid Chinese information retrieval model by incorporating word-based techniques with the traditional character-based techniques. The aim of this approach is to investigate the influence of Chinese segmentation on the performance of Chinese information retrieval. Two ranking methods are proposed to rank retrieved documents based on the relevancy to the query calculated by combining character-based ranking and word-based ranking. Our experimental results show that Chinese segmentation can improve the performance of Chinese information retrieval, but the improvement is not significant if it incorporates only Chinese segmentation with the traditional character-based approach.

Keywords: Chinese Segmentation, Information Retrieval, Chinese characters

1 Introduction

The amazing growth speed in the number of Chinese Internet users indicates that building Chinese information retrieval systems is in great demand. A major difference between Chinese (Asian language) information retrieval (IR) and IR in European languages lies in the absence of word boundaries in sentences. Words have been the basic units of indexing in IR. As Chinese sentences are written in continuous character strings, pre-processing is necessary to segment sentences into shorter units that may be used as indices. Hence, a segmentation processing for corpora is necessary before building indexing and ranking. Chinese word segmentation is a difficult, important and widely studied sequence modeling problem. Word segmentation is therefore a key precursor for language processing tasks in these languages. For Chinese, there has been significant research on finding word boundaries in un-segmented sequences [3], [5], [6], [8].

For Chinese information retrieval, the query is usually a set of Chinese words rather than a sequence of Chinese characters. For character-based Chinese information retrieval, since the texts are not segmented, the retrieved documents which contain the character sequence of the query may not be relevant to the query as they may not contain the words in the query. Therefore, the quality of character-based Chinese information retrieval is not satisfactory.

The impact of Chinese word segmentation on the performance of Chinese information retrieval has been investigated in previous research. [4] has conducted a series of experiments which conclude that word segmentation has a certain positive impact on the performance of Chinese information retrieval. However, [4] suggests that for Chinese IR, the relationship between word segmentation and retrieval performance is in fact non-monotonic. In this investigation, [4] used a wide range of segmentation algorithms with accuracies from 44% to 95%. The experimental results showed that retrieval performance increases with the increase of segmentation accuracy in the first part from the lowest segmentation accuracy of 44%. However, after a point around 77%, the retrieval performance decreases from plateaus with the increase of segmentation accuracy.

Both Chinese characters and words can be used as the indexing units for Chinese IR. Both of these have advantages and disadvantages. In general, character indexing based IR may achieve better recall since it can retrieve most of the relevant documents as long as they contain the query terms (the query terms are sequences of Chinese characters in the documents, not segmented words, since the documents are not segmented). However, the retrieval precision is not necessarily good. This is because many irrelevant documents are ranked high due to high query term frequency, since they have many instances of the query term sequences, many of which are actually not valid words. On the other hand, the word indexing based IR can make a better ranking and therefore achieve a little better performance than that of character indexing based IR, but the improvement is limited since some relevant documents may not contain the query terms as segmented words and thus will not be retrieved.

In this paper, we propose to combine the two approaches in order to achieve better retrieval performance. In our approach, we create two indexing tables, one a Chinese character indexing table and the other a segmented word indexing table. Three methods are proposed to make use of the two indexing tables, with the hope of improving the accuracy of ranking. We first briefly describe the segmentation system that we used for Chinese word segmentation, then in Section 3, we introduce our hybrid approach of Chinese IR based on both segmentation words and Chinese characters. After that, in Section 4, we represent the experimental results. Section 5 concludes the paper.

2 Segmentation Procedure

As Chinese word segmentation is not a research focus of this thesis, we have used the segmentation system developed by the Institute of Information science, Academia Sinica in Tai-Wan (http://ckipsvr.iis.sinica.edu.tw). In this system, the processes of segmentation could be roughly divided into two steps; one is resolving the ambiguous matches, the other is identifying unknown words. These processes adopt a variation of the longest matching algorithm, with several heuristic rules to resolve the ambiguities and achieve the high success rate of 99.77% reported in [7]. After a disambiguation process, for the needs of the unknown word extraction, a POS bi-gram model is applied to tag the POS of words. In the unknown-word extraction process, a bottom-up merging algorithm, which utilizes hybrid statistical and linguistic information, is adopted.

