Application of the Tightness Continuum Measure to Chinese Information Retrieval

Ying Xu[†], **Randy Goebel**[†], **Christoph Ringlstetter**[‡] and **Grzegorz Kondrak**[†]

[†]Department of Computing Science
University of Alberta

[‡]Center for Language and
Information Processing (CIS)
Ludwig Maximilians University

{yx2,goebel,kondrak}@cs.ualberta.ca kristof@cis.uni-muenchen.de

Abstract

Most word segmentation methods employed in Chinese Information Retrieval systems are based on a static dictionary or a model trained against a manually segmented corpus. These *general* segmentation approaches may not be optimal because they disregard information within semantic units. We propose a novel method for improving word-based Chinese IR, which performs segmentation according to the tightness of phrases. In order to evaluate the effectiveness of our method, we employ a new test collection of 203 queries, which include a broad distribution of phrases with different tightness values. The results of our experiments indicate that our method improves IR performance as compared with a general word segmentation approach. The experiments also demonstrate the need for the development of better evaluation corpora.

1 Introduction

What distinguishes Chinese Information Retrieval from information retrieval (IR) in other languages is the challenge of segmenting the queries and the documents, created by the lack of word delimiters. In general, there are two categories of segmenters: character-based methods and word-based methods. Despite the superior performance of bigram segmenters (Nie *et al.*, 2000; Huang *et al.*, 2000; Foo and Li, 2004), word-based approaches continue to be investigated because of their application in sophisticated IR tasks such as cross language IR, and within techniques such as query expansion (Nie *et al.*, 2000; Peng *et al.*, 2002a).

Most word-based segmenters in Chinese IR are either rule-based models, which rely on a lexicon, or statistical-based models, which are trained on manually segmented corpora (Zhang *et al.*, 2003). However, the relationship between the accuracy of Chinese word segmentation and the performance of Chinese IR is non-monotonic. Peng *et al.* (2002b) reported that segmentation methods achieving segmentation accuracy higher than 90% according to a manual segmentation standard yield no improvement in IR performance. They further argued that IR often benefits from splitting compound words that are annotated as single units by manual segmentation.

The essence of the problem is that there is no clear definition of *word* in Chinese. Experiments have shown only about 75% agreement among native speakers regarding the correct word segmentation (Sproat *et al.*, 1996). While units such as "花生" (peanut) and "月下老人" (match maker) should clearly be considered as a single term in Chinese IR, compounds such as "机器学习" (machine learning) are more controversial.[1]

Xu *et al.* (2009) proposed a "continuum hypothesis" that rejects a clean binary classification of Chinese semantic units as either compositional or non-compositional. Instead, they introduced the notion of a *tightness measure*, which quantifies the degree of compositionality. On this tightness continuum, at one extreme are non-

[1]This issue is also present to a certain degree in languages that do use explicit delimiters, including English (Halpern, 2000; McCarthy *et al.*, 2003; Guenthner and Blanco, 2004).

Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010), pages 55–63,
Beijing, August 2010

compositional semantic units, such as "月下老人" (match maker), and at the other end are sequences of consecutive words with no dependency relationship, such as "上海哪有" (Shanghai where). In the middle of the spectrum are compositional compounds such as "机器学习" (machine learning) and phrases such as "正当收入" (legitimate income).

In this paper, we propose a method to apply the concept of semantic tightness to Chinese IR, which refines the output of a general Chinese word segmenter using tightness information. In the first phase, we re-combine multiple units that are considered semantically tight into single terms. In the second phase, we break single units that are not sufficiently tight. The experiments involving two different IR systems demonstrate that the new method improves IR performance as compared to the general segmenter.

Most Chinese IR systems are evaluated on the data from the TREC 5 and TREC 6 competitions (Huang *et al.*, 2000; Huang *et al.*, 2003; Nie *et al.*, 2000; Peng *et al.*, 2002a; Peng *et al.*, 2002b; Shi and Nie, 2009). That data contains only 54 queries, which are linked to relevancy-judged documents. During our experiments, we found the TREC query data is ill-suited for analyzing the effects of compound segmentation on Chinese IR. For this reason, we created an additional set of queries based on the TREC corpus, which includes a wide variety of semantic compounds.

