Evaluating Chinese Text Retrieval with Multilingual Queries

Kuang-hua Chen

Department of Library and Information Science, National Taiwan University, TAIWAN

Kuang-hua Chen is an Associate Professor in the Department of Library and Information Science at National Taiwan University. He initiated a joint-force in evaluation of cross-language information retrieval with colleagues from Japan and Korea. His major research topics are natural language processing, information retrieval, digital libraries, and database systems.

Chen, K.-H. (2002). **Evaluating Chinese Text Retrieval with Multilingual Queries.** *Knowledge Organization, 29(3/4).* 156-170. 28 refs.

ABSTRACT: This paper reports the design of a Chinese test collection with multilingual queries and the application of this test collection to evaluate information retrieval systems. The effective indexing units, IR models, translation techniques, and query expansion for Chinese text retrieval are identified. The collaboration of East Asian countries for construction of test collections for cross-language multilingual text retrieval is also discussed in this paper. As well, a tool is designed to help assessors judge relevance and gather the events of relevance judgment. The log file created by this tool will be used to analyze the behaviors of assessors in the future.

1. Introduction

Information is critical for human beings to live from pre-historic times to the present time. Along with the passage of time, the form of information, the complexity of information, the ways in which information is transferred, and the means of information access have changed sharply. Although, many different technologies, methods, and means for information access have been devised, we have a consensus at this time that the World Wide Web (WWW) has greatly changed the ways in which laymen access information. It is very easy for us to find a lot of examples of many information services based on the WWW for users all over the world. How to provide fast, prompt, one-stop, up-to-date, and complete information services are the goals of service providers. The search engine is the basis for all services from the viewpoint of information access. Due to the global nature of the WWW, multi-lingual and multi-cultural issues are very important for search engines. The need to evaluate the performance of search engines from the multilingual and multi-cultural viewpoint is critical.

Basically, the search engine is an application of the information retrieval (IR) system on the Internet. The IR has been one of the core research issues in information services for a long time since the idea started in 1945 as described in an article by Vannevar Bush. (Bush, 1945) In the era of the information explosion, it is much more difficult to search information in efficient and precise ways. Although IR is an old concept, IR systems still play an increasingly significant role nowadays. Many researchers have devoted themselves to the research of IR and manage to design IR systems with high precision and high recall. In order to evaluate the precision and the recall of IR systems, IR evaluation has become one of the major research fields in the IR community since the 1950s.

IR evaluation is essential while designing and developing an IR system. Through the procedure of evaluation, researchers may examine the relative effectiveness of different IR techniques and improve their retrieval systems according to the results of evaluation. In fact, evaluation has great impacts upon the direction of the development of IR systems. Traditionally, evaluations are usually carried out in a so-called

normalized environment using test collections. Some evaluation rules, procedures, and criteria are designed to measure retrieval effectiveness. Cleverdon's experiment in 1966, Cranfield II, was considered to be the first one initiating such an evaluation model. (Cleverdon, 1967) The Cranfield II project used document set, query set, and relevance judgments and established effectiveness measurements to evaluate several different indexing models. Finally, the methodology adopted in Cranfield II has been regarded as a paradigm for IR evaluation. However, because the scales of test collections in the early days were usually small, a big gap between a normalized environment and a real environment exists. The effectiveness of using a test collection to evaluate IR systems is doubtful.

In 1992, the Defense Advanced Research Projects Agency (DARPA) and the National Institute of Standards and Technology (NIST) initiated the Text REtrieval Conference, called TREC. They established a large-scale test collection, provided various evaluation tasks, formalized test procedures, devised measurement criteria, and then built a normalized evaluation environment. The conference also sponsored a forum for participants to discuss numerous IR issues and share experiences. TREC has made a great contribution to research in IR and evaluation of IR systems.

IR research has been growing rapidly, and most people realize the urgency to build a unified evaluation environment. At present, in addition to TREC, some evaluation mechanisms are being initiated and some non-English test collections are being established. For example, NTCIR (Kageura, 1997) and IREX (IREX, 1999) have built Japanese test collections, and AMARYLLIS has built a French test collection. (Smeaton & Harman, 1997, p.173) A Chinese IR test collection in TREC (Ellen & Harman, 1996) has been constructed, but the volume is not large enough and the genres of documents are not varied enough. As a result, this author and his colleagues have decided to construct a Chinese test collection and plan to use this collection for evaluation of Chinese text retrieval systems.