In practice, the client sends a request, a piece of a Chinese document, to the system and the system then responses and returns the segmented document in XML style. For example, an original Chinese document (Figure 1-(a)) and the corresponding segmented document (Figure 1-(b)) are given below:

【美聯社里約熱內盧卅一日電】巴西球星羅納度在新年向球迷請命，請球迷耐心給他時間康復，並且無奈的說「我不是機械人」。耶誕節才新婚的羅納度帶著妻子在里約熱內盧過年，他接受「閱讀」雜誌訪問時說，幾乎每晚他都夢到在沙灘上踢足球，但醒後膝痛難擋，使他不得不面對殘酷的現實，11月動過手術的膝蓋尚未復元。羅納度說，球迷對他熱切期盼的心意令人感動，但「我不是機械人，不能說上緊螺絲就萬事OK」，他保證盡一切努力鍛鍊，希望在2000年恢復巔峰狀態。

(a) Original Chinese Document

(b) Segmented Chinese Document

Figure 1 Example of Segmented Document

After segmentation, words are separated with white blanks, shown in Figure 1-(b). Additionally, a POS tag is provided immediately following each word in the bracket.

3 Retrieval and Ranking Approaches

For retrieval process, we used the character-based Boolean model. The approach considers each query term as a sequence of Chinese characters, regardless of whether it is a word or not. If a document contains such a sequence, it will be retrieved as a candidate of relevant documents and will be ranked in the ranking procedure. Using the character-based retrieval model, guarantees that none of the documents containing query terms will be missed in the first retrieval procedure. In the following sub sections, we first describe the architecture of our hybrid Chinese IR model, and then discuss the three ranking methods in detail.

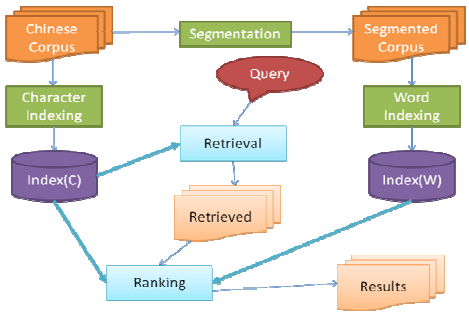

Figure 2 Architecture of the Hybrid Chinese IR Model

In Figure 2, the architecture of the hybrid Chinese IR model is depicted. In this model, we first construct the segmented corpus by segmenting all the documents in the original corpus. Then two indexing tables are built up from the original corpus and the segmented corpus; one is the character indexing table containing all characters appearing in the original corpus, the other is the Chinese word indexing table containing all segmented words appearing in the segmented corpus.

We have conducted an experiment to evaluate the retrieval performance by using either the character indexing table only or using segmented word indexing table only. The result, as given in table 1 below, shows that the recall and precision of the pure character-based retrieval model are better than those of the pure word-based model. This is because, in general, the character indexing based model can retrieve most of the relevant documents as long as they contain the query terms, while for the word-based retrieval model, in which the corpus contains segmented documents, the recall of retrieval is worse since the query terms very often are not segmented words. In our previous research, we have developed a Chinese information retrieval system that uses the traditional Boolean model to perform retrieval only, based on the character indexing table. As indicated in the architecture, in this research the retrieval step is performed based only on the character indexing table to generate candidate relevant documents. The previously developed character-based retrieval model is used to perform the retrieval task.

Table 1. Performance of Pure Character and Pure Word Models

Top N	Pure Character		Pure Word	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
10	39.7	19.6	23.9	13.7
15	35.2	26.1	20.1	15.8
20	32.7	30.9	18.0	18.9
30	27.5	36.7	15.5	21.2
100	14.1	53.8	8.7	32.6
Ave	29.9	33.4	17.2	20.4

Even though the character based retrieval model can retrieve most of the relevant documents, it doesn't necessarily mean that the most relevant documents will be delivered to the user. The relevancy ranking plays a very important role in determining the most relevant documents. In this research, we propose two ranking methods based on both character indexing and word indexing; these are described in the following sections.

