Comparing Representations in Chinese Information Retrieval

K.L. Kwok

Computer Science Dept., Queens College, CUNY
65-30 Kissena Blvd., Flushing, NY 11367, USA
Email: kklqc@cunyvm.cuny.edu

Abstract

Three representation methods are empirically investigated for Chinese information retrieval: 1-gram (single character), bigram (two contiguous overlapping characters), and short-word indexing based on a simple segmentation of the text. The retrieval collection is the approximately 170 MB TREC-5 Chinese corpus of news articles, and 28 queries that are long and rich in wordings. Evaluation shows that 1-gram indexing is good but not sufficiently competitive, while bigram indexing works surprisingly well. Bigram indexing leads to a large index term space, three times that of short-word indexing, but is as good as short-word indexing in precision, and about 5% better in relevants retrieved. The best average non-interpolated precision is about 0.45, 17% better than 1-gram indexing and quite high for a mainly statistical approach.

1. Introduction

While information retrieval (IR) in English has over thirty years of history, IR in Chinese is relatively recent. It is well-known that written Chinese consists of strings of ideographs separated by punctuation signs. An ideograph (or character) can function as a word with meaning(s), or it can act as an alphabet forming a 'short-word' with one or more adjacent characters and having more specific meaning. Short-words can be strung together to form 'long-words' that have more complex and precise meaning. Words with little content are stopwords and usually manifest as single ideograph or short-words. Determining the boundaries of single or multi-character words in a string, a process called segmentation, is difficult because no delimiter or white space is used in the text and one has to rely on context, see for example [ChKi92,JiCh95,WuTs95,NiBR96,PoCr96,SSGC96,SuSH97].

Since it has been known in English IR that retrieval using words can provide a sound basis from which other methodologies can improve on, it appears that successful segmentation of Chinese documents and queries into words for representation to diminish ambiguity may be an important initial step for Chinese IR. From a linguistics point of view, segmentation usually means determining the longest words with precise meaning in a character string. The corresponding English constructs for these long-words are often noun phrases. These long-words may be unambiguous and pleasing to read, but for retrieval they lead to the disadvantage of having to determine partial matching values when a query short-word matches part of a document long-word or vice versa. For IR purposes, where we are working at the content term level, it appears that using short-words for representation would be adequate as a first step.

Because text segmentation is not straightforward and the process itself can have ambiguous outcomes, several previous attempts make use of single characters or all bigrams (adjacent overlapping character pairs) as representation [Chie95,LiLy96,BGHR9x, BuSM9x] in Chinese and in Japanese [FuCr93,OgIw95]. These approaches are simple and efficient, provide exhaustive indexing, and do not rely on having a segmentation procedure nor large dictionaries. Thus, it would be interesting to compare the effectiveness of retrieval among these three types of representation. IR concerns the detection of relevant documents, not answering questions. Effective IR can be obtained using indexing features with appropriate statistical properties. Good indexing features may not necessarily be good linguistic constructs that are needed for human comprehension. Thus, if 1-gram or bigrams can provide comparable effectiveness, segmentation may not be as important as thought to be for IR.

2. Representation Methods for Chinese Text

2.1 Single Characters

The simplest method of indexing a piece of Chinese text is to use single characters or 1-gram as index terms. Punctuation signs are deleted but stopword removal is not performed. This guarantees that if there are correct word matches between queries and documents there will be 1-gram matches, and should be good for recall. However, single characters are ambiguous in meaning and would adversely affect precision. Also, the set of commonly used Chinese characters are fairly small: the GuoBiao-2312 coding scheme supports 6763 only. Our indexing includes some codes for English characters/words as well as some other symbols and ends up with 8093 for our collection. The usage pattern of these index terms over a large collection could be quite different from that of English content terms. Because of the imprecision associated with single characters, the result of 1-gram retrieval may provide a basis on which one can measure improvements by other representation methods.

2.2 Bigrams

Another popular and simple method of representation is to use all contiguous overlapping 2-character pairs as indexing features [Chie95,LiLy96]. Again no stopword removal is performed, only punctuation signs are deleted. [Lin72] estimates that 80% of

modern Chinese words are bisyllable, while [He73] observes that 67.8% of tokens and 33.68% of word types are two characters based on a 1 million word corpus. Thus, with bigrams most of the correct Chinese words in a piece of text will be generated and they are much more specific in meaning than single characters. The drawback is that many meaningless character-pairs would also be produced and they could lead to noisy matchings between queries and documents, adversely impacting precision. With the number of commonly used single characters being over 6700, there could in theory be 40 million or more bigrams. Even 2.5 percent of this would result in a million pairs, potentially much larger than normal common English content terms for a similar size collection.