This paper will report the construction of a Chinese test collection with multilingual queries for system performance evaluation. We also exepect to set up a feasible methodology and normalized procedures for construction of the test collection.

The rest of this paper is structured as follows. Section 2 will describe the previous works on evaluation of information retrieval and test collections. Section 3 will discuss the design of a Chinese test collection, in detail. Section 4 will depict the application of the test collection in the CHinese Text Retrieval (CHTR) task of the NTCIR Workshop 2 and discuss the run types, effective techniques, IR models, and search results. Section 5 will explain the necessity of the construction of a cross-language test collection with multilingual text and describe the plan for the NTCIR Workshop 3 using a real multilingual test collection. Section 6 will illustrate the design of ReleAssessor, a tool for relevance judgment. Section 7 will give short conclusions and suggest future research.

2. Previous Works

A typical information retrieval process could be outlined in the following steps:

(1) A user proposes a query that represents her/his information need.
(2) The retrieval system uses a matching algorithm to identify documents that are likely to satisfy the user's need.
(3) The user reads the retrieved documents to find answers to the original query.

Based on this simplified procedure of IR, an information retrieval system can be measured with respect to a test collection, which consists of a set of documents, a set of queries, and relevance judgment of each document with respect to each query. The IR system performs the retrieval according to the queries, and its performance is scored based on its ability to match the relevant documents.

A well-constructed test collection has been thought of as a good means for evaluation. In the early days, test collections were often established for individual IR evaluation projects such as Cranfield II, ADI, MEDLARS, TIME, CACM, CISI, NPL, INSPEC, ISILT, UKCIS, UKAEA, LISA, and so forth. (Harman, 1996; Sparck Jones & van Rijsbergen, 1976; Salton, 1972; Fox, 1983; Shaw et al., 1997) These test collections have different schemes based on 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. (Cleverdon, 1967) Because of the great gap between the IR test collection and the real IR environment, the validity of system evaluations based on such kinds of test collections had been doubted. (Bawden, 1990) Although many test collections have been constructed since, such as OHSUMED (Hersh, 1994), Cystic Fi-

brosis (Shaw et al., 1991), and BMIR-J2 (Kitani et al., 1998), they still have the same shortcomings mentioned above.

To establish a test collection may take significant time and cost great manpower, especially at the phase of relevance judgment. In Cranfield II for example, the relevance judgment has to be performed hundreds of thousands of times to consider the relationship between each query and document. Therefore, little research focuses on the development of test collection in the early periods. The test collection in Cranfield II experiments was the most famous and used widely. (Sparck Jones, 1981)

The design and composition of a test collection must be capable of handling different purposes, targets and measurements of evaluation. In addition, a "portable" test collection for different evaluation tasks is especially in demand. Sparck Jones and Van Rijsbergen (1976, p.67) are of the opinion that an ideal test collection must be of a certain scale. The contents, types, and sources of documents and queries of test collections should be heterogeneous to reflect the real environment. On the contrary, some subcollections have to be homogeneous for some special-purpose testing.

The first Text REtrieval Conference (TREC) built up a large-scale test collection with the different domains of documents and queries. It is deemed to have initiated a new landmark in IR evaluation research. The components and key characteristics of the test collection are briefly described below (Voorhees & Harman, 1998):

(1) Document Set
Presently, there are approximately 2,000,000 documents presented in various genres of contents, and the volume is increasing steadily at a high speed. SGML (Standard Generalized Markup Language) and DTD (Document Type Definition) is added in each document.

(2) Topic
The queries used in TREC are quite different from those in the past. The queries, called "Topics," represent information needs in various fields. TREC produces 50 new topics every year. The representation or structure of the topics might be revised according to the previous evaluation results or particular testing requests. Topics in TREC-1 and TREC-2 contain a complicated and detailed format with more than ten fields. The topics developed after TREC-3 are simplified. Generally, a topic consists of three main fields: <title>, <description>, and <narrative>, which can be seen as the various presentations of information needs.

