NTCIR-3 Chinese, Cross Language Retrieval Experiments Using PIRCS

K. L. Kwok
Computer Science Department, Queens College,
City University of New York, Flushing, NY 11367, USA
kwok@ir.cs.qc.edu

Abstract

We participated in the monolingual Chinese, English-Chinese cross language and multilingual retrieval tasks using our PIRCS retrieval system. For monolingual, bigram and short-word indexing (both with single characters) were employed for representation. Two separate retrieval lists were obtained and later combined as final result for some submissions. For cross-lingual and multilingual retrieval, only short-word indexing was used. We performed retrieval with two types of queries: queries from all sections of a topic, and from the description section only. The best monolingual mean average precision based on relax assessment is ~0.41 for long queries and ~0.36 for short description-only queries. These values are much less than those for NTCIR-2 and may indicate that NTCIR-3 environment is more difficult. For cross-lingual, we employed the query translation approach and concatenated outputs from MT-software and dictionary translation into one Chinese query. Results were also much inferior to those observed in NTCIR-2, achieving only about 56% of monolingual for long and 44% for short queries using relaxed judgment.

Post-judgment experiments show that monolingual retrieval can be improved for short-word indexing by employing a corpus-specific segmentation dictionary derived from the corpus itself. For cross-lingual retrieval, bigram indexing should also have been used to combine with short-word indexing. This can improve comparisons of cross language result with monolingual to 69% for long and 52% for short queries respectively.

Keywords: *monolingual Chinese retrieval; CLIR; MLIR; bigram indexing; short-word indexing.*

1 Introduction

We continue to employ our PIRCS retrieval system to participate in the NTCIR-3 CLIR track, performing the monolingual, cross-lingual and multilingual experiments. “Monolingual” means Chinese topics retrieving against Chinese documents; “cross-lingual” means using English topics to retrieve Chinese documents, and “multilingual” means using English topics to retrieve both English and Chinese documents and returning a single merged retrieval list. Each topic contains four textual sections: title (T), description (D), narrative (N) and concepts (C). Participants were required to return at least one result using D-queries formed from the description section only. Other returned results can freely use any or all of the topic sections to form queries, but only a maximum of three submissions were allowed for each separate task. There were 50 topics initially for the monolingual Chinese and the English-Chinese cross-lingual tasks. These were later trimmed to 42 for evaluation because some topics have few answer documents in the collection. For multilingual, 46 topics were used. The Chinese collections for retrieval came from United Daily News and CIRB001 sources totaling ~500 MB raw text. The English collection came from Mainichi, Taiwan News and Chinatimes English News totaling ~56 MB. All of them cover the period 1998-9.

2 Document Processing

The Chinese documents were pre-processed with our default procedure of breaking long documents into sub-documents of about 2000 bytes ending on a paragraph boundary. Sub-documents have been found to be useful for better pseudo-relevance feedback operation, and for retrieval with very long documents. Sub-document scores were combined after retrieval to report full document score for the final output rank list. Our collection contains 481,851 sub-documents, and were indexed two ways: first employing short-words with characters as index terms (to be referred to as short-word indexing: sw), and secondly using overlapping bigrams and 1-grams (to be referred to as bigram indexing: bi). The translation dictionary with 126,092 entries for cross language retrieval was also used as segmentation dictionary for convenience. The collection produced 139,572 unique index terms using short-word indexing. These were later trimmed to 71,830 using Zipf’s thresholds of 4 and 48,000. For bigram indexing, the corresponding unique terms were 2,900,147 before trimming and 1,246,287 after. Here, five common stop characters were used to gain more efficiency. In our system with some compression, short-word indexing without position information uses indexing space ~112% of raw document, while for bigram indexing it is ~250%. Much more is needed to include position information. Past and recent experiments have shown that bigram indexing is effective [1] but more costly. These two indexing structures were later used for our system to produce two separate retrieval lists.

PIRCS retrieval system is language independent once the documents and queries are processed into indexing terms. For the English collection, we used the same processing as in our TREC experiments – stop-word removal, Porter stemming, and breaking long documents into ~500 word sub-documents on paragraph boundaries. The statistics for the English collection are: 34,394 sub-documents, 166,803 unique index terms thresholded to 63,326.

