NTCIR-6 Chinese Monolingual and English-Chinese Cross Language Retrieval Experiments using PIRCS

Kui-Lam Kwok and Norbert Dinstl

Computer Science Department, Queens College,
City University of New York, Flushing, NY 11367, USA
kwok@ir.cs.qc.edu, emc21@earthlink.net

Abstract

In NTCIR-6, our Stage-1 results which consist of using old queries retrieving on a different old collection, were not official because of late submission. Stage-2 submissions, which consists of repeating previous experiments, were on time. These NTCIR-6 experiments were conducted as new without referring to any previous knowledge about the runs. Comparisons with old results however were less favorable for about half the runs. We traced this to the accidental use of an out-dated module which sets the Zipf high frequency threshold too low, and leads to too many high frequency terms being removed from a query. Some runs are new and not submitted previously by us. These include: 'title' queries for NTCIR-3 monolingual Chinese and English-Chinese CLIR, and the English-Chinese CLIR runs for NTCIR-4.

Keywords: *Monolingual Chinese IR; English-Chinese CLIR.*

1 Introduction

We participated in NTCIR-6 in both Stage-1 and -2 experiments with the Chinese collections. Because of time conflicts, we could not complete the Stage-1 experiments before the deadline. They were submitted later, about the same time as Stage-2 submissions. Stage-2 runs are on time.

Since NTCIR-6 CLIR tasks involved old queries and collections, there are specific email discussions and guidelines concerning the use of old knowledge [1]. We decided to perform all tasks as independent new ad hoc retrievals without incurring any knowledge from past NTCIR experiments.

Our retrieval strategy has not changed much during the past years. PIRCS algorithm and system was used for retrieval, which is a two-way activation-spreading implementation of the probabilistic retrieval model [2]. Two-way mean using a query as focus and spreading activation from documents through terms to it to evaluate query-focused retrieval status value (RSV), and using a document as focus and spreading activation from query to obtain document-focused RSV. Final RSV is a linear combination of the two, equivalent to combining a basic language model with a probabilistic retrieval model [3]. PIRCS also counts with document components (which are single terms) rather than whole documents for probability estimation. It employs neural network learning procedure to implement second stage pseudo-relevance feedback.

For Stage-1, two monolingual Chinese (RunID's: pircs-C-C-T-01, pircs-C-C-D-02), and two English-Chinese CLIR runs (RunID's: pircs-E-C-T-03, pircs-E-C-D-04) using 'title' and 'description' field as queries were completed. These results are presented in Section 2.

For Stage-2, similar runs were completed for all three sets of queries N3 (RunID's: pircs-C-C-T-01-N3, pircs-C-C-D-02-N3, pircs-E-C-T-03-N3, pircs-E-C-D-04-N3), N4 (RunID's: pircs-C-C-T-01-N4, pircs-C-C-D-02-N4, pircs-E-C-T-03-N4, pircs-E-C-D-04-N4), and N5 (RunID's: pircs-C-C-T-01-N5, pircs-C-C-D-02-N5, pircs-E-C-T-03-N5, pircs-E-C-D-04-N5). These are discussed in Section 3. Section 4 has some observations and conclusions.

2 Stage-1 Experiments

2.1 Monolingual Chinese IR

These Stage-1 results are not official because we submitted them after the deadline, and are included for reference. We performed two indexing for each collection, viz., bigram and 1-gram indexing as well as short-word and single character indexing. Long documents are segmented into subdocuments of about 3000 bytes ending on a paragraph boundary. Two retrieval lists were obtained using our probabilistic PIRCS engine for each query, which are linearly combined with a ratio of 6:4 in favor of bigram. Moreover, pseudo-relevance feedback was done by expanding each original query with 100 terms from top 10 retrieved documents. Retrieval using both 'title' and 'description' queries were done. (This approach was applied to all of our NTCIR-6 experiments for both monolingual and cross language runs.) Results are tabulated in Table 1. For example, for Rigid evaluation our MAP values for 'title' and

pircs-	R%	MAP	P10	P20	R.Pre
Rigid (#relevant = 2598)					
C-C-T-01	78	.2435	.3320	.3000	.2640
C-C-D-02	83	.2592	.3640	.3250	.2903
Relax (#relevant = 4405)					
C-C-T-01	79	.3399	.5020	.4640	.3634
C-C-D-02	84	.3660	.5420	.5110	.3908