(3) Relevance Judgment
TREC adopts binary decisions in relevance judgment. That is to say, all documents are divided into two groups: "relevant" and "irrelevant". A partially relevant document is seen as relevant. TREC uses the "pooling method" to create the candidate set for relevance judgments. The top *n* documents in each submitted run from various participating groups for a given topic are merged into a pool for judgment. The main purpose of the pooling method is to locate the relevant documents as exhaustively as possible through different IR systems and IR techniques. Thus the workload of the relevance judgment may be alleviated substantially.

Many test collections have been constructed based on the model of TREC especially in the design of topics. For example, Institute of Information Scientific and Technique (INIST) in France has initiated the AMARYLLIS project and built a TREC-like test collection (AMARYLLIS, 1996); the CLEF (Cross Language Evaluation Forum), a joint-force of Europe, has constructed multilingual test collections (CLEF, 2000); and the NTCIR (NACSIS Test Collection for IR Systems) in Japan (NTCIR, 2002), and HANTEC in Korea (HANTEC, 1999) have also developed the TREC-like test collection.

In ideal circumstances, the queries in test collections should be constructed according to the real information needs of users, but it is not easy to collect such information. In addition, not all of them are applicable for the purposes of IR evaluation. Therefore, the queries are usually built by simulation or revision of original users' queries such as that done in Cranfield II and TREC. Consequently, some researchers argue that the queries are too artificial to facilitate the validity of IR systems evaluation. (Sparck Jones, 1995; Borlund & Ingwersen, 1997; Salton, 1992) The following section will describe the design of a Chinese test collection.

3. The Design of CIRB

In this era of the information explosion, the significance of creating a large-scale test collection for IR evaluation is identified worldwide. However, an appropriate test collection for Chinese text retrieval has

not yet been established in the Chinese research community. Although TREC has initiated a Chinese IR track (Voorhees & Harman, 1996), the size of the test collection used is not big enough. Therefore, we have decided to construct a test collection for Chinese text retrieval, the Chinese Information Retrieval Benchmark (CIRB). A test collection for information consists of three parts: a document set, a topic set, and an answer set (relevance judgment). This section will clearly describe the three components of CIRB test collection.

3.1 Document Set

The documents of CIRB are news articles. We contacted three news agencies and received permission to use five newspapers for research and academic purposes. Table 1 lists the statistics of the document set.

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

Table 1. Document Set

In order to facilitate the process of identification and analysis of the contents, we add tags to identify each part of a document and each document is encoded in BIG5 with XML-style tags. The meaning of each tag is described below. Figure 1 shows a sample document.

- <doc> </doc>: Denote the beginning and the ending of a document.
- <id> </id>: Denote the identification code of the document, which is composed of the source, the subject category, and the serial number of the document. The code can also be seen as the full file path of each document in our data CD. The document is assigned a 7-digit serial number in each category, so we can identify each document and recognize its source by this unique id code.
- <date> </date>: Denote the date of the news using ISO8601 format. The date 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.

```
<doc>
<id>cts_foc_0005657</id>
<date>1999-05-07</date>
<title>解決高鐵融資 尋求第三管道</title>
<text>
<p>
【記者羅兩莎台北報導】據負責台灣高速鐵路聯合貸款的主辦銀行表示，高鐵融資問題目前仍卡在銀行團、交通部高鐵局以及台灣高鐵公司「三方合約」內容的訂定。在銀行團和交通部一直未能就相關歧見達成共識之下，三大主辦銀行原則決定，將尋求行政院經建會等第三管道與交通部協調，以儘早解決銀行團和交通部之間對融資問題的歧見。</p>
<p>
高鐵案將向國內銀行融資二千八百多億元，這項聯貸案確定由交銀、台銀和中國國際商業銀行共同主辦。不過，由於高鐵是國內首宗ＢＯＴ案，潛在風險究竟有多高，銀行無從評估。
</p>
<p>
據主辦銀行主管表示，銀行當然希望債權確保不會有問題，譬如，在三方合約中訂定，由政府出面保證萬一將來台灣高鐵公司蓋不下去時，政府可以出面買下，負責把工程完成等。
</p>
</text>
</doc>
```

Figure 1. Example of CIRB Document Tagging

3.2 Topic Set

Three main procedures of topic construction for CIRB are shown as follows.