This paper is organized as follows. After summarizing related work on Chinese IR and word segmentation studies, we introduce the measure of semantic tightness. Section 4 describes the integration of the semantic tightness measure into an IR system. Section 5 discusses the available data for Chinese IR evaluation, as well as an approach to acquire new data. Section 6 presents the results of our method on word segmentation and IR. A short conclusion wraps up and gives directions for future work.

2 Related Work

The impact of different Chinese word segmentation methods on IR has received extensive attention in the literature (Nie *et al.*, 2000; Peng *et al.*, 2002a; Peng *et al.*, 2002b; Huang *et al.*, 2000; Huang *et al.*, 2003; Liu *et al.*, 2008; Shi and Nie, 2009). For example, Foo and Li (2004) tested the effects of manual segmentation and various character-based segmentations. In contrast with most related work that only reports the overall performance, they provide an in-depth analysis of query results. They note that a small test collection diminishes the significance of the results.

In a series of papers on Chinese IR, Peng and Huang compared various segmentation methods in IR, and proposed a new segmentation method (Peng *et al.*, 2002a; Peng *et al.*, 2002b; Huang *et al.*, 2000; Huang *et al.*, 2003). Their experiments suggest that the relationship between segmentation accuracy and retrieval performance is non-monotonic, ranging from 44%-95%. They hypothesize that weak word segmenters are able to improve the accuracy of Chinese IR by breaking compound words into smaller constituents.

Shi and Nie (2009) proposed a probability-based IR score function that combines a unigram score with a word score according to "phrase inseparability." Candidates for words in the query are selected by a standard segmentation program. Their results show a small improvement in comparison with a static combination of unigram and word methods.

Liu *et al.* (2008) is the research most similar to our proposed method. They point out that current segmentation methods which treat segmentation as a classification problem are not suitable for Chinese IR. They propose a ranking support vector machine (SVM) model to predict the internal association strength (IAS) between characters, which is similar to our concept of tightness. However, they do not analyze their segmentation accuracy with respect to a standard corpus, such as Chinese Treebank. Their method does not reliably segment function words, mistakenly identifying "的人" ('s people) as tight, for example. Unlike their approach, our segmentation method tackles the problem by combining the tightness measure with a general segmentation method.

Chinese word segmentation is closely related to multiword expression extraction. McCarthy *et al.* (2003) investigate various statistical measures of compositionality of candidate multiword verbs.

Silva *et al.* (1999) propose a new compositionality measure based on statistical information. The main difference with Xu *et al.*'s measure is that the latter is focused on word sense disambiguation. In terms of multiword expressions in IR, Vechtomova (2001) propose several approaches, such as query expansion, to incorporating English multiword expressions in IR. Braschler and Ripplinger (2004) analyze the effect of stemming and decompounding on German text retrieval. However, Chinese compound segmentation in IR is a thorny issue and needs more investigation for the reasons mentioned earlier.

3 Semantic Tightness Continuum

We adopt the method developed by (Xu *et al.*, 2009) for Chinese semantic unit tightness measure, which was shown to outperform the pointwise mutual information method. For the sake of completeness we briefly describe the basic approach here. The input of the measure is the probability distribution of a unit's segmentation patterns, i.e., potential segmentation candidates. The output is a tightness value; the greater the value, the tighter the unit. In this paper, we focus on 4-gram sequences because 4-character compounds are the most prominent in Chinese. There are eight possible segmentations of any 4-character sequence: "ABCD," "A|BCD," "A|B|CD," etc. For a sequence of n characters, there are 2^{n-1} potential segmentations. Equation 1 below defines the tightness measure.

$$ratio = \begin{cases} \frac{\sharp Pt(s)}{\max(\sharp Pt(s_1|s_2)) + \frac{1}{N}} & \text{if } \sharp Pt(s) > \sigma \\ \text{undef} & \text{otherwise} \end{cases} \quad (1)$$

In Equation 1, $\sharp Pt(s)$ stands for frequencies of segmentation patterns of a potential semantic unit s; $Pt(s_1|s_2)$ is a pattern which segments the unit s into two parts: s_1 and s_2; σ is a threshold to exclude rare patterns; and N is a smoothing factor which is set as the number of documents. Note that when the first part of the denominator is zero, the ratio of the unit will be very high. Intuitively, the lack of certain separating patterns in the data is evidence for the tightness of the units.

