Cross-Language Chinese Text Retrieval in NTCIR Workshop – Towards Cross-Language Multilingual Text Retrieval

Kuang-hua Chen[+] and Hsin-Hsi Chen[*]

[+]Department of Library and Information Science
National Taiwan University
1, Sec. 4, Roosevelt Road
Taipei 10617, Taiwan
khchen@ccms.ntu.edu.tw

[*]Department of Computer Science and Information Engineering
National Taiwan University
1, Sec. 4, Roosevelt Road
Taipei 10617, Taiwan
hh_chen@csie.ntu.edu.tw

Abstract

This article reports the results of Chinese Text Retrieval (CHTR) tasks in NTCIR Workshop 2 and the future plan of NTCIR workshop. CHTR tasks fall into two categories: Chinese-Chinese IR (CHIR) and English-Chinese IR (ECIR). The definitions, schedules, test collection (CIRB010), search results, evaluation, and initial analyses of search results of CHIR and ECIR are discussed in this article. The new plan of NTCIR towards multilingual Cross-Language Information Retrieval (CLIR) is also described.

1. Introduction

The evaluation of information retrieval has been an important issue since the first IR system built in 1954. However, how to evaluate has also been a controversial issue. A well-constructed test collection has been thought as a good mean for evaluation. Test collections often established on purpose of individual IR evaluation project like Cranfield II, ADI, MEDLARS, TIME, CACM, CISI, NPL, INSPEC, ISILT, UKCIS, UKAEA, LISA, and so on [1, 2, 3, 4, 5]. These test collections have different scheme according to the different purposes and the variant goals. However, they share some characteristics: small scale and strong homogeneity. For instance, the test collection of Cranfield II only has 1400 documents with the similar document length and the same subject of Aeronautics [6]. Because of the great gap between the IR test collection and the real IR environment, the evaluation based on such kind of test collections had been ever doubted of the validity [7]. Although many test collections had been constructed afterwards such as OHSUMED [8], Cystic Fibrosis [9], and BMIR-J2 [10], they still have the same shortcomings mentioned above.

To establish a test collection may cost significant time and manpower, especially at the phase of relevance judgment. Take Cranfield II for example, the relevance judgment has to be performed hundred thousands of times to consider the relationship between each query and document. Therefore, few researches focus on the development of test collection in early periods. The test collection in Cranfield II experiments was the most famous and used widely [11].

In 1992, the first Text REtrieval Conference (TREC) [12] built up a large-scale test collection with the different domain of document and query set. It is deemed to initiate a new landmark in IR evaluation research. Many test collections have been constructed based on the scheme and the model of TREC especially in the design of topics. For example, Institute of Information Scientific and Technique (INIST) in France has initiated AMARYLLIS project and built a TREC-like test collection [13]; CLEF (Cross-Language Evaluation Forum), a joint-force of Europe, has constructed multilingual test collections [14]; the NTCIR (NACSIS Test Collection for IR Systems) in Japan [15], CIRB (Chinese Information Retrieval Benchmark) in Taiwan [16] and HANTEC in Korea [17] have also developed the TREC-like test collection. Now, NTCIR, CIRB, and HANTEC have been a joint-force for multilingual CLIR.

NTCIR Workshop 1 is the first evaluation workshop designed to enhance research in Japanese text retrieval [18]. The authors have discussed a kind of joint efforts in evaluating Eastern-Asia text retrieval with Dr. Kando for a long time. NTCIR Workshop 2 is the result of the attempt and is the first evaluation workshop designed to enhance research in Japanese and Chinese text retrieval. The CHTR tasks fall into two categories: Chinese queries against Chinese documents (CHIR, a monolingual IR task) and English queries against Chinese documents (ECIR, a cross-language IR task). Both CHIR and ECIR are ad hoc IR tasks, i.e., the

document set is fixed for various topics.

The test collection used in CHTR tasks of NTCIR Workshop 2 is called CIRB010. It contains 132,173 documents. These documents are all news stories downloaded from web sites of Chinatimes [19], Chinatimes Commercial [20], Chinatimes Express [21], Central Daily News [22], and China Daily News [23] during the period of May 1998 to May 1999. We are authorized to use these news stories for evaluation in NTCIR. The advantages of using news articles are manifold. They are usually quite novel, quickly-updated and with multiple subjects in contents.