(1) Collect information request

In order to simulate the real world, the topic set is constructed based on genuine users' requests. A questionnaire is posted on a portal website in Taiwan and on our website. The questionnaire consists of closed and open questions, which are the types and subject of requests, brief description requests, detailed description of requests, and other related information. The assumption of the method is that users could state their specific information request distinctly and exhaustively. In total, 405 requests were finally collected.

(2) Select information request

The responses of questionnaires gained from the Internet were not as qualitative, complete and exhaustive as required, so we screened the 50 best requests out of the 405 collected requests based on

criteria in three phases as the following explanation. In the first phase we examined the statements and narratives of the questionnaire, and then deleted simple, short, ambiguous, and subjective requests. We also excluded the requests with the following criteria: (a) the coverage of subject is too broad; (b) the request shows a great gap from the document set; (c) possible answers to the request change rapidly from time to time. In the second phase, a full-text information retrieval system is used to search possible relevant documents. The motive of this method is to determine if the request is too broad or too narrow in subject coverage. In the final phase, we selected the 50 remaining requests best fitting the following criteria: (a) explicitness of requests; and (b) distinctness of request. The results of the request selection are shown as Table 2.

	1st selection	2nd selection	3rd selection
Selection method	by persons	by IR system	by persons
# of topics deleted	163	173	19
# of topics remained	242	69	50

Table 2. The Selection of Information Request

(3) Construct topics

The main task of this procedure is to establish the topics based on the content of the 50 final requests. Four fields: title, question, narrative, and concepts are used to represent topics in accordance with the TREC convention. The meaning, content, syntax and resources of each field are shown in Table 3. The "title" field has the widest coverage in its content in comparison to the other three fields. The coverage of the "question" field is second to the "title" field. The "narrative" field is the most specific because of its detailed description. The keywords in the "concepts" field touch on the contents of the above three fields. The 50 topics can be roughly classified into nine categories. The average number of words in a topic is 169. No significant differences in length are found in the corresponding fields among different topics with comparison to other test collections. The statistics are shown in Table 4. Figure 2 shows a sample topic.

After finalizing the topics, the English counterpart of the Chinese topic set is created as well. As a result, Chinese topics and English topics could be used at the same time to measure the performance of IR systems for the comparison of language factors.

Fields	Content	Syntax	Resources
<title>	Concise representation of the information request subject.	Noun or Noun Phrase	The subject of information request
<question>	Brief descriptions of the content of the information request.	One or two sentences	The whole information request
<narrative>	Narratives of the information request such as the further interpretation to request and proper nouns, the list of relevant or irrelevant information, and the specific requirements or limitations of relevant documents.	Several sentences	The whole information request
<concepts>	Keywords relevant to the whole topic.	One or more keywords	Relevant keywords about information request

Table 3. Fields of CIRB Topics

	Fields	Minimum	Maximum	Average	Standard Deviation	Standard Deviation/Average
CIRB	<title>	3	13	6.52	2.23	0.34
	<question>	12	37	23.64	5.92	0.25
	<narrative>	57	141	93.90	20.43	0.22
	<concepts>	26	74	44.68	11.58	0.26
	Total	103	244	168.74	27.77	0.16
TREC Chinese Topic 1-54	<title>	4	29	12.30	5.58	0.45
	<desc>	6	35	17.48	7.40	0.43
	<narr>	31	174	81.54	30.28	0.37
	Total	53	204	111.32	31.36	0.28
TREC-6 Topic 301-350 (Chinese translation)	<title>	3	13	6.80	2.28	0.34
	<desc>	7	87	30.14	16.88	0.56
	<narr>	26	217	94.56	42.15	0.45
	Total	64	237	131.5	42.03	0.32

Table 4. Document Length of CIRB and TREC

<topic>
<number> CIRB010TopicZH011</number>
<title>金融機構合併。</title>
<question>
查詢我國政府單位鼓勵金融機構合併之各項措施。
</question>
<narrative>
財政部等相關單位爲健全金融市場、改善金融體質，推動了一連串鼓勵銀行、證券商及保險公司等金融機構合併的措施。相關文件內容包括各項具體的獎勵優惠辦法、施行細節、法令中明定之規範條文、以及各界對相關政策的討論與評估。若文件中只陳述金融機構合併之個案，視爲不相關。
</narrative>
<concepts>
金融機構、合併、銀行合併、租稅優惠、租稅減免、稅前盈餘、低利融資、促進產業升級條例、財政部、經濟部、央行、中央銀行、增值稅、印花稅、證交稅。
</concepts>
</topic>