3 Chinese Monolingual Retrieval

Monolingual Chinese retrieval is important in itself because evaluated experiments and query variety seen so far were much fewer compared to English. It is also important as a basis for measuring cross-lingual results. Recent cross-lingual retrievals have been shown to surpass monolingual [2,3], but they may be due to the monolingual basis being too low. We intend to establish a high basis by employing multiple indexing strategies that were successfully deployed in our TREC-5&6 Chinese retrieval experiments.

We submitted three official monolingual runs called **pircs-C-C-D-001** (description-only D-query using bigram indexing), **pircs-C-C-D-002** (D-query, and combining retrieval lists from bigram and short-word indexing), and **pircs-C-C-TDNC-003,** which is similar to 002 except that all sections of a topic are used to form a TDNC-query. We followed our English query processing practice to remove functional phrases (such as: 查询, 相关报道, 之文章 ..), and 'sentences' in the narrative section that contains references to 'irrelevant data' (such as: 不相关）. Table 3.1a,b show results for relax and rigid assessment respectively. Each run is identified with a shortened label such as bolded D-001 together with the indexing method employed. Additional un-submitted results were also tabulated with un-bolded row labels such as D-001w (description-only query using short-word indexing), and TDNC-001 and TDNC-001w (all-section query using bigram or short-word indexing respectively).

In Table 3.1, the rows with an exclamation ! for D-query using short-word indexing (D-001w) denote a faulty run in our system. The error was unknown at the time and discovered only after evaluation was made available. This batch produced essentially random ranked lists with small retrieval status values, which when combined with the bigram run (D-001) did not hurt the resultant (D-002) too much. For the TDNC-query, both the bigram and short-word indexing were shown as TDNC-001 and TDNC-001w and no error was made. The faulted D-query experiment was later repeated and tabulated in Table 6.1.

pircs-C-C-	**MAP**	**% imp**	**P@10**	**% imp**	**RPre**	**% imp**
D-001 bi	**.3617**	*****	**.5000**	*****	**.3773**	*****
!D- 001w sw	.0180	!	.0238	!	.0137	!
D-002 cmb	.3576	-1	.4976	-0	.3672	-3
TDNC-001 bi	.4060	*	.5548	*	.4202	*
TDNC-001w sw	.3583	-12	.5000	-10	.3724	-11
TDNC-003 cmb	**.4077**	**+0**	**.5571**	**+0**	**.4221**	**+0**

a) Relax Assessment

pircs-C-C-	**MAP**	**% imp**	**P@10**	**% imp**	**RPre**	**% imp**
D-001 bi	**.2928**	*****	**.3643**	*****	**.3009**	*****
!D-001w sw	.0082	!	.0143	!	.0121	!
D-002 cmb	.2902	-1	.3595	-1	.2955	-2
TDNC-001 bi	.3395	*	.4214	*	.3390	*
TDNC-001w sw	.2899	-15	.3833	-9	.3099	-9
TDNC-003 cmb	**.3435**	**+1**	**.4238**	**+1**	**.3448**	**+2**

b) Rigid Assessment

Table 3.1: Monolingual MAP Retrieval Results (Bolded row headings denote official submissions; '% imp'rovement calculated from a nearest row with * as basis; rows with ! denote faulty runs)

pircs-C-C-	**>AvgPre**	**=AvgPre**	**<AvgPre**
D-001	5, 25	0	12
TDNC-003	4, 34	0	4

a) Relax Assessment

pircs-C-C-	**>AvgPre**	**=AvgPre**	**<AvgPre**
D-001	28	0	14
TDNC-003	7, 28	0	7

b) Rigid Assessment

Table 3.2: Monolingual MAP Retrieval Results: Compared to Average of 33 Submissions

For this monolingual environment, it is observed that bigram indexing provides superior results, about 13% to 17% better mean average precision (MAP), compared to the short-word indexing for long queries (TDNC-001 vs. TDNC-001w). This has also occurred in the TREC-9 [4] and is different to our previous experience with TREC-5 and 6 collections. Post-judgment experiments in Section 6 describe the cause of this problem. Directly combining the retrieval status values of bigram and short-word runs using a factor of 0.5 for D-query and 0.6 for TDNC-query in favor of bigram does not materially change the bigram-only result. Thus, bigram indexing by itself is effective. This is good news in that bigram indexing does not need any segmentation dictionary. It does require more disk space and processing time, but the costs of these factors are being reduced constantly.