Each participant could conduct ECIR task, CHIR task or both tasks. Sixteen groups from seven countries or areas had enrolled CHTR tasks. Among them, 14 groups enrolled CHIR task and 13 groups enrolled ECIR task. However, not all enrolled groups submit search results. Table 1 shows the distribution of groups enrolling CHIR and ECIR tasks and groups submitting search results at final. The search results of 115 runs were submitted from 11 groups. 98 runs from 10 groups are for CHIR task; 17 runs from 7 groups are for ECIR task. Table 2 shows the detailed statistics.

Table 1. Distribution of Participants

	Enrolled		Submitted	
	CHIR	ECIR	CHIR	ECIR
Canada	1	1	0	0
China	2	1	2	1
Hong Kong	2	1	1	0
Japan	3	2	3	1
Taiwan	2	2	2	2
UK	1	1	0	0
USA	3	5	2	3

Table 2. Participants of CHTR Tasks

	CHIR	ECIR	Total
# of groups enrolled	14	13	16
# of groups submitted	10	7	11
# of submitted runs	98	17	115

The rest of this report will focus on the test collection (CIRB010), the CHIR task, ECIR task, and future plan. Section 2 will introduce the task of CHTR in NTCIR workshop 2. Section 3 will describe the test collection used in the CHTR tasks. Section 4 will give a picture of the evaluation mechanism. Section 5 will analyze the search results in a broad view. Section 6 will talk about the new plan of NTCIR towards CLIR. Section 7 will give a conclusion.

2. CHIR Task and ECIR Task

Two kinds of IR tasks have been arranged for NTCIR Workshop 2 Chinese Text Retrieval. The first is Chinese IR Task (a monolingual IR task) and the second is English-Chinese IR Task (a cross-language IR task). Both tasks are ad-hoc-based tasks, that is to say, the document set is fixed against the different topics.

2.1 Schedule

2000-07-15:	Application due.
2000-08-31:	Test data are distributed to participants.
2000-09-30:	Results and system description forms submission.
2001-01-10:	Results of Relevance Assessments will be distributed to the participants.
2001-02-12:	Papers for the working-note proceedings submission.
2001-03-07/ 2001-03-09:	Workshop meeting at NII, Tokyo, Japan.
2001-03-16:	Camera-ready copies for the proceedings.

2.2 Task Type

- Chinese IR Task (“CHIR”)
  The Chinese IR Task is to assess the capability of participating systems in retrieving Chinese documents using Chinese queries. Chinese texts, which are composed of characters without explicit word boundary, make the retrieval task more challengeable than English ones. The participating systems can employ any approaches. Either word-based or character-based systems are acceptable. The organizer will not provide any segmentation tools and Chinese dictionaries.
- English-Chinese IR Task (“ECIR”)
  The English-Chinese IR Task is to assess the capability of participating systems in retrieving Chinese documents using English queries. The organizer will not provide any segmentation tools and English-Chinese dictionaries.

2.3 Query Type

We distinguish each run according to the length of query. Three different types of run are defined as follows.

- Long query (“LO”): Any query uses <narrative> field..
- Short query (“SO”): Any query uses no <narrative> field.
- Very short query (“VS”): Any query uses neither <narrative> nor <question> fields.
- Title query (“TI”): Any query uses the <title> field only.

The participating group could use any type of query to carry out the IR tasks.

3. The Test Collection: CIRB010

3.1 Document Set

In order to facilitate the process of identification and analysis of the contents, documents are supposed to be consistent in their format. Therefore, we edit the html documents downloaded from web and delete the noises. In addition, we add tags to mark the designated fields, which are document id-number, news reporting date, title, paragraphs, etc. Consequently, every document has the same format and tags. The documents are encoded in BIG5 with XML-style tags. We add tags to documents to mark their specifications and sections. The meaning of each tag is described below:

- <doc> </doc>: Denote the beginning and the ending of a document.

- <id> </id>: Denote the document identifier, which is composed of the source, the subject category, and the serial number of document.
- <date> </date>: Denote the date of the news using ISO8601 format. It is presented in the format of "year (in 4 digits)-month (in 2 digits)-day (in 2 digits)".
- <title> </title>: Denote the title of news.
- <text> </text>: Denote the text of news.
- : Denote the paragraphs of news.

Table 3 shows the statistics of CIRB document set. CIRB010 contains 132,173 documents with the size of 200MB. The subjects of documents are various, such as politics, finance, social, life, sports, entertainment, international issue, and information technology and so on.