Figure 2. Example of CIRB Topic

3.3 Answer Set (Relevance Judgments): Pre-Task

In general, the answer set of a test collection is constructed using a pooling method during system evaluation. (Voorhees & Harman, 1996) This kind of answer set could be called a post-task answer set. Some answer sets of test collections are constructed before they are used to evaluate IR systems. As a result, the answer set of this kind of test collection could be called a pre-task answer set. For example, the AMARYLLIS test collection has a pre-task answer set. Its post-task answer set is constructed based on the pre-task answer set and the search results submitted by participating IR systems. The following will describe how we constructed the pre-task answer set.

Basically, the main function of relevance judgments is to establish the relation between topics and documents, and it is constructed based on the following assumptions:

- Relevance is meaningful for information retrieval, and we may appropriately evaluate the performance of IR systems with respect to the concept of relevance.
- The assessors can make objective judgments based on the content of topics without being influenced by outside factors or by their own personal factors.
- The results of relevance judgments are stable, and the assessors do not have to repeat judgments several times.
- The assessors can transform the relation between topic and document into corresponding relevant scores or categories.

Relevance judgment is the important part of a test collection. As a result, we should set up criteria and procedures in the first place, including choosing the assessors, deciding measuring scales, and establishing judging rules. Next we have to build a relevant document pool (the candidate set of relevant documents) for each topic to narrow down the scope and number of documents being judged. After judging, we calculated the relevance score based on the judgments. The procedures are explained below:

(1) Identify rules for relevance judgments
Since the topics are constructed based on the genuine users' requests, it is difficult to ask the providers of these requests to carry out relevance judgments for each document. Therefore, we invite three assessors in the roles of subject expert, retrieval expert, and lay user to minimize negative factors. This methodology may enhance both the practicability and reliability of relevance judgments at the same time. The three assessors with different backgrounds and characteristics in information retrieval may reflect the divergent perceptions in common IR circumstances. The "subject relevance" concept is adopted in relevance judgment. The assessors have to objectively link the document and the topic in relation to four relevance categories: highly relevant, relevant, partially relevant, and irrelevant. These categories are assigned relevance scores from 3 to 0, respectively.

(2) Build a relevant document pool
It takes time and manpower to carry out relevance judgments. It is impossible to judge the relevance of all documents for topics. The "pooling method" is used to construct a candidate set of relevant documents. In order to do the work more completely, we use various search strategies such as query expansion by hand.

(3) Execute relevance judgments
While performing relevance judgments, every assessor has to understand the meaning of the topic, peruse document contents in the pool, and assign each of them the most appropriate category based on <question> and <narrative> fields. The documents in the pool are ranked according to

their identifiers. There are approximately 5,000 documents in 50 pools, and every document has to be judged three times. The total number of relevance judgments is around 15,000 and it takes about 230 working hours. Table 5 shows the statistics of the pool size.

# of Docs	# of pools		
31-50	14	Average	93.82
51-100	15	Maximum	198.00
101-150	12	Minimum	30.00
151-200	9	Standard deviation	47.14
Total	50	Total	4691.00

Table 5. Pools of relevant documents

(4) Compute relevance scores

A list of relevance scores for documents for a given topic has to be prepared for the test collection. Since three assessors are employed to carry out relevance judgments, we have to combine three scores into an integrated score. The integration conforms to the following two considerations:

- Each assessor has an equal contribution to the final score.
- Each judgment is independent.

According to the above considerations, we combine three judgments of the same document-topic pair using the following formula:

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

where X denotes the relevance category assigned by an assessor, and A,B,C represents each assessor. The closer the score is to 1, the higher the relevance is.

Since the TREC scoring program will be used for this performance measurement (Voorhees & Harman, 1996), thresholds have to be decided for binary relevance judgments. Two thresholds are decided: one is 0.6667; the other is 0.3333. The first threshold is for "rigid relevance," the second is for "relaxed relevance." The so-called rigid relevance means the final relevance score should be between 0.6667 and 1. That is to say, it is expected that each assessor will assign a "relevant (2)" to the document.