Using bigram-indexing results as reference, it is observed that the longer TDNC queries outperform the shorter D-queries as known before [5]. For example, MAP improves from 0.3617 (D-001) to 0.4060 (TDNC-001): over 12% for the relax assessment. Relax assessment provides about 20% better results than rigid assessment.

The RPre column represents precision at a retrieved number exactly equal to the number of relevant documents for each query. For perfect retrieval, the precision would have maintained a value of 1.0 from the first retrieved document up to this point. In comparison, our bigram run for short queries has dropped to 37.73%, and long queries to 42.02%, of this perfect value using relaxed assessment.

The comparison of our monolingual retrieval to the average of 33 submitted runs is tabulated in Table 3.2 using the MAP value. For example, our all-section query (TDNC-003) has 38 retrievals above average and 4 below for relax assessment. 4 of these 38 equals the best average precision achieved. The description-only query (D-001) retrieval has 30 above average with 5 being best, and 12 below. These have lower values with respect to the long queries because it is compared to all the 33 runs that include many that use longer and more effective queries such as TDNC type. The MAP value of this description-run (D-001) represents the top submitted result for this query size.

4 English-Chinese CLIR

Our translation strategy for cross language retrieval is essentially similar to NTCIR-2, but using concatenation of translation output from dictionary lookup and from MT-software. The translation dictionary is still the one from LDC [6], but MT-software is a package called HuaJian from Mainland China [7]. We replaced the Transperfect package used before because HuaJian actually gives better result for the NTCIR-2 experiments. (This was a bit surprising since it was assumed that Transperfect output should be more compatible with the text collection because both came from Taiwan and there could be more agreement in dialect.)

Three runs were submitted: **pircs-E-C-D-001** (description-only query with normal post-translation expansion), **pircs-E-C-D-002** (description-only query with both pre- and post-translation expansion), and **pircs-E-C-TDNC-003**, which is similar to pircs-E-C-D-001 except that all query sections were used. All runs use short-word indexing only. The reason is that we were pressed for time, and we thought that short-words could provide better precision. This turns out to be a wrong assumption as shown by the post-judgment bigram runs tabulated later in Table 6.2.

Table 4.1 shows our cross-lingual results with both intra-table improvement comparisons "% imp", and inter-table comparison with monolingual bigram-indexing runs pircs-C-C-D-001 and pircs-C-C-

pircs-E-C-	**MAP**	**% imp**	**P@10**	**% imp**	**RPre**	**% imp**
D-001 sw	.1334	*	.2214	*	.1629	*
% of mono D-001	37		44		43	
D-002 sw	**.1587**	**+19**	**.2643**	**+19**	**.1846**	**+13**
% of mono D-001	44		53		49	
TDNC-003 sw	**.2259**	**+69**	**.3571**	**+61**	**.2448**	**+50**
% of mono TDNC-001	56		64		58	

a) Relax Assessment

pircs-E-C-	**MAP**	**% imp**	**P@10**	**% imp**	**RPre**	**% imp**
D-001 sw	.1021	*	.1548	*	.1268	*
% of mono D-001	35		42		42	
D-002 sw	**.1150**	**+13**	**.1667**	**+8**	**.1381**	**+9**
% of mono D-001	39		46		46	
TDNC-003 sw	**.1751**	**+71**	**.2476**	**+60**	**.2047**	**+61**
% of mono TDNC-001	52		59		60	

b) Rigid Assessment

Table 4.1: Cross-lingual MAP Retrieval Results ('% imp'rovement calculated from the nearest row with * as basis; '% of Mono' rows compare results with D-001 bigram runs in Table 3.1)

pircs-E-C-	**>AvgPre**	**=AvgPre**	**<AvgPre**
D-002	1,21	0	2,18
TDNC-003	6,23	0	1,12

a) Relax Assessment

pircs-E-C-	**>AvgPre**	**=AvgPre**	**<AvgPre**
D-002	4,12	1	3,22
TDNC-003	1,30	0	1,10

a) Rigid Assessment

Table 4.2: Cross-lingual MAP Retrieval Results: Compared to Average of 16 Submissions