Table 1: Stage-1 Monolingual Chinese IR Results for 50 Queries

pircs-	> median	=median	<median
Rigid (#relevant = 2598)			
C-C-T-01	31+4=35	0	15
C-C-D-02	22+6=28	7	15
Relax (#relevant = 4405)			
C-C-T-01	27+4=31	0	19
C-C-D-02	25+3=28	10	12

Table 2: Stage-1 Monolingual Chinese IR Results Compared with Median

'description' queries are .2435 and .2592 respectively In comparison, the median 'title' and 'description' MAP results from all sites [1] are: .2339, .2284 for Rigid evaluation, and .3262, .3385 for Relax respectively. Our results are above median, as also shown in Table 2 where the numbers of queries performing above, equal and below median are tabulated. The +n numbers under '>median' means number of queries equaling the best MAP attained. Our results are much less than the all-site maximum MAP results which are: .3097, .3136 (Rigid) and .4013, .4118 (Relax).

After the macro statistics of Stage-2 were available, it was discovered that an experimental processing module (which is less effective) was inadvertently used for *all* the runs in NTCIR-6 (see Sec.3). It is possible that results in Table 1 may be improved by a few percent.

2.2 English-Chinese CLIR

An English query (from either the 'title' or 'description' section of a topic) was translated two ways: first by Systran MT software, and secondly, entities were extracted by BBN's IdentiFinder, and these English entity terms were rendered into Chinese via our web-based entity-oriented translation/ transliteration procedure [4]. The two outputs were merged to form our translated queries. For queries from 'description', common English introduction phrases (e.g. 'Find articles') were also removed. Once a query is defined, retrieval was done as in Chinese monolingual processing using both bigram and short-word indexing, and pseudo-relevance feedback. Results are tabulated in Table 3. Except for the recall (R%), the precision values for 'title' queries attain the high-sixty percent of monolingual results,

pircs-	R%	MAP	P10	P20	R.Pre
Rigid (#relevant = 2598)					
E-C-T-03	64	.1686	.2300	.2080	.1896
%mono	82	69	69	69	72
E-C-D-04	71	.1671	.2220	.2120	.1873
%mono	86	64	61	65	65
Relax (#relevant = 4405)					
E-C-T-03	64	.2237	.3480	.3170	.2386
%mono	81	66	69	68	66
E-C-D-04	70	.2373	.3560	.3360	.2600
%mono	83	65	66	66	67

Table 3: Stage-1 English-Chinese CLIR Results for 50 Queries

pircs-	> median	=median	<median
Rigid (#relevant = 2598)			
E-C-T-03	11+14=25	0	25
E-C-D-04	16+14=30	0	20
Relax (#relevant = 4405)			
E-C-T-03	15+13=28	0	22
E-C-D-04	14+17=31	0	19

Table 4: Stage-1 English-Chinese CLIR Results Compared with Median

while the 'description' queries attain less at about the mid-sixty percent of monolingual. Our submissions compared to all-site median are tabulated in Table 4. It appears that 'description' query results have better comparison with median. Between 13 to 17 of the 50 queries attain the best average precision in our results. The corresponding all-site maximum MAP values for 'title' and 'description' queries are: .2013 and .1911 for Rigid, and .2931 and .2804 for Relax evaluation [1].

3 Stage-2 Experiments

Stage-2 tasks consist of repeating previous NTCIR experiments so as to make comparison based on methods of different years. Unfortunately, we retained only the old NTCIR-3 & 5 files and run parameters. NTCIR-4 files were lost, with only our workshop paper as record [5]. The following subsections summarize our results for N5, N4 and N3 monolingual and CLIR runs.