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

The so-called relaxed relevance means the final relevance score should be between 0.3333 and 1. That is to say, it is expected that each assessor will assign a "partially relevant (1)" to the document.

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

3.4 The Consistency of Relevance Judgment

Numerous factors may influence relevance judgments. People with different knowledge, intelligence, perception or experience may make different judgments. Owing to the complicated and uncertain factors mentioned above, we can't control all the variables involved in the experiment. In order to investigate the consistency of assessors, we use three different statistics to examine consistency of relevance judgments. These statistics are Kappa Coefficient of Agreement (*K*), Kendall Coefficient of Concordance (*W*) and Coefficient of Consistency (*C*) considering the variation in relevance scores. The definition of C is shown as follows:

$$C = 1 - \frac{|X_A - X_B| + |X_B - X_C| + |X_C - X_A|}{6}$$

Figure 3 shows the performance of 50 topics under the three statistics. The curves and distributions in the graph are quite similar. Values of *W* and *C* are both above 0.7 with an average exceeding 0.8, and the

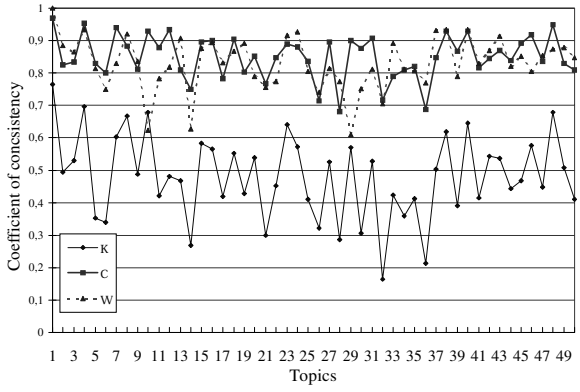

Figure 3. Consistency of Assessor

standard deviation coefficient is less than 0.1. In addition, we also compute the significance of Kendall (*W*) and Kappa (*K*). The results show that the values of significance are enough, while α is equal to 0.001. Accordingly, it could be inferred that the consistency

between the three assessors is significant, and the relevance judgments are sufficiently reliable.

For detailed information on CIRB the reader can be referred to the CIRB website. (CIRB, 2000)

4. CIRB in NTCIR Workshop

NTCIR-1 is the first evaluation workshop designed to enhance research in Japanese text retrieval. (Kando, 1999) We have negotiated this kind of joint effort in evaluating East Asian language text retrieval with Dr. Kando over a long period of time. NTCIR-2 is the result of this attempt and is the first evaluation workshop designed to enhance research in Japanese and Chinese text retrieval. A Chinese Text Retrieval (CHTR) task using the CIRB test collection has been initiated in NTCIR-2. The CHTR task falls 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). That is to say, the CHTR task could be regarded as an evaluation of Chinese text retrieval systems with multilingual queries. Both CHIR and ECIR are ad hoc IR tasks, that is, the document set is fixed for various queries (topics).

The goals of CHIR and ECIR tasks in NTCIR are shown as follows:

- Promote Chinese IR research.
- Investigate effective techniques for Chinese IR.
- Construct a mechanism for Chinese IR evaluation.
- Provide a forum to present research results and share research ideas.

Sixteen groups from seven countries or areas had enrolled in CHTR tasks. Among them, 14 groups enrolled in the CHIR task and 13 groups enrolled in the ECIR task. However, not all enrolled groups submitted search results. The search results of 115 runs were submitted from 11 groups; 98 runs from 10 groups are for the CHIR task; and 17 runs from 7 groups are for the ECIR task.

We distinguish each run according to the use of the topic. Four different types of run are defined as the follows.

- Long query ("LO"): Any query using the <narrative> of the topics.
- Short query ("SO"): Any query using no <narrative> of the topics.
- Very short query ("VS"): Any query using neither <narrative> nor <question> of the topics.
- Title query ("TI"): Any query using the <title> of the topics only.

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

Since the CIRB test collection is used in the CHTR task, we could construct a post-task answer set based on the search results submitted by all of the participating IR systems. The pooling method is also used here. On average, the size of pool for each topic is about 900 documents. Table 6 shows the total number of documents in the pool.