2.3 Simple Segmentation of Text

Because Chinese long-words are difficult to isolate correctly and would lead to a partial matching dilemma during retrieval, we believe that short-words would be the appropriate level of representation to use and they are comparable to English words. Although the bigram approach (Section 2.2) can account for most of the Chinese short-words, it also generates a large number of un-intended terms that may adversely affect precision and efficiency. Our strategy is to segment texts into meaningful short-words of one to four characters long to be used as indexing terms and perform exact matching during retrieval [NiBR96,AlCC9x, KwGr9x]. To achieve segmentation for a sentence in a relatively efficient manner, we implement a four-step procedure with a mixture of expert knowledge and statistical processing as follows:

* Step 1 uses 'facts' - lookup on a small (about 2000) manually created lexicon list of commonly used short-words of one to three characters. Some 4-character names and proper nouns are also included. Each item is tagged as useless (stopword), useful, numeric, punctuation and a few other codes. Given a piece of text, a window of maximum 4-character size slides through it and the longest match on the lexicon is identified. The matched entry could be a stopword, or it could be useful in which case it is retained, and this results in breaking a sentence into smaller chunks of texts. During searching of the lexicon, multiple entries of different sizes might match the text at the same position, causing ambiguities. A strategy of conflict resolution such as an HMM model [PoCr96] or using the path with the fewest components [NiBR96] can be implemented, but it is not done in our simple approach.

* Step 2 involves 'rules' - common ad-hoc language usage rules, also manually determined, are then employed to further split the chunks into short words of 2 or 3 characters. Examples of rules that serves to segment text are: a) any two adjacent similar characters XX; b) AX, where A is a small set of special characters; c) a remaining sequence of even number of characters is segmented two by two. These rules are of course not perfect, but they may be correct sufficiently often.

* Step 3 applies 'frequency filtering' - a first pass through the test corpus using steps 1 and 2 results in breaking the Chinese text strings into short-words of one to four characters long (some longer). Many of them may be meaningless. A threshold on the frequency of occurrence in the corpus is then used to extract the most common ones. These become our additional legitimate short-words.

* Step 4 is 'iteration' - we then expand the initial lexicon list in step 1 using those discovered in step 3 (which are all marked useful) and re-process corpus. The result is indexing of the corpus based on common short-words in general and those discovered in the corpus. The iteration process can be repeated in theory, but only one round is done and we believe it is sufficient.

Our procedure tries 'data-mining' the useful short-words from a collection and no training corpus is needed. Naturally no segmentation algorithm is perfect and our simple approach is no exception. For IR purposes, the error rate may be tolerable. The words detected in the above 4-step procedure are used for document and query representation directly. A major advantage is that the number of distinct terms for indexing should be substantially less than with bigrams, which is one purpose for segmentation. In addition, to safeguard against wrong segmentation and therefore missing matchings that might affect recall, we further break up each identified short-word into single characters and use them for representation as well. Unlike 1-gram representation in Section 2.1, many common single characters would not appear as indexing terms because of stopword removal, or their usage statistics is vastly different. Other investigators just add all single characters [Chie95,LiLy96]. With this scheme, we try to achieve both specificity and generality in representation.

3. The Chinese Retrieval Environment

Over the past several years the NIST-DARPA sponsored Text REtrieval Conference (TREC) have provided the IR community with large, realistic document collections for experimentation as well as unbiased manual relevance judgments between queries and documents for evaluation purposes. In the 1996 TREC-5 cycle [Harm9x], Chinese language IR experiments were included for the first time. This provides an opportunity to study these representation issues employing a large collection with respect to a set of queries that has unbiased judgments for evaluation.

3.1 Document and Query Collections

The collection of documents consists of the TREC-5 24,988 Xinhua and 139,801 People's Daily news articles. To guard against very long documents which can lead to outliers in frequency estimates, these are segmented into subdocuments of about 550 characters in size ending on a paragraph boundary. The total number of subdocuments is 231,527. These subdocuments are indexed in three ways as discussed in Section 2: single characters, bigrams and segmented short-words. The number of unique index terms are respectively: 8093, 1,482,172 and 494,288. Our segmented text approach reduces the index term size by about 1/3 compared to bigrams.