TDNC-001 of Table 3.1. We used monolingual bigram indexing as basis because it differs little from the combination run. It is seen that our standard cross-lingual D-001 run provides a low MAP value of 0.1334 using relax assessment. This is only 37% of our monolingual basis for these fairly short queries. Employing the TREC-8 disk-5 FBIS collection [8] to do pre-translation query expansion of 15 terms manages to produce between 12 to 19% improvement in precision values (pircs-E-C-D-002). The MAP value of 0.1587 is now about 44% of monolingual, still quite far from the ~55% seen in NTCIR-2 for title queries. The long query pircs-E-C-TDNC-003 run is better, giving a MAP of 0.2259, which is a 69% improvement over the short description-query result pircs-E-C-D-001. This reinforces our NTCIR-2 observation that longer queries are recommended for cross-lingual retrieval. Compared to monolingual result it achieves 56%, much less than the 75-85% for NTCIR-2. It appears that this NTCIR-3 environment is harder than the NTCIR-2 experiments. See also Section 6 for some improved post-judgment results that include bigram indexing.

Table 4.2 shows our pircs-E-C-TDNC-003 run compared to the average of 16 submissions. (Again, the pircs-E-C-D-002 values are distorted because it is compared to a mean in Table 4.2 that incorporates runs with more effective query types like TDNC).

5 English-English/Chinese MLIR

MLIR is a new sub-task in NTCIR-3. Its aim is to access both English and Chinese documents using an English query. Since retrieving English documents with English queries is not an issue, the problem is how to i) retrieve the Chinese documents; then ii) merge with the English list to form a single retrieval list. The first problem is similar to CLIR. One could translate the documents to English, but we use query translation since their results are already in place for CLIR in Section 4. We directly make use of pircs-E-C-D-001 and pircs-E-C-D-002 results of Section 4 as the retrieval lists from the Chinese collection. English monolingual retrieval is performed using English D-queries to form an English retrieval list pircs-E-E-D-002. This run has MAP value of .3709 and was not submitted.

The second issue is how to merge retrieval lists: one set coming from monolingual English retrieval and another coming from the Chinese collection via Chinese queries translated from the same English query. Because these are two queries of the same topic in different languages and the collection statistics are different, the document retrieval status values (RSV) from the two collections in general may not be comparable.

pircs-E-EC-	**MAP**	**%**	**P@10**	**%**	**RPre**	**%**
D-001 sw	.1620	*	.2848	*	.1988	*
D-002 sw	.1577	-3	.2870	+1	.1938	-3
D-001-p sw	.1346	*	.2457	*	.1746	*
D-003 sw	.1320	-2	.2348	-4	.1724	-1

a) Relax Assessment

pircs-E-EC-	**MAP**	**%**	**P@10**	**%**	**P@1K**	**%**
D-001 sw	.1198	*	.1761	*	.1446	*
D-002 sw	.1158	-3	.1739	-1	.1403	-3
D-001-p sw	.1020	*	.1609	*	.1342	*
D-003 sw	.0986	-3	.1630	-1	.1312	-2

b) Rigid Assessment

Table 5.1: Multilingual Retrieval Results (‘% imp’rovement calculated from the nearest row with * as basis)

Since the PIRCS model in theory evaluates a log odds value as the RSV for each document and it is related to its probability of being relevant, we simply assume that the RSV’s are comparable and directly merge the lists from pircs-E-C-D-002 (the E-C cross language run with pre-translation expansion) with pircs-E-E-D-002 to form a combined list that is our first submitted result: **pircs-E-EC-D-001.** Another list that was not submitted similarly merges pircs-E-C-D-001 (the E-C run without pre-translation expansion) with the same English pircs-E-E-D-002. This is shown in Table 5.1 as D-001-p. These are the basis runs from which evolved two other submissions with adjustments based on the following considerations.