Table 3. Document Set

News Agency	# of Document	Percentage
Chinatimes	38,163	28.8%
Chinatimes Commercial	25,812	19.5%
Chinatimes Express	5,747	4.4%
Central Daily News	27,770	21.0%
China Daily News	34,728	26.3%
Total	132,173	(200MB)

3.2 Topic

Three main procedures constructing topics of CIRB010 are shown as follows:

(1) Collecting information request

In order to increase the similarity between our benchmark and real environment, we build the topics using real users' information requests. We collected 405 requests through questionnaire on web. There are both closed and open-ended questions about the types and subject of requests, narratives of requests, and other related information. The basic assumption of the method is that users may state their specific information request distinctly and exhaustively.

(2) Selecting information request

The responses of questionnaires gained from Internet was not entirely so qualitative, complete and exhaustive. In addition, the type and subject of the request provided by user is not necessarily suitable for evaluation purpose in our IR benchmark. Therefore, we pick out 50 best requests from 405 collected requests according to some criteria.

(3) Constructing Topics

The main task of this phase is to establish the topics in accordance with the 50 final requests. We use four fields: title, question, narrative, and concepts to represent topics in accordance with the TREC's convention. The "title" field has the widest coverage in its content with comparison to the other three fields. The coverage of "question" field is the second to "title" field. The "narrative" field is the most specific because of its detailed description. The keywords in "concepts" field touch on the contents of above three fields. The average number of words in a topic is 169.

4. Evaluation

This is our first attempt to organize Chinese IR evaluation workshop. We follow the method used in TREC and NTCIR Workshop 1. The TREC's evaluation program is used to score the research results. It provides the interpolated recall and precision at 11 points, average precision (non-interpolated) over al relevant documents and precision at 5, 10, 15, 20, 30, 100, 200, 500, and 1000 documents. Each participating group has to submit its search results in the designated format. The result file is a list of tuples in the following form:

qid iter docid rank sim runid

giving CIRB010 document "docid" (a string extracted from the <id> </id> field, e.g. <id> cts_cec_1999111514 </id>) retrieved by query "qid" (an integer extracted from the last 3 digits in <number> </number> field of topic, e.g., <number> CIRB010TopicEN002 </number>, the "qid" is 002) with similarity *sim* (a float). The result file is assumed to be sorted numerically by "qid". "Sim" is assumed to be higher for the documents to be retrieved first. The "iter" and "rank" could be regarded as the dummy filed in tuples. In addition, each field in tuples is separated by inserting 'TAB' (\x0A, \t) character.

A list of relevance between each topic and documents in a benchmark is needed to facilitate the comparison and evaluation of IR system effectiveness. This is the so-called "relevance judgment." While performing relevance judgments, every judge should read and understand the meaning of the topic carefully and assign each of them to the most appropriate category (mentioned below) from their viewpoint mainly according to "question" field of topic. In order to keep judges' criterion consistent, the judges must complete the judgments for a topic in a period of time. Each topic is judged by 3 judges. In total, 23 judges spend 799 hours in relevance judgment.

The "subject relevance" concept is adopted in relevance judgment. That is to say, we pay more attention to the concrete meaning, which can be perceived from the text. Based on this concept, the judges should make an objective link between document and topic. This will increase the consistency and reliability of judgments performed by different judges. As for measurement granularity, it is supposed that some distinct definitions of relevance degree should be identified to keep judgment objective. 4 categories of relevance are identified: "Very Relevant", "Relevant", "Partially relevant", and "Irrelevant." Each kind of relevance is assigned a relevance score. "Very relevant" is 3, "Relevant" is 2, "Partially relevant" is 1, and "Irrelevant" is 0.

Since one unified relevance score have to be produced for final relevance judgment using TREC's scoring program, we combine judgment results of three judges, and then decide how to interpret the meaning of the score and how can it be applied to IR evaluation. Based on the following philosophy, we devise a method to integrate 3 relevance scores to form one relevance score.

- Each judge has equal contribution to final relevance score.

● Each judgment is independent.

The following formula is used to combine 3 judges' relevance score,

$$R = \frac{(X_A + X_B + X_C)/3}{3}$$

where X means the relevance category assigned by each judge, and A, B, C represent the three different judges. The value of R will be between 0 and 1.

As mentioned above, TREC scoring program is used to calculate the recall and precision. Since it uses binary relevance judgment, we have to decide the threshold. Two thresholds are decided: one is 0.6667, the other is 0.3333. The so-called rigid relevance means the final relevance score should between 0.6667 and 1. That is to say, it is equivalent that each person assigns "relevant (2)" to the document.

[(2+2+2)/3/3=0.6667]

The so-called relaxed relevance means the final relevance score should between 0.3333 and 1. That is to say, it is equivalent that each person assigns "partially relevant (1)" to the document.