There are 28 queries, mostly on current affairs, and examples are shown in **Fig.1**. They are indexed like documents using the three derived index term sets. **Fig.1** also shows how our simple segmentation procedure segments it with all stopwords kept. We also did a manual segmentation on the 28 queries based on short-word identification, and which are regarded as correct. Evaluation shows that we achieve a recall of 91.3% and precision of 83%. The worst behavior is, as expected, with name recognition such as for countries, person, organization, etc. unless they happen to be on our lexicon (see for example Query #11, **Fig.1**). Numeric entities are also erroneously separated into single characters, but for retrieval purposes they are removed as stopwords anyway. The average size of queries are 19.3, 74.2 and 24.8 based on the three indexing procedures. These are long queries, rich in words and concepts.

3.2 PIRCS Retrieval Engine

For retrieval, we use our PIRCS (acronym for Probabilistic

Indexing and Retrieval - Components - System) engine that has been documented elsewhere [Kwok90,95]. Our practice for ad-hoc retrieval in English is based on two stages and is also employed for Chinese. The first is the initial retrieval where a raw query is used directly. The d best-ranked documents for this retrieval are then regarded as relevant without user judgment, and employed for feedback to train the initial query term weight and to do query expansion. This process has been called pseudo-feedback. This expanded query retrieval then provides the final result. This second retrieval in general provides substantially better results than the initial if the initial retrieval is reasonable and has some relevant documents within the best d. Since the raw Chinese queries are long and rich, we expect our two stage strategy to work well.

4. Results and Discussion

As is with IR in general, many experimental parameters need to be investigated and optimized for Chinese retrieval even though the general probabilistic model is retained. For example, because of the diverse indexing term set sizes for the three representations, an upper threshold for removing high frequency terms need to be determined for each. Because the initial retrieval quality may vary, the number of d-best documents used for pseudo-feedback may also vary. Also the number and types of terms to be used for query expansion need also be studied. These issues are investigated in the following sub-sections. The purpose of these experiments is not to do an exhaustive search of parameter values (which would lead to overfitting just for this collection), but to give a general idea of what parameter ranges would be reasonable and appropriate.

4.1 Retrieval Based on 1-gram Indexing

We first investigated a suitable upper threshold for 1-gram indexing by doing a set of initial retrievals using the raw queries at various cut-offs from document frequency 15k to 35k. Since the total number of subdocuments in our collection is about 230k, this represents a range from about 7% to 16%. Lower threshold is fixed at 3. The results are tabulated in **Table 1**. It can be seen that at a threshold of about 30k, the initial retrieval appears to do well in most measures except the relevants retrieved at 1000 documents. Fewer than 25k is not good.

With thresholds set at 30k3, we next investigate the number of 'feedback' documents d to use for our 2nd stage query expansion retrieval. This would also depend on the number of feedback terms to use. To make experimentation manageable we decide to use a fix value of 40 terms at this point, since this value has given reasonable results in other experiments. The result is tabulated in **Table 2**. It can be seen that d=20 returns the most favorable results, and varying d above 20 does not substantially change the picture. Comparing this with the initial retrieval at column (30k3) in **Table 1**, we see that significant improvements are made in pseudo-feedback retrieval as in the English case, thus: 6.4% in relevants retrieved (1917 vs 1802), 18.6% in average precision (.384 vs .324), and 13.2% in precision @ 30 (.455 vs .402).

Hi-Thold	15k3	20k3	25k3	30k3	35k3
RR	1558	1551	1731	**1802**	1810
AvP	.261	.295	.309	**.324**	.315
P@10	.379	.418	.432	**.471**	.469
P@20	.355	.389	.423	**.438**	.425
P@30	.339	.364	.391	**.402**	.394
RP	.286	.317	.338	**.352**	.352

RR = Tot Relevants at 1000 Docs Retrieved (Maximum 2182).
AvP = Average Non-Interpolated Precision.
P@d = Average Precision at d = 10,20,30 Docs Retrieved.
RP = Average Precision at Exact #of Relevants Retrieved.