3.1 Ranking Methods

(1) Word-enhanced Ranking Method

In our previous research [1], [2], we used a *tf-idf* based algorithm to calculate the ranking value of a retrieved document to determine the top N relevant documents. The document rank score for a query of *m* terms is calculated with the following equation:

$$d_{rank} = n^5 \times \sum_{i=1}^{m} tf_i \times idf_i \qquad (3.1)$$

Here, *m* is the number of query terms, *n* is the number of distinct query terms that appear in the document as character sequences which are not necessarily segmented words (in this thesis we refer to the query terms appearing in the document as query character sequences, to differentiate from the segmented words), tf_i is the frequency of the i^{th} term in the document and idf_i is the inverse document frequency of the i^{th} term in the collection. The equation can ensure two things: firstly, the n^5 strongly rewards the documents that contain more query terms. The more the distinct query terms are matched in a document, the higher the rank of the document. For example, a document that contains four distinct query terms will almost always have higher rank than a document that contains three distinct query terms, regardless of the frequency of the query terms in the document. Secondly, when documents contain a similar number of distinct terms, the score of a document will be determined by the sum of the *tf-idf* value of the query terms, as in traditional information retrieval. According to our experiments, n^5 is the best value for the NTCIR5 Chinese collection that we used in our experiments.

The word-enhanced ranking method proposed here is an extension of the traditional *tf-idf* based ranking method mentioned above. The equation to calculate the document rank score is given in Equation (3.2), in which not only the frequency of the query character sequences but also the frequency of the segmented words is taken into consideration.

$$d_{rank} = (n_c + n_w)^5 \times \left(\sum_{i=1}^{m} tf_i^c \times idf_i^c + \frac{n_w}{n_c + n_w} \times \sum_{i=1}^{m} tf_i^w \times idf_i^w \right) \qquad (3.2)$$

In Equation (3.2), *m* is the number of query terms, n_c is the number of query terms that appear in the document as character sequences but not segmented words, n_w is the number of query terms that appear in the document as segmented words, tf_i^c is the frequency of the i^{th} query term appearing as a sequence in the document and idf_i^c is the inverse document frequency of the i^{th} query term appearing as a sequence in the collection. Similarly, tf_i^w and idf_i^w are the frequency of the i^{th} query term appearing as a segmented word in the document and the inverse document frequency of the i^{th} query term in the collection, respectively.

In the equation (3.2), the ranking score d_{rank} is derived from two parts, the frequency of the query terms as character sequences (the first part of the equation) and the frequency of query terms as segmented words (the second part of the equation). The first part is actually Equation (3.1). The second part is an additional contribution to the ranking score from the query terms which are also segmented words. The intention of adding the second part to Equation (3.1) is to enhance the impact of the segmented words on the ranking. If there are no query terms that appear in the document as segmented words, $n_w = 0$, the second part becomes 0 and Equation (3.2) becomes Equation (3.1). The higher the n_w, the more the ranking score is increased by the

second part. The idea behind this strategy is to increase the rank of the documents which contain query terms that are segmented words. This method emphasizes the importance of the segmented query terms in calculating the document rank score. If a document contains the query terms as segmented words, the document will get a higher ranking score than that scored by using Equation (3.1).

(2) Average based Ranking Method

This method is also an extension of the *tf-idf* based ranking method. It simply calculates the average of the frequencies of the query terms as character sequences and as segmented words. The equation to calculate the document rank score is given below:

$$d_{rank} = \left(n_c^5 \times \sum_{i=1}^{m} tf_i^c \times idf_i^c + n_w^5 \times \sum_{k=1}^{m} tf_k^w \times idf_k^w \right) / 2 \qquad (3.3)$$

Equation (3.3) also consists of two parts, the frequency of query terms as character sequences and the frequency of query terms as segmented words. Similar to the word-enhanced method, if there is no query terms that appear in the document as segmented words, n_w = 0, the second part becomes 0 and only the first part is used to derive the ranking score. When the number of query terms that are segmented words increases, the contribution of the first part decreases, while the contribution of the second part increases. The higher the n_w, the more the contribution from the second part increases. Different from the word-enhanced method, in which the number of query terms as segmented words increases, this method not only increases the contribution of the segmented words but also decreases the contribution of the query term as character sequences only. The idea is that if a document contains fewer query terms that there are segmented words, the topic of this document may be irrelevant to the query terms; hence, the contribution from these query terms should be reduced.