There are 22927 full documents in the English corpus and 381681 in the Chinese. If one performs a random selection of 100 documents in the combined collection, the ratio of English to Chinese documents in the retrieval list would be quite small: r~0.06. For different samples, the number of English documents would approximate a Poisson distribution with mean and standard deviation of ~6 and ~2.4 respectively. Since topics are Asia-oriented, it is natural during retrieval to have more documents from Chinese sources than from English, and r could be smaller than expected. True document relevance in the English corpus may increase r above the mean. But, incompatible statistics between collections or false term matches may increase r substantially above expectation, which we assume should be avoided. For a query, if the English document occurrence ratio in the basis were higher or equal to a threshold, our strategy is to lower their RSV values in the basis so as to diminish this ratio; otherwise they are left untouched. We arbitrarily chose the threshold as 21 (mean+6σ) for the top 100 documents of the basis run, as these affect precision most. Adjustments were made as follows: RSV(new) = RSV(old)*f*g. f is chosen as a linear function of the number of English documents (Ne) in the top 100 of the basis list, and drops from about 0.95 (Ne=21) to 0.8 (Ne=100). The adjustment is made larger by g if the difference between the English and Chinese RSV’s is bigger. We use $g = 1-|E_{10}-C_{10}|/(E_{10}+C_{10})$, where E_{10} and C_{10}

are respectively the average top 10 RSV values of the English and Chinese lists. **pircs-E-EC-D-002** is the second submission employing this adjustment on the first basis. The third submission **pircs-E-EC-D-003** made adjustments on the second basis.

Results in Table 5.1 shows that the strategy failed since both adjusted runs are worse than just merging the RSV's from the two retrievals directly.

6 Post-Judgment Experiments

6.1 Chinese Monolingual Retrieval with New Short-Word Indexing

As discussed before, we had a faulty run in our monolingual D-query with short-word indexing. This experiment was corrected after relevance judgment is available and D-001w tabulated in Table 6.1 is the result. (Underscore is used to differentiate from submitted runs). We also show the result of combining its retrieval list with the bigram: D-002. This corrected short-word indexing retrieval has substantial deficit compared to bigram indexing (MAP .2732 vs. .3617 relax assessment), worse than those of its TDNC counterpart in Table 3.1. Its combination with the bigram run D-001, shown as D-002, produces a decrement (.3521 vs. .3617) in contrast to the submitted TDNC case in Table 3.1.

pircs-C-C-	MAP	% imp	P@10	% imp	RPre	% imp
D-001 bi	.3617	*	.5000	*	.3773	*
D-001w sw	.2732	-24	.4095	-18	.3071	-19
D-002 cmb	.3521	-3	.4857	-3	.3672	-3

a) Relax Assessment

pircs-C-C-	MAP	% imp	P@10	% imp	RPre	% imp
D-001 bi	.2928	*	.3643	*	.3009	*
D-001w sw	.2052	-30	.2952	-19	.2263	-25
D-002 cmb	.2824	-4	.3619	-1	.2955	-4

b) Rigid Assessment

Table 6.1: Monolingual Retrieval Results -- Corrected for D-001w (Bolded row headings denote official submissions; '% imp' rovement calculated from the nearest row with * as basis)

Why is the short-word indexing much worse than bigram indexing results? As noted in Section 2, we had used the translation dictionary for segmentation for convenience, and this turns out to be not suitable. A similar situation also occurred in our Trec-9 [4] experiments. In Trec-5 & 6, the segmentation dictionary was self-generated from the corpus. We further repeated the monolingual experiments by using the procedures described in our Trec-5 work [9] to generate from this NTCIR-3 Chinese corpus a dictionary of 86K size based on a seed dictionary of 33K. Results of retrieval using this segmentation dictionary for the same NTCIR-3 experiments are shown in Table 6.2.

pircs-E-C-	MAP	% imp	P@10	% imp	RPre	% imp
D-001 bi	.3617	*	.5000	*	.3773	*
D-001w sw	.3194	-12	.4405	-12	.3507	-7
D-002 cmb	**.3705**	**+2**	**.4952**	**-1**	**.3884**	**+3**
TDNC-003 bi	.4060	*	.5548	*	.4202	*
TDNC-003w	.3811	-6	.4952	-11	.4053	-4
TDNC-003 cmb	**.4102**	**+1**	**.5595**	**+1**	**.4257**	**+1**

a) Relax Assessment

pircs-E-C-	MAP	% imp	P@10	% imp	RPre	% imp
D-001 bi	.2928	*	.3643	*	.3009	*
D-001w sw	.2423	-17	.3214	-12	.2561	-15
D-002 cmb	**.2979**	**+2**	**.3786**	**+4**	**.3135**	**+4**
TDNC-003 bi	.3395	*	.4214	*	.3390	*
TDNC-003w	.3093	-9	.3762	-11	.3154	-7
TDNC-003 cmb	**.3402**	**+0**	**.4214**	**+0**	**.3376**	**-0**

b) Rigid Assessment

Table 6.2: Monolingual Retrieval Results -- Self-Generated Segmentation Dictionary for Short-word Indexing (Bolded row headings denote official submissions; '% imp' rovement calculated from the nearest row with * as basis)