[(1+1+1)/3/3=0.3333]

5. Search Results

We will report the search results in a broad view and analyze some of runs using different query types in this section. The different techniques which each participating group took could be referred to each paper in workshop proceedings [24].

5.1 CHIR Task

The search results of CHIR task are submitted from 10 participating groups listed in Table 4. Some groups are "Full Participation" and some are "Anonymous Participation". The number of submitted runs by the query types is shown in Table 5. The query types have been mentioned in Section 2.

The recall/precision graphs of top runs of CHIR task are showed in Figure 1 (relaxed relevance) and Figure 2 (rigid relevance). The techniques used in these top runs are showed in Appendix I. It is easy to find out that all runs use query expansion techniques except Brkly-CHIR-LO-01. CRL group performs well in all query types, since it uses automatic feedback to carry out query expansion. We also find that stop-word list seems a good resource for Chinese information retrieval. Brkly group applies logistic regression technique to tune the various parameters. The unique technique shows good performance in CHIR task.

Basically, most of the participating groups use tf/idf approach in a little different form. This shows that the long-history tf/idf approach still play an important role in information retrieval.

Both CRL group and PIRCS group adopt probabilistic model and they are the leading groups in CHIR task. This implies that the probabilistic model shows good performance in this task at least.

In general, the performance of "Very Short Query" is the best; the performance of "Short Query" is better than that of "Long Query". The performance of "Title Query" is worst. It seems that the long query conveys many noises. On the contrary, the title query conveys little information.

Table 4. List of CHIR Participating Groups

1	Communications Research Laboratory
2	University of California at Berkeley
3	Queens College, CUNY
4	Chinese research group of Lab Furugori, the University of Electro-Communications
5	Department of Computing, Hong Kong Polytechnic University
6	Umemura Lab. Department of Information and Computer Sciences, Toyohashi University of Technology
7	NTHU NLP Lab and Knowledge Express Technology Inc
8	Institute of Software, Chinese Academy of Sciences
9	Trans-EZ Information Technology Inc.
10	Fujitsu R&D Center

Table 5. Number of Submitted Runs of CHIR

# of Groups	10
# of Total RUN	98
# of "LO" RUN	30
# of "SO" RUN	12
# of "VS" RUN	27
# of "TI" RUN	29

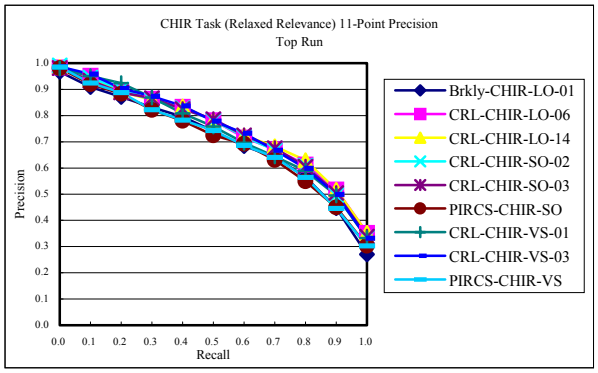

Figure 1. CHIR Task (Relaxed Relevance)

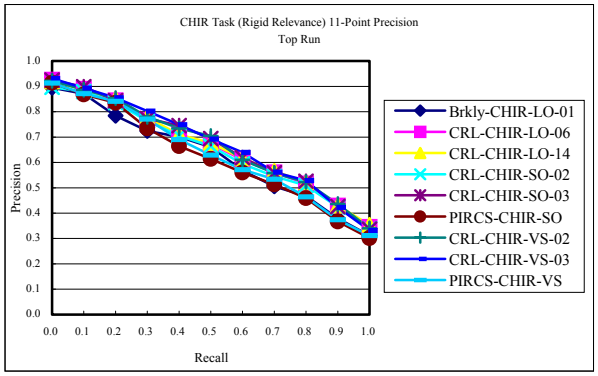

Figure 2. CHIR Task (Rigid Relevance)

Observing the search results, we conclude that the short query is appropriate for CHIR task. It seems that the name of "Very Short Query" will mislead us to draw a direct conclusion that query should be short. In fact, the "Very Short Query" means that the participants could use the concepts identified in the topic. In the case of CIRB010, the <concepts> field contains many significant keywords. As a result, the runs applying "Very Short Query" perform well.

Since the leading groups show good performance in all runs, we will not explain the details of each query type. The details could be referred to the workshop proceedings.

5.2 ECIR Task

The search results of ECIR task are submitted from 7 participating groups listed in Table 6. All groups are "Full Participation". The number of submitted runs by the query types is shown in Table 7.