4 Application to Chinese IR

We propose a novel approach to segmentation for Chinese IR which is based on the tightness measure. Our segmenter revises the output of a general segmenter according to the tightness of units. The intuition behind our method is that segmentation based on tightness of units will lead to better IR performance. For example, keeping "皮纳图博" (Pinatubo) as a unit should lead to better results than segmenting it into "皮(skin)|纳(include)|图(picture)|博(large)". On the other hand, segmenting the compositional phrase "科威特国" (Kuwait country) into "科威特(Kuwait)|国(country)" can improve recall. We revise an initial segmentation in two steps: first, we combine components that should not have been separated, such as "皮纳图博" (Pinatubo); second, we split units which are compositional, such as "科威特国" (Kuwait country).

In order to combine components, we first extract 4-gram non-compositional compounds whose tightness values are greater than a threshold σ_1 in a reference corpus, and then revise a general segmenter by combining two separated words if their combination is in the list. This approach is similar to the popular longest match first method (LMF), but with segmentation chunks instead of characters, and with the compound list serving as the lexicon. For example, consider a sequence "ABCDEFGHIGK," which a general segmenter annotates as "ABC|D|E|F|G|HI|GK." If our compound list constructed according to the tightness measure contains {"DEFG"}, the revised segmentation will be "ABC|DEFG|HI|GK." Units of length less than 4 are segmented by using the LMF rule against a dictionary.

In order to split a compositional unit, we set the additional thresholds σ_2, σ_3, and σ_4, and employ the segmentation rules in Equation 2. The intuition comes from the pattern lattice of a unit (Figure 1). For the patterns on the same level, the most frequent pattern suggests the most reasonable segmentation. For the patterns on different levels, the frequency of each level indicates the tightness of the unit.

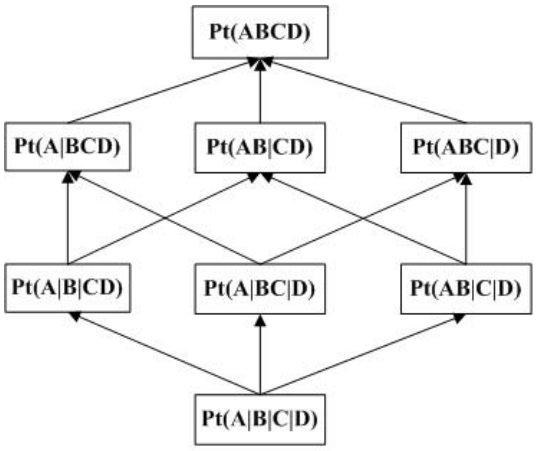

Figure 1. The Lattice of the 8 Patterns.

$$
\begin{array}{l}
\text{if} \\
\quad v_1 = \frac{\sharp Pt(ABCD)}{\max(\sharp Pt(A|BCD), \sharp Pt(AB|CD), \sharp Pt(ABC|D)) + \frac{1}{N}} > \sigma_2 \\
\text{then ``ABCD'' is one unit;} \\
\\
\text{else if} \\
\quad v_2 = \frac{\max(\sharp Pt(A|BCD), \sharp Pt(AB|CD), \sharp Pt(ABC|D)) + \frac{1}{N}}{\max(\sharp Pt(A|B|CD), \sharp Pt(A|BC|D), \sharp Pt(AB|C|D)) + \frac{1}{N}} > \sigma_3 \\
\text{then ``ABCD'' is segmented into two parts;} \\
\\
\text{else if} \\
\quad v_3 = \frac{\max(\sharp Pt(A|B|CD), \sharp Pt(A|BC|D), \sharp Pt(AB|C|D)) + \frac{1}{N}}{\sharp Pt(A|B|C|D) + \frac{1}{N}} > \sigma_4 \\
\text{then ``ABCD'' is segmented into three parts;} \\
\\
\text{else} \\
\text{``ABCD'' is segmented into four parts;}
\end{array}
\tag{2}
$$