4 Experimental Results and Evaluation

We have conducted an experiment to evaluate the performance of the proposed three ranking methods. The experiment is conducted on a DELL PC within Pentium 4 processor and 1GB physical memory, 80GB hard disk space. The system is implemented in C# by using MS Visual Studio 2005, MS SQL Server 2005 and Windows XP Professional Operating System.

4.1 Data Collection

The Chinese corpus obtained from the NTCIR5 (http://research.nii.ac.jp/ntcir/) is used as the testing data: it contains 434,882 documents of news articles in traditional Chinese. Even though the experiments are conducted in traditional Chinese, the techniques proposed in this work can be applicable to simplified Chinese. The detailed information of the test set is as shown in table 2 below.

Table 2. Statistic of Chinese Corpus

Document collection	Year 2000	Year 2001	No. of articles
United Express (ude)	40445	51851	92296
Ming Hseng News (mhn)	84437	85302	169739
Economic Daily News (edn)	79380	93467	172847
Total	204262	230620	434882

The document itself is in XML format with the following tags:

- <DOC> </DOC> The tag for each document
- <DOCNO> </DOCNO> Document identifier
- <LANG> </LANG> Language code: CH, EN, JA, KR
- <HEADLINE> </HEADLINE> Title of this news article
- <DATE> </DATE> Issue date
- <TEXT> </TEXT> Text of news article
- <P> </P> Paragraph marker

Queries used in the experiment are from the NTCIR5 CLIR task. There are a total of 50 queries created by researchers from Taiwan, Japan and Korea. NTCIR5 provided both English queries and the corresponding Chinese queries. The Chinese queries are used in this research.

4.2 Retrieval model and Evaluation Measures

We use the traditional Boolean model as our retrieval model to retrieve potential candidate relevant documents. If a document contains one or more query terms, no matter whether as character sequences or segmented words, the document will be retrieved as a candidate relevant document. For all the ranking methods tested in this experiment, the same retrieval model is used to ensure a fair comparison. For indexing units, all segmented words and characters in the whole collection of Chinese documents are extracted as units. We create two indexing tables for characters and segmented words, respectively. In our experiment, the traditional precision and recall evaluation metrics are used to measure the effectiveness of the proposed ranking methods. The evaluation is done to various numbers of retrieved documents, ranging from top 10, 15, 20, 30 to 100 documents.

4.3 Performance Comparison

The baseline model used in this experiment to compare with the proposed methods is the traditional character-based ranking model described in Equation (3.1). The experiment results are given in the following tables (3 and 4) and figures (3 to 6).

Table 3. Precision Comparison

TOP N	Character (%)	Word-Enhanced (%)	Average (%)
10	39.7	39.3	39.9
15	35.2	35.5	35.5
20	32.7	32.5	32.9
30	27.5	27.8	27.5
100	14.1	14.5	14.6
Ave	29.9	29.9	30.1

Table 4 Recall Comparison

TOP N	Character (%)	Word-Enhanced (%)	Average (%)
10	19.6	19.7	20.1
15	26.1	25.9	26.5
20	30.9	30.4	31.3
30	36.7	36.6	35.7
100	53.8	53.9	53.9
Ave	33.4	33.3	33.5