Table 1: 1-gram Indexing 1st Stage Retrieval - Results Averaged over 28 Queries for Various High Frequency Thresholds (Lower Threshold = 3).

FB #Doc	15d40t	20d40t	30d40t	40d40t	50d40t
RR	1888	**1917**	1909	1909	1911
AvP	.362	**.384**	.377	.378	.378
P@10	.521	**.550**	.532	.550	.554
P@20	.486	**.495**	.489	.491	.486
P@30	.439	**.455**	.446	.450	.450
RP	.375	**.400**	.392	.393	.391

RR = Tot Relevants at 1000 Docs Retrieved (Maximum 2182).
AvP = Average Non-Interpolated Precision.
P@d = Average Precision at d = 10,20,30 Docs Retrieved.
RP = Average Precision at Exact #of Relevants Retrieved.

Table 2: 1-gram Indexing 2nd Stage Retrieval - Results Averaged over 28 Queries for Various Number of 'Feedback' Documents.

With the feedback document number fixed at 20, we finally vary the number of terms t to expand our queries. t ranges from 20 to 60 and are tabulated in **Table 3**. It appears that one should use at least 40 expansion terms and any number between 40 to 60 give quite similar results. We regard the 20d50t column as the representative effectiveness that can be achieved using 1-gram indexing. The average precision of 0.386, relevants retrieved of 1919 (88% of the pooled 2182), and nearly 17 of the first 30 documents retrieved are relevant means very good retrieval. It is surprising that single Chinese characters, though highly ambiguous, can give such respectable results. It appears they have retrieval power similar to English words.

4.2 Retrieval Based on Bigram Indexing

Similar testing was done for bigrams as for 1-gram indexing. The optimal high document frequency threshold is found to be 20k, significantly different from character indexing and results are

FB #Term	20d20t	20d30t	20d40t	20d50t	20d60t
RR	1895	1908	1917	**1919**	1921
AvP	.376	.379	.384	**.386**	.384
P@10	.521	.529	.550	**.543**	.543
P@20	.482	.484	.495	**.493**	.496
P@30	.444	.454	.455	**.456**	.457
RP	.395	.397	.400	**.401**	.400

RR = Tot Relevants at 1000 Docs Retrieved (Maximum 2182).
AvP = Average Non-Interpolated Precision.
P@d = Average Precision at d = 10,20,30 Docs Retrieved.
RP = Average Precision at Exact #of Relevants Retrieved.

Table 3: 1-gram Indexing 2nd Stage Retrieval - Results Averaged over 28 Queries for Various Number of 'Feedback' Terms.

Hi-Thold	15k3	20k3	25k3	30k3
RR	1995	**2021**	2024	2010
AvP	.345	**.350**	.346	.340
P@10	.482	**.493**	.482	.475
P@20	.464	**.466**	.461	.454
P@30	.421	**.435**	.433	.433
RP	.385	**.388**	.384	.383

RR = Tot Relevants at 1000 Docs Retrieved (Maximum 2182).
AvP = Average Non-Interpolated Precision.
P@d = Average Precision at d = 10,20,30 Docs Retrieved.
RP = Average Precision at Exact #of Relevants Retrieved.

Table 4: Bigram Indexing 1st Stage Retrieval - Results Averaged over 28 Queries for Various High Frequency Thresholds (Lower Threshold = 3).

tabulated in **Table 4**. This cut-off is about 9% of all 230k subdocuments, and is more like English IR. For 1-gram (Section 4.1), the index term space is only about 8k. Many useful content-bearing characters also attain high frequencies, hence the higher threshold of 30k.

The initial retrieval effectiveness is better than those in Table 1 for 1-gram. For example, comparing the results for 20k3 bigram with 30k3 1-gram, we see improvements of 12.2% in relevants retrieved (2021 vs 1802), 8.2% in average precision (.350 vs .324), and 8% in precision @ 30 (.435 vs .402). We attribute this to bigram indexing terms being more specific and useful for retrieval purposes. Moreover, recall of relevants improve also.