The recall/precision graphs of top runs of ECIR task are showed in Figure 3 (relaxed relevance) and Figure 4 (rigid relevance). Appendix II shows the techniques which leading participating groups used in ECIR task. Observing Figure 3, we find that PIRCS group outperforms other groups using relaxed relevance metric. Further investigating Appendix II, we have an idea that PIRCS uses MT software to carry out translation commission. On the contrary, most groups use dictionaries with select-all, select-top-1, select-top-n, or select-all approaches. Among select-X approach, select-all is better than select-top-3; select-top-3 is better than select-top-2; select-top-2 is better than select-top-1. We could not conclude directly that select-all is the best among all select-X approaches, since some groups also apply corpus-based approach at the same time. However, no enough information shows how participating groups utilize corpus. Did they calculate mutual information? Did they calculate the bilingual mutual information? The detailed information should be referred to the papers in workshop proceedings.

Observing the index unit, we find that word-based approaches are much better than other approaches in ECIR task. In addition, PIRCS group combines word-based and character-based approaches to construct index file.

Since only PIRCS group submits runs of all query types in ECIR task, we will compare the performances of each query type based on the search results of PIRCS. The "Title query" is the worst among all query types. The difference among "Long Query", "Short Query", and "Very Short Query" is little. However, the "Short Query" is better than others. As mentioned before, the "Very Short Query" in CIRB010 conveys many important keywords, so the performance is good. This phenomenon is different from the observation pointed out in Japanese IR task of NTCIR workshop 1 [18].

Since the leading groups show good performance in all runs, we will not explain the details of each query type. Again, the interested readers have to refer to the corresponding papers in the conference proceedings for technical details.

Table 6. List of ECIR Participating Groups

1	University of California at Berkeley
2	University of Maryland
3	Queens College, CUNY
4	Umemura Lab. Department of Information and Computer Sciences, Toyohashi University of Technology
5	NTHU NLP Lab and Knowledge Express Technology Inc
6	Institute of Software, Chinese Academy of Sciences
7	Trans-EZ Information Technology Inc.

Table 7. Number of Submitted Runs of ECIR

# of Groups	7
# of Total RUN	17
# of "LO" RUN	8
# of "SO" RUN	2
# of "VS" RUN	6
# of "TI" RUN	1

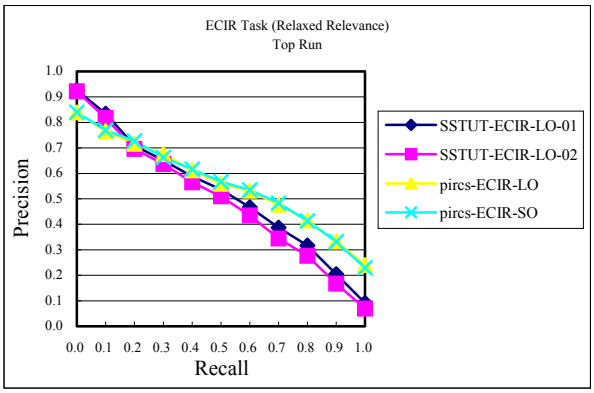

Figure 3. ECIR Task (Relaxed Relevance)

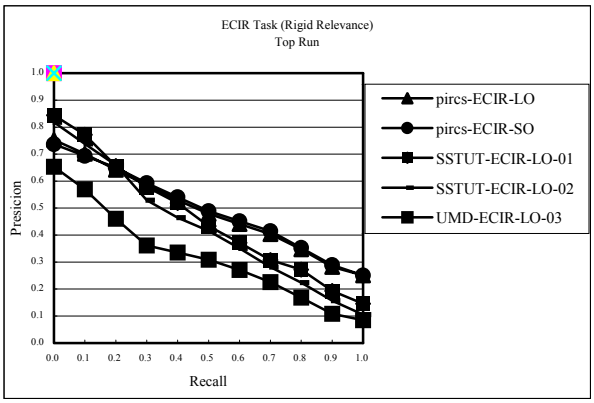

Figure 4. ECIR Task (Rigid Relevance)

6. Multilingual Cross-Language IR

NTCIR has become a joint-force since the second workshop. The organizers of NTCIR would like to extend the coverage of languages and propose a multilingual cross-language information retrieval task (at least 4 languages, Chinese, English, Japanese, and Korean). The third workshop of

NTCIR will provide the following tasks: CLIR Task, Patent Retrieval Task, Question Answering Task, Automatic Text Summarization Task, and Web Retrieval Task. The authors will be in charge of the cross-language information retrieval task (CLIR).