We apply the rules in Equation 2 to the sequence of 4-grams, with simple voting for selecting the segmentation pattern. For example, within the sequence "ABCDEF," three 4-gram patterns are considered: "ABCD," "BCDE," and "CDEF." If only one of the 4-grams contains a segmentation delimiter, the insertion of the delimiter depends only upon that 4-gram. If two 4-grams contain the same delimiter, the insertion of the delimiter depends upon the two 4-grams. If the two 4-grams disagree on the segmentation, a confidence value is calculated as in Equation 3,

$$\text{confidence} = v_i - \sigma_{i+1}, \tag{3}$$

where $i \in [1, 2, 3]$. If three 4-grams contain the same delimiter, voting is employed to decide the segmentation. Returning to our example, suppose that the first 4-gram is segmented as "A|B|C|D," the second as "BC|DE," and the third as "C|DE|F." Then the segmentation delimiter between "A" and "B" is inserted, but the delimiter between "B" and "C" depends on the confidence values of the first two segmentation patterns. Finally, the delimiter between "C" and "D" depends on the result of voting among the three 4-gram segmentations.

The two steps of combining and splitting can either be applied in succession or separately. In the former case, σ_1 must be greater or equal to σ_2. In the remainder of this paper, we refer to the first step as "Tight_Combine," and to the second step applied after the first step as "Tight_Split." Note that the second method can be used to segment sentences directly instead of revising the output of a general segmenter. This method, which we refer to as "Online_Tight," has the same shortcoming as the method of Liu *et al.* (2008), namely it frequently fails to segment function words. For example, it erroneously identifies "的人" ('s people) as tight. Therefore, we do not attempt to embed it into the IR systems discussed in Section 6.

5 Test Collection

We analyzed the currently available Chinese test collection of TREC, and found it unsuitable for evaluating different strategies of compound segmentation. One problem with the TREC data is that the Chinese queries (topic titles) have too many keywords. According to the output of ICTCLAS, a general segmenter, the average length of Chinese queries is 12.2 words; in contrast, the average length of English ad-hoc queries in TREC-5 and 6 (English_topics 251-350) is 4.7. Even if we use English translation of the Chinese queries instead, the average length is still more than 7 words. The problem with long queries is that they introduce complicating effects that interact in ways difficult to understand. An example is the co-occurrence between different keywords in the base corpus. Sometimes a completely correct segmentation causes a decrease in IR performance because the score function assigns a higher score to less important terms in a topic. For example, for query 47 (Trec-6 dataset), "菲律宾，皮纳图博火山，火山灰，岩浆，爆发" (Philippines, Mount Pinatubo, volcanic ash, magma, eruption), preserving the unit Pinatubo makes the average precision drop from 0.76 to 0.62 as compared to the segmentation "皮|纳|图|博". The score of the

unit is lower than that the sum of its components, which results in a relatively low ranking for some relevant documents. Another problem with the TREC Chinese test collection is the small number of queries (54). The number of of queries containing non-compositional words is smaller still. Similarly, the other available corpus, NTCIR, comprises only 50 queries. In order to be confident of our results, we would like to have a more substantial number of queries containing units of varying tightness.

Because of the shortcomings of available data sets, we created our own test collection. There are three components that define an IR test collection: a query set, a corpus from which relevant documents are retrieved, and relevance judgements for each query. Our criteria for gathering these components are as follows.