For the d-best 'feedback' documents we fix d = 40 since it has been shown to be good when we experimented initially with 19 queries only. Moreover, results do not vary appreciably for d = 25 to 50. **Table 5** tabulates results as we vary the number of terms t used for query expansion. It is interesting to note that effectiveness continues to increase with the number of query expansion terms. This is reminiscent of previous observation with massive expansion in routing [BuSA94] and probably is similar because the initial retrieval is so good for these rich and long queries: at the 20k3 column of **Table 4**, the precision @ 40 should be about 0.4, meaning about 16 of the 40 'feedback' documents are relevant on average. Pseudo-feedback again provides very significant improvements over initial retrieval: comparing the 20k3 column results of **Table 4** with the 40d120t expansion of **Table 5**, we see improvements of 5% in relevants retrieved (2125 vs 2024), 28% in average precision (.448 vs .350), and 19.8% in precision @ 30 (.521 vs .435). We did not pursue higher expansions because of time constraints, but it appears not to be changing much.

FB #Term	40d20t	40d40t	40d60t	40d80t	40d100t	40d120t
RR	2104	2114	2121	2122	2122	**2125**
AvP	.426	.435	.443	.445	.448	**.448**
P@10	.571	.571	.582	.604	.607	**.604**
P@20	.509	.534	.554	.558	.555	**.563**
P@30	.495	.513	.513	.513	.518	**.521**
RP	.439	.448	.447	.455	.453	**.456**

RR = Tot Relevants at 1000 Docs Retrieved (Maximum 2182).
AvP = Average Non-Interpolated Precision.
P@d = Average Precision at d = 10,20,30 Docs Retrieved.
RP = Average Precision at Exact #of Relevants Retrieved.

Table 5: Bigram Indexing 2nd Stage Retrieval - Results Averaged over 28 Queries for Various Number of Expansion Terms.

Hi-Thold	15k3	20k3	25k3	30k3
RR	1947	**1944**	1932	1927
AvP	.380	**.392**	.381	.374
P@10	.536	**.546**	.532	.511
P@20	.480	**.484**	.470	.463
P@30	.451	**.457**	.439	.437
RP	.397	**.403**	.401	.398

RR = Tot Relevants at 1000 Docs Retrieved (Maximum 2182).
AvP = Average Non-Interpolated Precision.
P@d = Average Precision at d = 10,20,30 Docs Retrieved.
RP = Average Precision at Exact #of Relevants Retrieved.

Table 6: Segmented Text Indexing Initial Retrieval - Results Averaged over 28 Queries for Various High Frequency Threshold (Lower Threshold = 3).

Apparently, even though there is generated a lot of character-pairs (~1.5 million) from the collection, bigrams do not lead to much noisy matchings between queries and documents. We regard the 40d120t column **Table 5** as the representative effectiveness that can be achieved using bigram-based indexing. It improves substantially over 1-gram results of **Table 3**: 10.7% in relevants retrieved (2125 vs 1919), 16.0% in average precision (.448 vs .386) and 14.3% in precision @ 30 (.521 vs .456).

4.3 Retrieval Based on Short-word Indexing

Using our segmentation procedure and the resulting short-words with characters as indexing terms, we obtain results in **Table 6** for the initial retrieval with various high frequency thresholds. Again, the same 20k threshold as for bigrams appears optimal for this collection. The initial retrieval average precision using short-word indexing is over 10% better than bigrams (.392 vs .350), and lesser improvements for precision. Relevants retrieved however is worse. We attribute this to bigrams doing much more exhaustive indexing.

As before, we fix the number of 'feedback' documents at 40 and perform the 2nd stage retrieval with various number of query

FB #Term	40d30t	40d40t	40d50t	40d60t	40d70t	40d80t
RR	2016	2022	2022	**2021**	2023	2024
AvP	.439	.443	.442	**.443**	.442	.442
P@10	.593	.593	.596	**.600**	.596	.593
P@20	.530	.538	.532	**.539**	.543	.536
P@30	.507	.505	.510	**.505**	.506	.507
RP	.433	.437	.435	**.440**	.439	.441

RR = Tot Relevants at 1000 Docs Retrieved (Maximum 2182).
AvP = Average Non-Interpolated Precision.
P@d = Average Precision at d = 10,20,30 Docs Retrieved.
RP = Average Precision at Exact #of Relevants Retrieved.

Table 7: Segmented Text Indexing 2nd Stage Retrieval - Results Averaged over 28 Queries for Various Number of Query Expansion Terms (any terms).