The CLIR task is a joint-effort of Japan, Korean, and Taiwan. The executive committee consists of 9 persons: Dr. Hsin-Hsi Chen (Co-chair, Taiwan), Dr. Kuang-hua Chen (Co-chair, Taiwan), Dr. Koji Eguchi (Japan), Dr. Noriko Kando (Japan), Dr. Hyeon Kim (Korea), Dr. Kazuaki Kishida (Japan), Dr. Kazuko Kuriyama (Japan), Dr. Suk-Hoon Lee (Korea), and Dr. Sung Hyon Myaeng (Korea). In order to discuss the details of CLIR task in NTCIR workshop 3, the members of executive committee met in Tokyo to decide the potential tracks, document set, topic set, criteria of relevance judgment, policy, schedule, etc. The following will describe the details of CLIR task.

Three tracks are identified: 1) Multilingual Cross-Language Information Retrieval (MLIR); 2) Bilingual Cross-Language Information Retrieval (BLIR); 3) Single Language Information Retrieval (SLIR). The participants could make their own mind to join any one, any two, or all tracks. The document set consists of Chinese, English, Japanese, and Korean news articles. All but Korean documents published between 1998 and 1999. Table 8 shows the document set used in CLIR task. The tag set is shown in Table 9.

Table 8. Document Set

Japan	Mainichi Newspaper (1998-1999): Japanese	230,000
	Mainichi Daily News (1998-1999): English	14,000
Korea	Korea Economic Daily (1994): Korean	66,146
Taiwan	CIRB010 (1998-1999): Chinese	132,173
	United Daily News (1998-1999): Chinese [25]	249,508
	Taiwan News and Chinatimes English News (1998-1999): English	10,204

Table 9. Tags for Document Set

Mandatory tags		
<DOC>	</DOC>	The tag for each document
<DOCNO>	</DOCNO>	Document identifier
<LANG>	</LANG>	Language code: CH, EN, JA, KR
<HEADLINE>	</HEADLINE>	Title of this news article
<DATE>	</DATE>	Issue date
<TEXT>	</TEXT>	Text of news article
Optional tags		
<P>	</P>	Paragraph marker
<SECTION>	</SECTION>	Section identifier in original newspapers
<AE>	</AE>	Contain figures or not
<WORDS>	</WORDS>	Number of words in 2 bytes (for Mainichi Newspaper)

The topics are contributed by each country. Topics are the information need of users, which are represented in different level of details. A number of tags are used to denote the information need in topics. Table 10 shows the tags for topic.

Table 10. Tags for Topic Set

<TOPIC>	</TOPIC>	The tag for each topic
<NUM>	</NUM>	Topic identifier
<SLANG>	</SLANG>	Source language: CH, EN, JA, KR
<TLANG>	</TLANG>	Target language: CH, EN, JA, KR
<TITLE>	</TITLE>	The concise representation of information request, which is composed of noun or noun phrase.
<DESC>	</DESC>	A short description of the topic. The brief description of information need, which is composed of one or two sentences.
<NARR>	</NARR>	The <NARR> has to be detailed, like the further interpretation to the request, the list of relevant or irrelevant items, the specific requirements or limitations, etc.
<CONC>	</CONC>	The keywords relevant to whole topic.

The different run types based on the combination of variant fields of topic are allowed in CLIR. For example, participants could submit T run, D run, N run, C runs, TD run, TN run, TC run, DN run, DC run, NC run, TDN run, TDC run, TNC run, DNC run, and TDNC run. However, the D run is a must-do run, i.e., each participant has to submit a D run. In addition, each participant at most submits 3 runs for each language pair. Here language pair means topic language and document language. For example, C-JE is a language pair, i.e., topic language is Chinese and document languages are Japanese and English. The submitted runs have to be assigned a unique identifier. The format of identifier is

GroupId-TopicLanguage-DocLanguage-RunType-dd,

where *GroupId* is a group identifier named by participating group itself; *TopicLanguage* is the language code (CH, EN, JA, or KR) for query language; *DocLanguage* is the language code (CH, EN, JA, or KR) for document language; The "dd" is two optional digits used to distinguish runs with the same run type but using different techniques. For example, a participating group, LIPS, submits 2 runs. The first is a D run for C-->CJ track and the second is a DN run for J-->C track. Therefore, the RunID for each run is LIPS-C-CJ-D and LIPS-J-C-DN. However, if this group uses different ranking techniques in LIPS-C-CJ-D, the RunID for each run has to be LIPS-C-CJ-D-01, LIPS-C-CJ-D-02, etc.