First, the set of queries should contain both tight queries and loose queries. For example, there should be tight queries such as "月下老人" (match maker), loose queries such as "上海海关" (Shanghai customs), and queries with tightness values in between, such as "机器学习" (machine learning). Furthermore, the queries should be realistic, rather than constructed by introspection. In order to meet these requirements we randomly chose 4-gram noun phrases (tagged by ICTCLAS) from the TREC corpus. 51 queries are from a real data set, the Sogou query logs[2]. The remaining 152 queries, which are selected manually based on the initial 51 queries, represent queries that IR system users are likely to enter. For example, queries of locations and organizations are more likely than queries such as "how are you." Finally, the queries should not be too general (i.e., resulting in too many relevant documents found), nor too specific (no relevant documents). Therefore, we selected the 4-grams which had the corresponding document frequency in the TREC corpus between 30 and 300.

The second set of criteria concerns the relevance judgements of documents. As our retrieval corpus, we adopted the TREC Mandarin corpus, which contains 24,959 documents. Because of resource limitation, we used the Minimum Test Collection (MTC) method (Carterette *et al.*, 2006). The method pools documents in such a way that the documents which are best for discriminating between different IR systems are judged first. We applied this method on a document set that contains all of the top 100 results of 8 IR systems (two score functions, tf*idf and BM25, 4 indexing methods, unigram, bigram, ICTCLAS segmentation, and our Tight_Combine segmentation). The systems were implemented with the Lucene framework (http://lucene.apache.org/).

The last criterion determines which document is relevant to a query. Annotators' opinions vary about whether a document is relevant to a topic. Is having the query in a document enough to be the criterion of relevance? For the query "Beijing airport," should the document that contains the sentence "Chairman Mao arrived at the Beijing airport yesterday," be classified as relevant? Since our goal is to analyze the relationship between Chinese word segmentation, and IR, we use weak relevant judgements. It is more related to score functions to distinguish weak relevance from strong relevance, that is, whether the query is the topic of the document. This means the above document is judged as relevant for the query "Beijing airport."

In summary, our own test collection has about 200 queries, and at least 100 judged documents per query with the TREC corpus as our base corpus[3].

6 Experiments

We conducted a series of experiments in word-based Chinese information retrieval, with the aim of establishing which segmenter is best for CIR, while pursuing the best segmentation performance in terms of segmented corpus is not the main crux. In this section, we first present the accuracy of different segmentation methods, and then discuss the results of IR systems.

6.1 Chinese Word Segmentation

ICTCLAS is a Chinese segmentation tool built by the Institute of Computing Technology, Chinese Academy of Sciences. Its segmentation model is a

[2]Sogou query logs 2007 can be downloaded at http://www.sogou.com/labs/dl/q.html.

[3]The query set and relevance judgements are available at http://www.cs.ualberta.ca/~yx2/research.html

class-based hidden Markov model (HMM) model (Zhang *et al.*, 2003). The segmenter is trained from manually segmented corpus, which makes it ignore both the tightness of units and unknown words such as "皮纳图博" (Pinatubo), which are difficult to identify.

In this experiment, we segmented the Chinese Treebank using ICTCLAS and our three methods that employ the tightness measure. The evaluation is based on the manual segmentation of the corpus. We evaluated the methods on the entire Treebank corpus, employing 10-cross validation for result significance verification.

In order to measure the tightness of Chinese semantic units, pattern distributions of every 4-gram were extracted from the Chinese Gigaword corpus. Tight_Combine is the ICTCLAS refined segmentation that employs the non-compositional compound list from the Chinese Gigaword corpus. The threshold for non-compositional compound σ_1 is set to 11. Tight_Split is the refined segmentation of Tight_Combine using Equation 2. Online_Tight is the segmentation using Equation 2 directly. For Tight_Split and Online_Tight, we employed a lexicon which contains 41,245 words, and set the thresholds σ_2, σ_3, and σ_4 to 11, 0.01, and 0.01, respectively. The parameters σ_1 and σ_2 are set according to the observation that the percentage of non-compositional units is high when the tightness is greater than 11 for all the 4-grams in the Chinese Gigaword corpus. The other two parameters were established after experimenting with several parameter pairs, such as (1,1), (0.1, 0.1), and (0.1, 0.01). We chose the one with the best segmentation accuracy according to the standard corpus.