FB #Term	40d30t	40d40t	40d50t	40d60t	40d70t	40d80t
RR	1998	2006	2008	2010	2010	**2015**
AvP	.438	.446	.445	.446	.450	**.452**
P@10	.564	.579	.593	.600	.593	**.600**
P@20	.521	.529	.546	.546	.548	**.548**
P@30	.506	.514	.513	.518	.526	**.532**
RP	.438	.447	.443	.447	.447	**.452**

RR = Tot Relevants at 1000 Docs Retrieved (Maximum 2182).
AvP = Average Non-Interpolated Precision.
P@d = Average Precision at d = 10,20,30 Docs Retrieved.
RP = Average Precision at Exact #of Relevants Retrieved.

Table 8: Segmented Text Indexing 2nd Stage Retrieval - Results Averaged over 28 Queries for Various Number of Query Expansion Terms (t/2 single characters and t/2 not).

expansion terms. Two strategies are used: **Table 7** tabulates results using any best t terms, and **Table 8** shows the same using the best t/2 terms that are single character and t/2 that are not. Previously [BCPW et.al.96] has reported that bigrams for feedback is substantially better than characters. When using the first strategy for query expansion, effectiveness remains practically constant when 40 or more terms are used. For expansion using more specific non-single character short-words, effectiveness appears to increase up to 80 terms, and has a slight edge in precision over expansion using the first strategy. However, relevants retrieved also suffer slightly.

4.4 Summary of Results

We have gathered in **Table 9** a side-by-side comparison of the detailed recall precision values for the best 2nd stage retrieval results of the three representation methods. The interesting observation is that of the three representations, 1-gram indexing alone is surprisingly good but not sufficiently competitive. Bigrams perform very well, only slightly worse in precision but about 5.5% better in relevants retrieved compared to short-word indexing for this collection. It is not clear if better word segmentation can improve retrieval effectiveness. The achieved average non-interpolated precision of about 0.45 is quite high for this mainly statistical approach, and probably not likely to be improved much without more detailed syntactic and semantic considerations. If this holds up in further experiments and other environments, then word segmentation may not be as important for IR as first thought. Bigrams have the advantage of easy generation without the need to worry about segmentation correctness nor maintaining a large dictionary. The disadvantage is in efficiency: one has to deal with a large indexing space, and might have to use large query expansion to achieve good effectiveness.

5. Conclusion

Within the TREC-5 Chinese retrieval environment using our PIRCS engine and where the queries are long and rich, bigram indexing is effective and performs as well as short-word indexing. The latter approach is however more efficient, dealing with an indexing term space that is about one third smaller. 1-gram indexing is also good but not sufficiently competitive and performed about 17% worse in average precision and about 5 to 11% worse in relevants retrieved compared to the previous two

	Representation Method		
	1-gram	**Bigram**	**Short-Word**
#of Topics	28	28	28
Relevant:	2182	2182	2182
Rel_ret:	1919	2125	2010
Recall Level Precision Averages:			
0.0	.781	.855	.855
0.1	.615	.698	.652
0.2	.519	.610	.613
0.3	.483	.543	.565
0.4	.432	.497	.522
0.5	.393	.448	.472
0.6	.356	.410	.420
0.7	.318	.369	.362
0.8	.264	.326	.314
0.9	.181	.259	.249
1.0	.046	.081	.087
Non-Interpolated Average Precision			
	.386	.448	.452
Document Level Precision Averages:			
5 docs	.579	.643	.664
10 docs	.543	.604	.600
15 docs	.502	.579	.567
20 docs	.493	.563	.548
30 docs	.456	.521	.532
100 docs	.330	.380	.369
200 docs	.237	.279	.260
500 docs	.125	.142	.134
1000 docs	.069	.076	.072
Average Exact R-precision:			
	.401	.456	.452

Table 9: Details of the Best Retrieval Results for the Three Representation Methods.

representation methods. It appears that word segmentation may not be a pre-requisite for Chinese IR, although it is still worthwhile as a future investigation to see if better and more accurate word segmentation will influence retrieval results.

It does not mean that word segmentation is not needed for IR. In the future, when improvements in Chinese IR via thesauri, syntactic or semantic methods are performed, bigrams would probably be useless or much less effective than correct segmented words from text.

ACKNOWLEDGMENT

This work is partially supported by a Tipster grant from the Department of Defense. Mr. JunHui Xu assisted in part of the programming.