It is seen that this self-generated segmentation dictionary has a large positive effect on short-word retrieval effectiveness, improving over 17% or more (D-001w MAP values: .3194 Table 6.2 vs. .2732 Table 6.1). These results are more comparable to, though still less than, those of bigram D-001 (MAP .3617 relax assessment). Combining the new short-word with bigram retrieval leads to results D-002 and TDNC-003 slightly better than both alone. They are also slightly better than those obtained previously using the translation dictionary for segmentation except for long query rigid assessment. This shows

pircs-E-C-	MAP	% imp	P@10	% imp	RPre	% imp
D-001 sw	.1494	*	.2381	*	.1687	*
D-001b bi	.1691	+13	.2262	-5	.1892	+12
% of mono D-001	46		46			
D-002 sw	.1784	+19	.2786	+17	.2075	+23
D-002b bi	.1909	+28	.2667	+12	.2045	+21
% of mono D-001	52		54		53	
D-002c bi+sw	**.1925**	**+29**	**.2929**	**+23**	**.2204**	**+31**
% of mono D-001	52		77		70	
TDNC-003 sw	.2447	*	.3786	*	.2581	*
TDNC-003b bi	.2538	+4	.3476	-8	.2743	+6
% of mono TDNC-001	**62**		62		64	
TDNC-003c bi+sw	**.2807**	**+15**	**.4143**	**+9**	**.3037**	**+18**
% of mono TDNC-001	**69**		74		72	

a) Relax Assessment

pircs-E-C-	**MAP**	**% imp**	**P@10**	**% imp**	**RPre**	**% imp**
D-001 **sw**	.1105	*	.1690	*	.1230	*
D-001b **bi**	.1161	+5	.1476	-13	.1318	+7
% of mono D-001	39		39		42	
D-002 **sw**	.1264	+14	.1690	+0	.1230	+0
D-002b **bi**	.1244	+13	.1595	-6	.1353	+10
% of mono D-001	42		42		43	
D-002c bi+sw	**.1314**	**+19**	**.1786**	**+6**	**.1509**	**+23**
% of mono D-001	44		47		48	
TDNC-003 sw	.1967	*	.2714	*	.2122	*
TDNC-003b **bi**	.1988	+1	.2524	-7	.2162	+2
% of mono TDNC-001	58		60		64	
TDNC-003c bi+sw	**.2214**	**+13**	**.2929**	**+8**	**.2438**	**+15**
% of mono TDNC-001	64		69		71	

b) Rigid Assessment

Table 6.3: Cross-lingual Retrieval Results using Bigram Indexing ('% imp'rovement calculated from the nearest row with * as basis; '% of Mono' compares result with combination runs in Table 6.2)

the usefulness of a segmentation dictionary tailored to the corpus for monolingual IR.

6.2 CLIR using Bigram Indexing and New Short-Word Indexing

Our submitted cross-lingual runs do not compare well with our own monolingual results, and with other top submissions. Initially it was assumed that our translation procedures might not be competitive. Quite surprisingly, it turns out that a large part of the deficit was simply because of inadequate indexing. Employing both bigram indexing and the short-word indexing with the self-generated segmentation dictionary, the same experiments were repeated and some of the new cross-lingual results are tabulated in Table 6.3.

The bigram results (MAP .1691 for D-001b, and .1909 for D-002b with pre-translation expansion, relax assessment Table 6.3a) are competitive with other submissions. Short-word indexing results also improve because of the new segmentation dictionary (e.g. Table 6.3a D-001 MAP .1494 vs. 1334 in Table 4.1a, D-002 MAP .1784 vs. .1587, and D-003 MAP .2447 vs. .2259). Combining them lead to a MAP of .1925 for D-type with pre-translation expansion, and .2807 for TDNC-type queries using relax assessment, Table 6.3a. These are about 52% and 69% of the best monolingual effectiveness based on combination strategy of Table 6.2. These ratios are about 10% worse than for NTCIR-2. For rigid assessment, bigram indexing is not as effective as using relax. Combination helps TDNC-type but not much for D-type queries.