Table 1 shows the mean accuracy result over the 10 folders. The accuracy is the ratio of the number of correctly segmented intervals to the number of all intervals. The result shows that our method improves over the ICTCLAS segmentation result, but the improvement is not statistically significant (measured by t-test). The only significant result is that Online_tight is worse than other methods.

Surprisingly, there is a large gap between Tight_Split and Online_Tight, although they employ the same parameters. It turns out the major difference lies in the segmentation of function words. Since it is based on ICTCLAS, Tight_Split does a good job in segmenting function words such as verbal particles which represent past tense "了" and the nominalizer "的." Online_Tight tends to combine these words with the consecutive one. For example, considering "积累了" (cumulated), the Treebank and Tight_Split segment it into "积累|了" (cumulate + particle); while Online_Tight leaves it unsegmented.

ICTCLAS	88.8%
Tight_Combine	89.0%
Tight_Split	89.1%
Online_Tight	80.5%

Table 1. Segmentation accuracy of different segmenters.

6.2 IR Experiment Setup

We conducted our information retrieval experiments using the Lucene package (Hatcher and Gospodnetic, 2004). The documents and queries were segmented by our three approaches before indexing and searching process. In order to analyze the performance of our segmentation methods with different retrieval systems, we employed two score functions: the BM25 function (Peng *et al.*, 2002b) [4]; and BM25Beta (Function 4), which prefers documents with more query terms.

$$Score(Q,D) = \begin{cases} \frac{T}{(1+\beta)*N} \sum_{i=0}^{T} score(t_i, D) & \text{if } T < N \\ \sum_{i=0}^{N} score(t_i, D) & \text{if } T = N \end{cases} \quad (4)$$

In the above equation, $score(t_i, D)$ is the score of the term t_i in the document D. Although we used BM25 as our base score function for $score(t_i, D)$, it can be replaced by other score functions, such as tf*idf, or a probability language model. β is a parameter to control a penalty component for those documents that do not contain all the query terms; T is the number of distinctive query terms in the document; and N is the number of query terms. The function penalizes documents that do not contain all the query terms,

[4] An implementation of BM25 into Lucene can be downloaded at http://arxiv.org/abs/0911.5046

	BM25	BM25Beta
ICTCLAS	62.78%	70.79%
Tight_Combine	65.92%	71.19%
Tight_Split	63.40%	70.95%

Table 2. MAP of different IR systems with different segmenters.

which is an indirect way of incorporating proximity distance [5].

6.3 IR Experiment Results

Table 2 shows the comparison of our two segmenters to ICTCLAS on the IR task. The performance of IR systems was measured by mean average precision (MAP) of the query set. The results show that Tight_Combine is better than the ICTCLAS segmentation, especially when using BM25. The relationship between Tight_Split and ICTCLAS is not clear.

In order to give a more in-depth analysis of the word segmentation methods with respect to the targeted phenomenon of semantic units, we classified the 200 queries into three categories according to their tightness as measured by function 1. The three classes are queries with tightness in ranges [$+\infty$, 10), [10, 1), and [1, 0), which contain 54, 41, and 108 queries respectively. Queries in the range [$+\infty$, 10) are tight queries, such as "弗吉尼亚" (Virginia). Queries in the range [1, 0) are loose queries, such as "广告公司" (advertising company). Other queries are those compounds which have ambiguous segmentations, such as "连锁反应" (chain reaction). Because the classification was based on the tightness measure, there are some errors. For example, "人民大学" (Renmin University) was classified as a loose query although it should at least be in the middle range. The three classes cover the whole tightness continuum, i.e. the whole possible query set. Table 3 shows the MAP with respect to these classes for the word segmentation methods. For queries with tightness less than 10, the results of ICTCLAS and Tight_Combine are approximately equal, which is not surprising since with few exceptions they have the same segmentation for both queries and documents.