REFERENCES

[AlCC9x] Allan, J, Callan, J & Croft, W.B (199x). Inquery at TREC-5. In: The Fifth Text REtrieval Conference (TREC-5). Harman, D.K. (Ed.). to be published.

[BGHRWW9x] Beaulieu, M.M, Gatford, M, Huang, X, Robertson, S.E, Walker, S & Williams, P (199x). Okapi at TREC-5. In: The Fifth Text REtrieval Conference (TREC-5). Harman, D.K. (Ed.).

to be published.

[BCPW et.al.96] Boisen, S, Crystal, M, Petersen, E, Weischedel, R, Broglio, J, Callan, J, Croft, B, Hand, T, Keenan, T & Okurowski, M.E. Chinese information extraction and retrieval. In: Proc. of Tipster Text Phase 2 Workshop, May 6-8, 1996. pp.109-119.

[BuSA94] Buckley, C., Salton, G & Allan, J (1994). The effect of adding relevance information in a relevance feedback environment. In: Proc. 17th ACM SIGIR Conf. on R&D in IR. Croft, W.B & van Rijsbergen, C.J (eds.) Springer-Verlag: London. pp.292-300.

[BuSM9x] Buckley, C., Singhal, A & Mandar, M (199x). Using query zoning and correlation within SMART: TREC 5. In: Overview of the Fifth Text REtrieval Conference (TREC-5). Harman, D.K. (Ed.). to be published.

[ChKi92] Chen, K.J & Kiu, S.H (1992). Word identification for Mandarin Chinese sentences. Proc. Conference on Computational Linguistics. pp.101-7.

[Chie95] Chien, L.F (1995). Fast and quasi-natural language search for gigabytes of Chinese texts. In: Proc. 18th ACM SIGIR Conf. on R&D in IR. Fox, E., Ingwersen, P. & Fidel, R. (eds.) ACM:NY, NY. pp.112-120.

[Harm9x] Harman, D.K. Overview of the Fifth Text REtrieval Conference (TREC-5). to be published.

[He73] He, W.H (1973). Automatic recognition of Chinese words. Master Thesis, National Taiwan Institute of Technology, 1973. (Thanks to Richard Jones of InTEXT for this reference).

[JiCh95] Jin W. & Chen, L (1995). Identify unknown words in Chinese corpus. In: Proc. of 3rd NLP Pacific-Rim Symposium (NLPRS'95). Seoul, Korea. Vol.1, pp.234-9.

[KwGr9x] Kwok, K.L. & Grunfeld, L (199x). TREC-5 English and Chinese Retrieval Experiments using PIRCS. In: The Fifth Text REtrieval Conference (TREC-5). Harman, D.K. (Ed.). to be published.

[Kwok95] Kwok, K.L (1995). A network approach to probabilistic information retrieval. ACM Transactions on Office Information Systems, 13:325-353.

[Kwok90] Kwok, K.L (1990). Experiments with a component theory of probabilistic information retrieval based on single terms as document components. ACM Transactions on Office Information Systems, 8:363-386.

[LiLY96] Liang, T, Lee, S.Y & Yang W.P (1996). Optimal weight assignment for a Chinese signature file. Info. Proc. Mgmnt. 2:227-237.

[Lin72] Lin, Y.T (1972). Chinese English Dictionary of Modern Usage. Chinese University of Hong Kong Press: Hong Kong.

[NiBR96] Nie, J.Y, Brisebois, M & Ren, X (1996). On Chinese text retrieval. In: Proc. 19th Annual Intl. ACM SIGIR Conf. on R&D in IR. Frei, H.P, Harman, D, Schauble, P & Wilkinson, R (eds.) ACM:NY, NY. pp.225-233.

[OgIw95] Ogawa, Y & Iwasaki, M (1995). A new character-based indexing method using frequency data for Japanese documents. In: Proc. 18th ACM SIGIR Conf. on R&D in IR. Fox, E., Ingwersen, P. & Fidel, R. (eds.) ACM:NY, NY. pp.121-128.

[PoCr96] Ponte, J & Croft, W.B (1996). USeg: a retargetable word segmentation procedure for information retrieval. In: Symposium on document analysis and information retrieval (SDAIR '96).

[SSGC96] Sproat, R., Shih, C., Gale, W. & Chang, N (1996). A stochastic finite state word segmentation algorithm for Chinese. Computational Linguistics, 22:377-404.