Relevance judgments will be done in four grades, Highly Relevant, Relevant, Partially Relevant, and Irrelevant. Evaluation will be done using trec_eval and new metrics for multigrade relevance.

The detailed schedule is shown as follows.

- 2001-09-30 Application Due
- 2001-10-01 Deliver Dry Run data
- 2001-10-19 Submit search result of Dry Run
- 2001-11-30 Deliver evaluation result of Dry Run
- 2001-12-22 Deliver Formal Run data
- 2002-01-25 Submit search result of Formal Run
- 2002-07-01 Deliver the evaluation results
- 2002-08-20 Paper Due
- 2002-10-08 NTCIR Workshop 3

7. Conclusions

The NTCIR Workshop 2 is the first international joint effort in providing an evaluation mechanism for Japanese and Chinese Text Retrieval. We hope this mechanism could encourage the IR researches in Eastern Asia, promote the concept of IR evaluation, provide an opportunity to share the research ideas and results, investigate the useful techniques for IR researches, and enhance the effectiveness of IR.

Through the initial analyses on the submitted runs, some findings are shown as follows.

- Most participating groups apply inverted file approach.
- Many participating groups adopt tf/idf-based approaches.
- "Short Query" and "Very Short Query" performs well.
- Query expansion is good for system performance.
- In general, the probabilistic model performs well.
- For CHIR task, stop-word list is a good resource for enhancing system performance.
- For ECIR task, select-all approach seems to be better than other select-X approaches, if it uses no further techniques.
- For ECIR task, MT approach is much better than dictionary-based approach.
- For ECIR task, word-based indexing approach is better.

We would like to say again that these findings are drawn from the submitted runs using the "CIRB010" test collection. They cannot directly apply to other test collection, since each test collection has its own characteristics and each language also has its own characteristics.

As mentioned previously, the keywords in concepts filed of topic provide the crucial information and make the performance higher than other IR evaluation forum. We would like to explain the procedure for keywords preparation. We had executed a pre-test for CIRB010 test collection. As a result, the positive documents and negative documents for each topic have been constructed. We then analyze these documents and extract the good keywords for each topic. According to our analysis and Brkly's experiment, using concepts field will produce the best performance with comparison to the other fields. Therefore, we are considering the role of concepts field in the future.

Acknowledgments

We would like to thank Chinatimes, Chinatimes Commercial, Chinatimes Express, Central Daily News, China Daily News, and United Daily News for their kindly providing test materials. We are grateful to all pioneers in the area of IR evaluation for their efforts in paving a smooth way for followers. We would like to thank the participants for their contributions and the relevance judges for their hard working. Special thanks are due to Dr. Noriko Kando for her helps.

References

[1] Donna K. Harman, "Panel: Building and Using Test Collections," in *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, August 18-22, 1996, 337.

[2] Karan Sparck Jones and C. J. van Rijsbergen, "Information Retrieval Test Collections," *Journal of Documentation* 32 (1976): 63-73.

[3] Gerard Salton, "A New Comparison between Conventional Indexing (MEDLARS) and Automatic Text Processing (SMART)," *Journal of the American Society for Information Science* 23, no. 1 (1972): 75-84.

[4] Edward A. Fox, "Characteristics of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts," (Technical Report TR 83-561, Cornell University: Computing Science Department, 1983), <http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR83-561> (30 Aug. 2001).

[5] William M. Shaw, Robert Burgin, and Patrick Howell, "Performance Standards and Evaluations in IR Test Collections: Vector-Space and Other Retrieval Models," *Information Processing and Management* 33, no. 1 (1997): 15-36. <http://ruby.ils.unc.edu/~howep/perform/hypergeom.html> (30 Aug. 2001).

[6] Cyril W. Cleverdon, "The Cranfield Tests on Index Language Devices," *Aslib Proceedings* 19, no. 6 (1967): 173-194.

[7] David Bawden, *User-oriented Evaluation of Information Systems and Services*. (Aldershot: Gower, 1990), 87-88.

[8] William Hersh, "OHSUMED: An Interactive Evaluation and New Large Test Collection for Research," in *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, July 3-6, 1994, 192-201.

[9] William M. Shaw, Judith B. Wood, Robert E. Wood, and Helen R. Tibbo, "The Cystic Fibrosis Database: Content and Research Opportunities," *Library and Information Science Research* 13 (1991): 347-366.

[10] Tsuyoshi Kitani, et al., eds., "BMIR-J2: A Test Collection for Evaluation of Japanese Information Retrieval Systems," in *Proceedings of IPSJ SIG Notes*, DBS-114-3, 1998, 15-22.