[5] We also experimented with replacing β with the tightness value, but the results were not substantially different.

	[$+\infty$, 10)	[10, 1)	[1, 0)
BM25			
ICTCLAS	74.48%	60.28%	57.87%
Tight_Combine	86.44%	60.55%	57.70%
Tight_Split	88.86%	56.78%	53.17%
BM25_Beta			
ICTCLAS	84.60%	72.56%	63.28%
Tight_Combine	86.44%	72.70%	63.07%
Tight_Split	88.86%	74.80%	60.39%

Table 3. Results on three query categories.

For the interesting case of segmentation of tight units, i.e. queries in the range [$+\infty$, 10), the results show clear superiority for IR systems based on our segmentation methods. When using BM25, MAP is 86.44% for Tight_Combine, as compared to 74.48% for standard word segmentation . The advantage of Tight_Combine over ICTCLAS is that it combines units such as "平板玻璃" (plate glass) as the term is tight, while ICTCLAS segments that unit into "平板" (plate) and "玻璃" (glass). This is evidence that word segmentation models based on the tight measure are better than models trained on a human annotated corpus which ignored tightness information. Interestingly, Tight_Split is superior in the range [$+\infty$, 10), although the segmentation for these queries is the same as with Tight_Combine. When we analyzed the instances, we found it improved IR results of proper nouns. One possible explanation is that splitting of proper nouns such as "弗吉尼亚州" (Virginia state) in documents improved the recall even when the segmentation of the queries remained the same. For example, for query "弗吉尼亚" (Virginia), documents which contain "弗吉尼亚州" (Virginia state) should be retrieved. However, since ICTCLAS treats "弗吉尼亚州" as a word, those documents are missed. Instead, Tight_Split segments the sequence into "弗吉尼亚|州," which results in the retrieval of those documents.

In the range of [10, 1), the result is mixed. For some instances, Tight_Split is worse than Tight_Combine and ICTCLAS, as it segments queries such as "连锁反应" (chain reaction). However, in other instances, it is better than

Tight_Combine and ICTCLAS since it segments queries such as "国际象棋" (international chess). The result suggests that the setting of the threshold for non-compositional terms should be below 10.

In the range of [1, 0), the result is also mixed. One reason for the low performance of Tight_Split is that the tightness measure is not precise for those queries, which affects the segmentation. For example, splitting the queries "工人运动" (labor movement) and "中山大学" (Zhongshan University) decreases the IR performance dramatically. In future work, we would like to investigate this problem by segmenting queries manually according to their tightness. If the manual segmentation is superior, it would provided evidence for the hypothesis that segmentation based on tightness is superior.

The difference between BM25 and BM25_Beta in the range [10, 1) suggests that for Chinese IR, it is better to segment text in a more fine-grained way, and combine terms through a score function. For example, for queries such as "连锁反应" (chain reaction), for which splitting the unit is worse, BM25_Beta decreases the negative effect of splitting dramatically. For the query "人寿保险" (life insurance), when using BM25, Tight_Split is worse than ICTCLAS (average precision 0.59 vs. 0.66); but when using BM25_Beta, it is better than ICTCLAS (average precision 0.72 vs. 0.66).

7 Conclusion

For Chinese IR, we have developed a new method to segment documents based on the tightness of Chinese semantic units. The segmentation performance of our method is close to ICTCLAS, but the mean average precision of IR systems using our method is higher than for ICTCLAS when using BM25. In addition, we proposed a fine-grained segmenter plus a score function that prefers short proximity distance for CIR.

In the future, we plan to employ ranking SVM models with the tightness measure as one of the features for segmentation (Liu *et al.*, 2008). We hope that it can predict the tightness more precisely, by combining with other features. In terms of our test collection, the 203 query set clearly helps the in-depth analysis for the performance of different IR systems on different queries. We also plan to gather more queries and more judged documents in order to further analyze the influence of the proper treatment of semantic units in Chinese information retrieval. A large query set could also make it possible to employ machine learning models for IR (Song *et al.*, 2009).