[SUSH97] Sun, M., Shen, D. & Huang, C (1997). CSeg&Tag1.0: A practical word segmenter & POS tagger for Chinese texts. In: Proc. 5th Conference on Applied Natural Language Processing, Mar 31 - Apr 3, 1997. pp.119-124.

[WuTs95] Wu, Z & Tseng, G (1995). ACTS: An automatic Chinese text segmentation system for full text retrieval. J. ASIS, 46:83-96.

Query 1: Original English

U.S. to separate the most-favored-nation status from
human rights issue in China.

most-favored nation status, human rights in
China, economic sanctions, separate, untie

A relevant document should describe why the U.S.
separates most-favored nation status from
human rights. A relevant document should
also mention why China opposes U.S. attempts
to tie human rights to most-favored-nation status.

Original Chinese

美国决定将中国大陆的人权状况与其是否给予中共最惠国待遇分离.
最惠国待遇，中国，人权，经济制裁，分离，脱钩

相关文件必须提到美国为何将最惠国待遇与人权分离;
相关文件也必须提到中共为什么反对美国将人权与最
惠国待遇相提并论.

Segmented Chinese Text

美国 ¦ 决定 ¦ 将 ¦ 中国 ¦ 大陆 ¦ 的 ¦ 人权 ¦ 状况 ¦

与其 ¦ 是否 ¦ 给予 ¦ 中共 ¦ 最 ¦ 惠国 ¦ 待遇 ¦ 分离 ¦.

最 ¦ 惠国 ¦ 待遇 ¦ ， ¦ 中国 ¦ ， ¦ 人权 ¦ ， ¦ 经济 ¦ 制裁 ¦ ，

¦ 分离 ¦ ， ¦ 脱钩 ¦

相关文件 ¦ 必须 ¦ 提到 ¦ 美国 ¦ 为何 ¦ 将 ¦ 最 ¦ 惠国 ¦ 待遇

¦ 与 ¦ 人权 ¦ 分离 ¦ ;

相关文件 ¦ 也 ¦ 必须 ¦ 提到 ¦ 中共 ¦ 为 ¦ 什么 ¦ 反 ¦ 对 ¦ 美国

¦将 ¦ 人权 ¦ 与 ¦ 最 ¦ 惠国 ¦ 待遇 ¦ 相 ¦ 提并 ¦ 论 ¦ .

(^ means missing separator; ~ means spurious separator)

Fig.1a: TREC-5 Query 1 Showing Original English and Chinese Versions,
and Result of our Word Segmentation Procedure.

Query 11: Original English

UN Peace-keeping Force in Bosnia

Bosnia, Former Yugoslavia, Balkan, U.N.,
NATO, Muslim, weapon sanction, peace-keeping

A relevant document should contain information on
how UN peace-keeping troops carry out their mision
in the war-torn Bosnia.

Original Chinese

联合国驻波斯尼亚维和部队 .
波斯尼亚，前南斯拉夫,巴尔干,联合国,北约,武器禁运,维和,维持和平

相关文件必须包括联合国和平部队如何在战火蹂躏的波斯尼亚进行维持和平
的任务.

Segmented Chinese Text

联合国 ¦ 驻波 ¦ 斯 ¦ 尼亚 ¦ 维 ¦ 和 ¦ 部队 ¦ .

¦ 波斯尼亚 ¦ , ¦ 前南 ¦ 斯 ¦ 拉夫 ¦ , ¦ 巴 ¦ 尔 ¦ 干 ¦ , ¦ 联合国 ¦ ,

¦ 北 ¦ 约 ¦ , ¦ 武器 ¦ 禁运 ¦ , ¦ 维 ¦ 和 ¦ , ¦ 维持 ¦ 和平 ¦

相关文件 ¦ 必须 ¦ 包括 ¦ 联合国 ¦ 和平 ¦ 部队 ¦ 如何 ¦ 在 ¦

战火 ¦ 蹂躏 ¦ 的 ¦ 波斯尼亚 ¦ 进行 ¦ 维持 ¦ 和平 ¦ 的 ¦ 任务 ¦ .

(^ means missing separator; ~ means spurious separator)

Fig.1b: TREC-5 Query 11 Showing Original English and Chinese Versions
and Result of our Word Segmentation Procedure.