3.1 NTCIR-5 Experiments

3.1.1 Chinese Monolingual IR

Results of our current N5 runs are tabulated in Table 5. They are uniformly worse than those of last year's (NTCIR-5 [6] rows in italics) and were unknown to us until after Stage-2 macro statistics were distributed. We discovered that during preparation of the experiments, a previous trial

pircs-	R%	MAP	P10	P20	R.Pre
Rigid (#relevant = 1885)					
C-C-T-01-N5	86	.3741	.4520	.3750	.3619
NTCIR-5: T	*87*	*.3958*	*.4880*	*.3990*	*.3897*
C-C-D-02-N5	93	.3742	.4560	.3980	.3536
NTCIR-5: D	*94*	*.3897*	*.4780*	*.4090*	*.3727*
Relax (#relevant = 3052)					
C-C-T-01-N5	85	.4414	.5840	.5140	.4189
NTCIR-5: T	*86*	*.4651*	*.6080*	*.5410*	*.4485*
C-C-D-02-N5	90	.4477	.6000	.5420	.4311
NTCIR-5: D	*90*	*.4625*	*.6200*	*.5510*	*.4400*

Table 5: Stage-2 N5 Monolingual Chinese Retrieval Results for 50 Queries

processing module was erroneously linked instead of the latest version. This module creates a network in memory for activation spreading to calculate the retrieval status values of each document based on the connected terms (according to our PIRCS model). The number of connected edges can be substantial depending on the occurrence frequencies of terms activated by the query. In an effort to save memory space and time, we had experimented with reducing the number of network edges based on varying the Zipf's high frequency threshold (Zhi) with respect to the size of a query q, i.e. Zhi=α*N_d, N_d being the number of documents in the collection. This old module has a drastic policy: letting α=1 if |q|<=3, α=0.4 when |q| = 4 or 5, α=0.1 when |q| >5. This module applies to initial retrieval. It appears that too many query terms are filtered, and result in our current NTCIR-6 runs. This version has been superceded since, after larger machines are available, memory space is not as critical. The latest version of this module (used in NTCIR-5) employs a Zhi that screens out only a few of the highest frequency terms. The difference is 3-6% in MAP effectiveness when comparing these two.

pircs-	R%	MAP	P10	P20	R.Pre
Rigid (#relevant = 1885)					
E-C-T-03-N5	77	.2379	.2920	.2720	.2507
%mono	90	64	65	73	69
NTCIR-5: T	*75*	*.2459*	*.2880*	*.2530*	*.2478*
E-C-D-04-N5	82	.2423	.3180	.2740	.2389
%mono	88	65	70	69	68
NTCIR-5: D	*72*	*.2682*	*.3500*	*.2960*	*.2661*
Relax (#relevant = 3052)					
E-C-T-03-N5	76	.3060	.4220	.3910	.2945
%mono	89	69	72	76	70
NTCIR-5: T	*72*	*.2975*	*.4000*	*.3640*	*.2870*
E-C-D-04-N5	79	.2933	.4100	.3730	.2864
%mono	88	66	68	69	66
NTCIR-5: D	*80*	*.3235*	*.4380*	*.4050*	*.3275*

Table 6: Stage-2 N5 English-Chinese CLIR Results for 50 Queries

3.1.2 English-Chinese CLIR

The influence of the Zhi threshold (Sec.3.1.1) also applies to these CLIR experiments. As tabulated in Table 6, the 'title' query results are close to those obtained in NTCIR-5, but there are fairly substantial differences for 'description'. The MAP values for the latter drop by nearly 10% compared to old runs (.2423 vs. .2682 for Rigid, and .2975 vs. .3235 for Relax evaluation). However, both our 'title' and 'description' MAP values are the best among submitted sites.