[11] Karan Sparck Jones, "The Cranfield Tests," in *Information Retrieval Experiment, ed. Karan Sparck Jones* (London; Boston: Butterworths, 1981), 276.

[12] TREC (Text REtrieval Conference) Homepage, <http://trec.nist.gov/> (30 Aug. 2001).

[13] AMARYLLLIS Homepage, <http://www.inist.fr/accueil/profran.htm> (30 Aug. 2001).

[14] CLEF (Cross-Language Evaluation Forum) Homepage, <http://www.iei.pi.cnr.it/DELOS/CLEF/> (30 Aug. 2001).

[15] NTCIR Project (NACSIS Test Collection for IR Systems) Homepage, <http://research.nii.ac.jp/ntcir/> (30 Aug. 2001).

[16] CIRB (Chinese Information Retrieval Benchmark) Homepage, <http://lips.lis.ntu.edu.tw/cirb/index.htm> (30 Aug. 2001).
[17] HANTEC Homepage, <http://hantec.kordic.re.kr/> (30 Aug. 2001).
[18] N. Kando, et al. "Overview of IR Tasks at the First NTCIR Workshop," in *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 1-46, Tokyo, 1999.
[19] China Times Homepage, <http://news.chinatimes.com.tw/> (30 Aug. 2001).
[20] Chinatimes Commercial Homepage, <http://news.chinatimes.com.tw/> (30 Aug. 2001).
[21] Chinatimes Express Homepage, <http://news.chinatimes.com.tw/> (30 Aug. 2001).
[22] Central Daily News Homepage, <http://www.cdn.com.tw/> (30 Aug. 2001).
[23] China Daily News Homepage, <http://www.cdns.com.tw/> (30 Aug. 2001).
[24] N. Kando, et al., eds. *Proceedings of the Second NTCIR Workshop on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, 2000.
[25] United Daily News Homepage, <http://udnnews.com/> (30 Aug. 2001).

Appendix

I. Techniques Leading Participants Used in CHIR Task

RunID	IndexUnit	IndexTech	IndexStru	QueryUnit	IRModel	Ranking	QueryExpan
Brkly-CHIR-LO-01	bi-character	stopword	Inverted file	bi-character	logistic regression	tf/idf/dl/ql/cl/cf	NO
CRL-CHIR-LO-06	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-LO-14	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-SO-02	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-SO-03	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
PIRCS-CHIR-SO	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term
PIRCS-CHIR-VS	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ	tf/ ictf	top40doc+ 100term
CRL-CHIR-VS-02	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback
CRL-CHIR-VS-03	mostly bi-character + character	using all characters and bi-characters	inverted file	mostly bi-character + character	probabilistic model (okapi)	okapi weight	automatic feedback

II. Techniques Leading Participants Used in ECIR Task

RunID	IndexUnit	IndexTech	IndexStru	QueryUnit	IRModel	Ranking	QueryExpan	TransTech
Brkly-ECIR-LO-01	word	stopword+ dictionary	Inverted file	word	logistic regression	tf/idf/dl/ ql/cl/cf	NO	Dictionary-based, select two
IOS-ECIR-*	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
NTHU-ECIR-LO-01	bi-word	morphology	invertedfile	word	vector space module	tf/idf	No	dictionary-based, corpus-based
PIRCS-ECIR-*	word+char	dictionary, Zipf-thrhld	Inverted file, network	word+char	probabilistic model + spread-activ + retrieval combination	tf/ ictf	top40doc+ 100term	bi-word list + MT software
SSTUT-ECIR-LO-01	all n-grams	as is	suffix array	word	probabilistic model	tf, idf, burstiness	No	dictionary-based handmade
SSTUT-ECIR-LO-02	all n-grams	as is	suffix array	word	probabilistic model with dynamic programming	tf, idf, burstiness	No	dictionary-based handmade
Trans-ECIR-SO	bi-word	No	inverted file	word	vector space model	tf	no	Dictionary-based and corpus-based,select-top-1
UMD-ECIR-LO-01	overlapping character bigram	hexicification of Chinese characters	inverted file	within word overlapping character bigram	probabilistic model	tf/idf	no	dictionary-based, select-all
UMD-ECIR-LO-02	overlapping character bigram	Chinese character hexifying	inverted file	within word overlapping character bigram	probabilistic model	tf/idf	no	dictionary-based, select-top-3
UMD-ECIR-LO-03	word	Chinese character hexifying	inverted file	word	probabilistic model	tf/idf	no	dictionary-based, select-top-3