7 Conclusions

Bigram indexing seems to work much better for Chinese monolingual and for cross-lingual retrieval as well in our NTCIR-3 experiments. For short-word indexing, a short cut to use the translation dictionary for segmentation is not effective. A self-generated word dictionary from the corpus itself as segmentation dictionary can improve retrieval effectiveness substantially. Additional improvements can be obtained by data fusion of these two retrieval lists.

Using rigid assessment, comparison of NTCIR-3 results show that they are approximately only half of what has been achieved in NTCIR-2, whether monolingual or cross-lingual. For example, NTCIR-2 monolingual long queries had a best MAP value of about ~.62 using rigid assessment, much better than the current value of ~.34. Similarly, NTCIR-2 cross-lingual long queries had a best MAP value of ~.48 compared to ~.22 achieved in NTCIR-3. It was stated that this year, the concept section of a topic was formed in a more casual, user-oriented way compared with last year's [10]. This may account for some of the differences. However, last year's TQN runs (queries without the benefit of concept sections) still achieve monolingual and cross-lingual MAP (rigid assessment) values of ~0.55 and ~0.38 respectively, more than 60% and 70% better than this year's. Presumably this NTCIR-3 Chinese retrieval environment is more difficult than before, and it will be interesting to explore the characteristics that render these experiments hard.

Appendix

A note on relevance judgment data: we casually looked at Topic 038: 'Provide documents that describe Asian reactions to the terrorist bombings of the U.S. embassies in Kenya and Tanzania'. This has the least number of documents relevant (6 according to relax and 3 for rigid assessment). Two of the relevant documents are listed as from United Daily News, viz.: udn_xxx_19980810_0186 and udn_xxx_19980817_0186. The first document however has the text stored in reverse and bottom to top order in the document file and our text processing procedures are not capable to handle such a situation. The second article appears not related to the topic.

Acknowledgment

This work was partially supported by the Space and Naval Warfare Systems Center San Diego, under grant No. N66001-1-8912.

References

[1] K.L. Kwok. Employing multiple representations for Chinese information retrieval. *Journal of the American Society for Information Science*, 50(8): 709-723, 1999.

[2] Gey, F & Chen, A (2001). TREC-9 Cross Language Information Retrieval (English-Chinese) Overview. In: *Information Technology: The Ninth Text Retrieval Conference (TREC-9)*. NIST SP 500-249, pp.15-23. GPO: Washington, D.C.

[3] Xu, J & Weischedel, R. (2001). TREC-9 cross-lingual retrieval at BBN. In: *Information Technology: The Ninth Text Retrieval Conference (TREC-9)*. NIST SP 500-249, pp.106-115. GPO: Washington, D.C.

[4] Kwok, K.L. An Analysis of the TREC-9 English-Chinese CLIR Experiments. To be published in proceedings of IEEE SMC'02 Conference.

[5] Kwok, K.L. NTCIR-2 Chinese and Cross Language Experiments using PIRCS. In: Proc. of Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Summarization, K. Eguchi, N. Kando & J.Adachi (eds.) pp.111-118, NII: Tokyo, 2001.

[6] http://www.morph.ldc.edu/Projects/Chinese (Linguistic Data Consortium).

[7] http://www.altlan.com

[8] Voorhees, E.M & Harman, D.K (2000). Overview of the Eigth Text Retrieval Conference (TREC-8). In: *Information Technology: The Eigth Text Retrieval Conference (TREC-8). NIST SP 500-246, pp.1-23, GPO: Washington, D.C.*

[9] Kwok, K.L. & Grunfeld, L. TREC-5 English and Chinese retrieval experiments using PIRCS. In: *Information Technology: The Fifth Text REtrieval Conference (TREC-5)*, E.M. Voorhees & D.K. Harman, eds. NIST Special Publication 500-238, US GPO: Washington, DC. pp.133-142, 1997.

[10] Chen, K-H, Chen, H-H, Kando, N, Kuriyama, K, Lee S, Myaeng, S.H, Kishida, K, Eguchi, K and Kim H. Overview of CLIR Task at the Third NTCIR Workshop. In: Working Notes of the Third NTCIR Workshop Meeting, pp.1-38, NII, Tokyo, October 8-11, 2002.