3.2 NTCIR-4 Experiments

3.2.1 Chinese Monolingual IR

Results of our current N4 runs are shown in Table 7. The MAP values are fairly close to what was reported before [5]. Since our original NTCIR-4 processing files are no longer available, we cannot ascertain the reason for some of the larger differences like pircs-C-C-D-02-N4 P10 value of .2763 vs. previous NTCIR-4 value of .2475. Retrieval was done similarly as during NTCIR-3, which is discussed in Sec.3.3.

pircs-	R%	MAP	P10	P20	R.Pre
Rigid (#relevant = 1318)					
C-C-T-01-N4	82	.2060	.2441	.1839	.2040
NTCIR-4: T	*83*	*.2097*	*.2356*	*.1958*	*.2059*
C-C-D-02-N4	82	.2183	.2763	.2076	.2183
NTCIR-4: D	*85*	*.2150*	*.2475*	*.1975*	*.2010*
Relax (#relevant = 2085)					
C-C-T-01-N4	83	.2542	.3288	.2585	.2655
NTCIR-4: T	*84*	*.2673*	*.3373*	*.2864*	*.2725*
C-C-D-02-N4	83	.2818	.3797	.3059	.2990
NTCIR-4: D	*86*	*.2761*	*.3542*	*.2941*	*.2810*

Table 7: Stage-2 N4 Monolingual Chinese Retrieval Results for 59 Queries

pircs-	R%	MAP	P10	P20	R.Pre
Rigid (#relevant = 1318)					
E-C-T-03-N4	63	.1506	.1695	.1415	.1588
%mono	77	73	69	77	78
E-C-D-04-N4	69	.1427	.1661	.1339	.1476
%mono	84	65	60	64	68
Relax (#relevant = 2085)					
E-C-T-03-N4	65	.1847	.2475	.2042	.2040
%mono	78	73	75	79	77
E-C-D-04-N4	70	.1924	.2525	.2017	.2078
%mono	84	68	66	66	69

Table 8: Stage-2 N4 English-Chinese CLIR Results for 59 Queries

3.2.2 English-Chinese CLIR

During NTCIR-4, we did not submit English-Chinese CLIR results. However, the recorded runs from other sites had very low MAP values, the best being .0663 (Rigid evaluation) for the 'description' query [7]. Table 8 tabulates our N4 experiments for both 'title' and 'description' queries and results are much better. These also are the best results submitted from all sites.

3.3 NTCIR-3 Experiments

3.3.1 C-C Monolingual IR

Results of our current N3 runs are shown in Table 9. During the old NTCIR-3 time frame, 'title' queries were not submitted [8]. Also, the old runs made use of a single Zhi threshold, but they also did not use 'avtf' (average term frequency) weighting [9] of the query terms for initial retrieval. There was one 'title' run from other sites during NTCIR-3 [10] with a Relax MAP value of .2467 compared with our current MAP of .3180. For 'description' queries, the MAP values are quite close to our old runs, but the other precision values improved for the current runs.

pircs-	R%	MAP	P10	P20	R.Pre
Rigid (#relevant = 1928)					
C-C-T-01-N3	76	.2438	.2976	.2679	.2569
C-C-D-02-N3	81	.2940	.4000	.3429	.3232
NTCIR-3: D	*78*	*.2902*	*.3595*	*.3048*	*.2987*
Relax (#relevant = 3284)					
C-C-T-01-N3	72	.3180	.4452	.3976	.3343
C-C-D-02-N3	78	.3570	.5286	.4548	.3798
NTCIR-3: D	*75*	*.3576*	*.4976*	*.4167*	*.3749*

Table 9: Stage-2 N3 Monolingual Chinese Retrieval Results for 42 Queries

pircs-	R%	MAP	P10	P20	R.Pre
Rigid (#relevant = 1928)					
E-C-T-03-N3	64	.1609	.1810	.1679	.1606
%mono	84	66	61	63	63
E-C-D-04-N3	67	.1753	.2262	.2179	.1827
%mono	83	60	57	64	57
NTCIR-3: D	*54*	*.1150*	*.1667*	*.1476*	*.1381*
Relax (#relevant = 3284)					
E-C-T-03-N3	60	.2027	.2571	.2393	.2312
%mono	83	64	58	60	69
E-C-D-04-N3	65	.2145	.3000	.3071	.2422
%mono	83	60	57	68	64
NTCIR-3: D	*52*	*.1587*	*.2643*	*.2179*	*.1846*