References

Braschler, Martin, and Bärbel Ripplinger. 2004. How effective is stemming and compounding for German text retrieval? *Information Retrieval*, 7(3/4), 291-316.

Carterette, Ben, James Allan, and Ramesh Sitaraman. 2006. Minimal Test Collections for Retrieval Evaluation. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 268-275.

Chang, Pi-Chuan, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. *Proceedings of the Third Workshop on Machine Translation*, 224-232.

Foo, Schubert and Hui Li. 2004. Chinese word segmentation and its effect on information retrieval. *Information Processing and Management: an International Journal*, 40(1), 161-190.

Guenthner, Frantz and Xavier Blanco. 2004. Multilexemic expressions: an overview. *Linguisticae Investigationes Supplementa,* 239-252.

Halpern, Jack. 2000. Is English Segmentation Trivial? *Technical report, CJK Dictionary Institute*.

Hatcher, Erik and Otis Gospodnetić 2004. *Lucene in Action*. Manning Publications Co.

Huang, Xiangji, Stephen Robertson, Nick Cercone, and Aijun An. 2003. Probability-Based Chinese Text Processing and Retrieval. *Computational Intelligence*, 16(4), 552-569.

Huang, Xiangji, Fuchun Peng, Dale Schuurmans, Nick Cercone, and Stephen E. Robertson. 2003. Applying Machine Learning for Text Segmentation in Information Retrieval. *Information Retrieval, 6 (3-4), pp. 333-362, 2003*.

Jiang, Wenbin, Liang Huang, Qun Liu, and Yajuan Lv. 2008. A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*.

Liu, Yixuan, Bin Wang, Fan Ding, and Sheng Xu. 2008. Information Retrieval Oriented Word Segmentation based on Character Associative Strength Ranking. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 1061-1069.

McCarthy, Diana, Bill Keller, and John Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. *Proceedings Of the ACL-SIGLEDX (a Special Interest Group on the Lexicon Workshop) on Multiword Expressions*, 73-80.

Nie, Jian-Yun, Jiangfeng Gao, Jian Zhang, and Ming Zhou. 2000. On the use of words and N-grams for Chinese information retrieval. *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, 141-148.

Packard, Jerome L. 2000. *Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge University Press.

Peng, Fuchun, Xiangji Huang, Dale Schuurmans, Nick Cercone, and Stephen E. Robertson. 2002. Using Self-supervised Word Segmentation in Chinese Information Retrieval. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 349-350.

Peng, Fuchun, Xiangji Huang, Dale Schuurmans, and Nick Cercone. 2002. Investigating the Relationship between Word Segmentation Performance and Retrieval Performance in Chinese IR. *Retrieval Performance in Chinese IR, Coling2002*, 1-7.

Shi, Lixin and Jian-Yun Nie. 2009. Integrating phrase inseparability in phrase-based model. *Proceedings of the 32th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 708-709.

Silva, Joaquim, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. 1999. Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. *In Proceedings of 9th Portuguese Conference in Artificial Intelligence (EPIA 1999)*, 849.

Sproat, Richard, Chilin Shih, William Gale and Nancy Chang. 1996. A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. *Computational Linguistics*, 22(3), 377-404, 1996.

Song, Young-In, Jung-Tae Lee, and Hae-Chang Rim. 2009. Word or Phrase? Learning Which Unit to Stress for Information Retrieval. *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, 1048-1056.

Tao, Tao and ChengXiang Zhai. 2007. An exploration of proximity measures in information retrieval. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 295-302.

Vechtomova, Olga. 2001. Approaches to using word collocation in information retrieval. Ph.D. Thesis (City University, 2001).

Xu, Ying, Christoph Ringlstetter, and Randy Goebel. 2009. A Continuum-based Approach for Tightness Analysis of Chinese Semantic Units. *Proc. of the 23rd Pacific Asia Conference on Language, Information and Computation*, 569-578.

Zhang, Hua-Ping, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, 184-187.