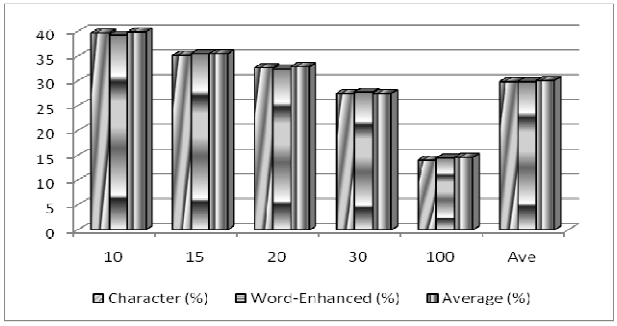

Figure 3 Precision between Ranking Models

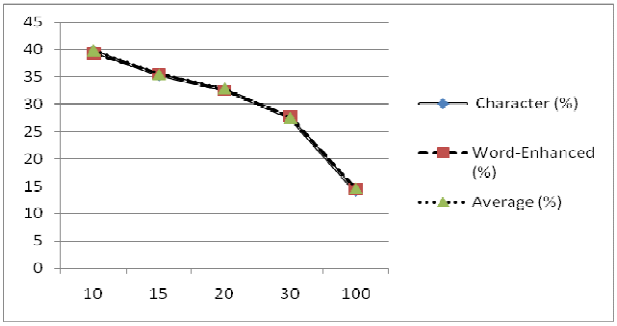

Figure 4 Curve of Precision

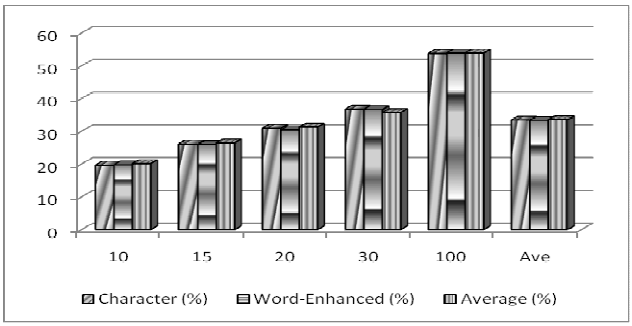

Figure 5 Recall between Ranking Models

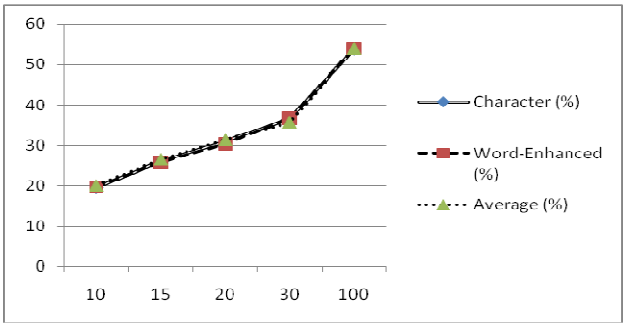

Figure 6 Curve of Recall

From the evaluation results of precision and recall, we find that the performances of the three models are very close; only the average model achieves slightly better results. On average the precision is improved by 0.2% and the recall is improved by only 0.1%. The experimental results show that Chinese word segmentation can improve the performance of Chinese information retrieval, but the improvement is not significant. This also confirms that a high accuracy of word segmentation, reported as 95% in [8] and used in our experiment, may not increase retrieval performance and may eventually decrease while the accurate of segmentation increases over a point [4].

5 Conclusion

In this chapter, we propose two methods for ranking retrieved documents by a hybrid of the character based relevancy measure and the segmented word based relevancy measure, in order to improve the performance of Chinese information retrieval. From the experimental results, we find that those approaches achieved slight improvement over the traditional character based approach, which indicates that the influence of taking segmented words into consideration in ranking retrieved documents is limited.

References

1. Chengye Lu, Yue Xu, Shlomo Geva: Translation disambiguation in web-based translation extraction for English-Chinese CLIR. Proceedings of the 2007 ACM symposium on applied computing, Pages 819–823 (2007)
2. Geva, S.: GPX - Gardens Point XML IR at INEX 2005, INEX 2005. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 240–253. Springer, Heidelberg (2006)
3. Jianfeng Gao and Mu Li: Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. Computational Linguistics, MIT. 531-574, Vol 31, Issue 4 (2005)
4. Peng F., Huang X., Schuurmans D. and Cercone N.: Investigating the relationship between word segmentation performance and retrieval performance in Chinese IR, Proceedings of the 19th international conference on Computational linguistics, Pages 1–7 (2002)
5. Nianwen Xue: Chinese Word Segmentation as Character Tagging. Computational Linguistics and Chinese Language Processing, Vol 8, No 1, Pages 29–48 (2003)
6. Richard Sproat and Chilin Shih: Corpus-Based Methods in Chinese Morphology and Phonology. AT&T Labs — Research (2002)

7. Wei-Yun Ma, Keh-Jiann Chen: Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff. Proceedings of the second SIGHAN workshop on Chinese language processing - Volume 17, Pages 168 - 171 (2003)
8. Xinjing Wang, Wen Liu and Yong Qin: A Search-based Chinese Word Segmentation Method. Proceedings of the 16th international conference on World Wide Web , Pages 1129 – 1130 (2006)