Table 10: Stage-2 N3 English-Chinese CLIR Results for 42 Queries

3.3.2 English-Chinese CLIR

Results of our current N3 CLIR runs are tabulated in Table 10. The methods used differ substantially from the old NTCIR-3 runs. For translation, NTCIR-3 runs employed Huajian MT concatenated with dictionary lookup. The current runs use Systran MT concatenated with web-based translation of entity terms. The old runs also employed pre-translation expansion which was not done during NTCIR-6. There were no 'title' runs in the old submissions. Comparing the 'description' runs, it can be seen that there is vast improvements in MAP from .1150 to .1753 (Rigid), and from .1587 to .2145 (Relax evaluation). Both the 'title' and 'description' runs are the best reported from all sites. We believe the entity translation contributes significantly to these results.

4 Discussion and Conclusion

Overall, we believe results of all our submissions could be better by a few percent if not for linking to an outdated module that was used for testing. Stage-1 results are unofficial because of late submission. They were above median for both monolingual Chinese IR and English-Chinese CLIR. Stage-2 tasks attempt to compare different methods for repeat retrievals using the same environments. All Stage-2 monolingual Chinese IR results are above median, and all E-C CLIR submitted results are the best among all sites. Comparison of N5 results with those of previous year show that setting the Zipf high threshold too low screens out too many high frequency terms (our out-dated module) and affect MAP values adversely. The use of 'avtf' (average term frequency) weighting during initial retrieval for N3 and N4 appears to help counteract this adverse effect, and return MAP results close to or exceeding those of NTCIR-3 and -4 by comparison. For CLIR, current translation procedure of using Systran MT with our web-based entity translation appears to be much superior to our older approach of using other MT systems with dictionary translation.

References

[1] K. Kishida. K.-H. Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, & S.-H .Myaeng. Overview of CLIR task at the sixth NTCIR Workshop. In Proceedings of the Sixth NTCIR Workshop 2007.

[2] K.L. Kwok. A network approach to probabilistic information retrieval. ACM Transactions on Office Information System, 13:324-353, 1995.

[3] K.L. Kwok. Improving English and Chinese Ad-Hoc Retrieval: A Tipster Text Phase 3 Project Report. Information Retrieval, 3:313-338, 2000.

[4] K.L. Kwok, P. Deng, H.L. Sun, W. Xu, N. Dinstl, P. Peng, & J. Doyon. CHINET – a Chinese name finder for document triage. Proc. of 2005 International Conference on Intelligence Analysis. 2005. (http://analysis.mitre.org/proceedings_agenda.htm#papers)

[5] K.L. Kwok, N. Dinstl & S. Choi. NTICR-4 Chinese, English Korean Cross Language Retrieval Experiments using PIRCS. In Proc. of the Fourth NTCIR Workshop. pp.186-92, 2004. (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/index.html)

[6] K.L. Kwok, S. Choi, N. Dinstl & P. Deng. NTICR-5 Chinese, English, Korean Cross Language Retrieval Experiments using PIRCS. In Proceedings of the Fifth NTCIR Workshop. pp.88-95, 2005. (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/index.html)

[7] K. Kishida. K.-H. Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, S.-H .Myaeng & K. Eguchi. Overview of CLIR task at the Fourth NTCIR Workshop. In Proceedings of the Fourth NTCIR Workshop. pp.1-59, 2004.

[8] K.L. Kwok. NTICR-3 Chinese, Cross Language Retrieval Experiments using PIRCS. In Proc. of the Third NTCIR Workshop. pp.45-49, 2002. (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/index.html)

[9] K.L. Kwok & M. Chan. Improving Two-Stage Ad-Hoc Retrieval for Short Queries. In: Proceedings of ACM SIGIR Conf. pp.250-6, 1998.

[10] K.-H. Chen, H.-H. Chen, N. Kando, K. Kuriyama, S. Lee, S.-H .Myaeng, K.Kishida, K.Eguchi & H. Kim. Overview of CLIR task at the Third NTCIR Workshop. In Proceedings of the Third NTCIR Workshop. pp.1-38, 2002