Topic 001	1035	topic 018	716	topic 035	1008
Topic 002	922	topic 019	896	topic 036	898
Topic 003	1016	topic 020	880	topic 037	918
Topic 004	956	topic 021	987	topic 038	540
Topic 005	1015	topic 022	1271	topic 039	720
Topic 006	703	topic 023	845	topic 040	1134
Topic 007	1175	topic 024	905	topic 041	935
Topic 008	1177	topic 025	895	topic 042	912
Topic 009	663	topic 026	1158	topic 043	765
Topic 010	1145	topic 027	812	topic 044	897
Topic 011	995	topic 028	807	topic 045	881
Topic 012	1086	topic 029	830	topic 046	764
Topic 013	748	topic 030	923	topic 047	720
Topic 014	928	topic 031	704	topic 048	1017
Topic 015	882	topic 032	736	topic 049	989
Topic 016	926	topic 033	710	topic 050	754
Topic 017	889	topic 034	776	average	898.48

Table 6. The Size of Pool for each Topic

4.1 Results of CHIR Evaluation

The recall/precision graphs of the top runs of the CHIR task are shown in Figure 4 (relaxed relevance) and Figure 5 (rigid relevance). Automatic feedback to carry out query expansion shows better performance. We also find that the stop-word list is a good resource

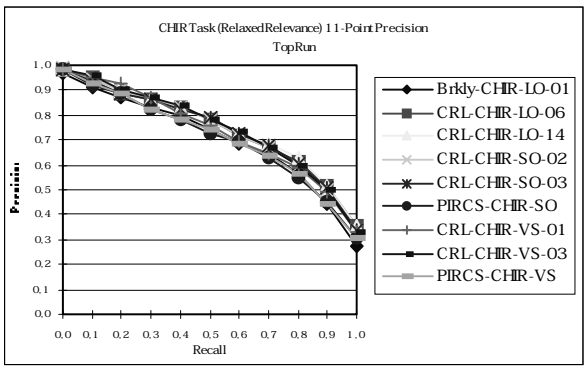

Figure 4. CHIR Task (Relaxed Relevance)

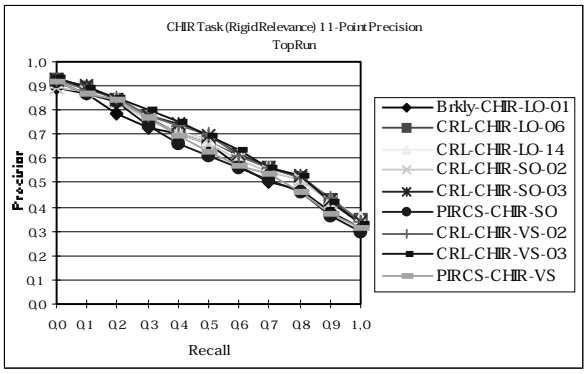

Figure 5. CHIR Task (Rigid Relevance)

for Chinese information retrieval and produces good results. A logistic regression technique also shows good performance in the CHIR task.

Basically, most of the participating groups use a *tf/idf*-based approach in a little different modification. This shows that the long-history *tf/idf* approach still plays an important role in information retrieval. Some participating groups adopting a probabilistic model show good performance in the CHIR task.

In general, the performance of the "Very Short Query" is the best; the performance of the "Short Query" is better than that of the "Long Query". The performance of the "Title Query" is the worst. It seems that the long query conveys considerable noise. Conversely, the title query conveys little information. Observing the search results, we conclude that the short query is appropriate for the CHIR task. It seems that the name of the "Very Short Query" will mislead us to draw a quick conclusion that the query should be short. In fact, the "Very Short Query" means that the participants could use the keywords shown in <concepts> field of topics. In the case of CIRB010, the <concepts> field contains many significant keywords. As a result, the runs applying "Very Short Query" perform well. (Chen & Chen, 2001)

4.2 Results of ECIR Evaluation

The recall/precision graphs of top runs of the ECIR task are shown in Figure 6 (relaxed relevance) and Figure 7 (rigid relevance). One participating group applies the Machine Translation (MT) approach to translating queries and shows good performance. However, most groups use dictionaries with select-all, select-top-1, select-top-n, or select-all approaches. Using the 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 a corpus-based approach at the same time. However, there is not enough information to show how participating groups utilize the corpus. Did they calculate mutual information? Did they calculate the bilingual mutual information?

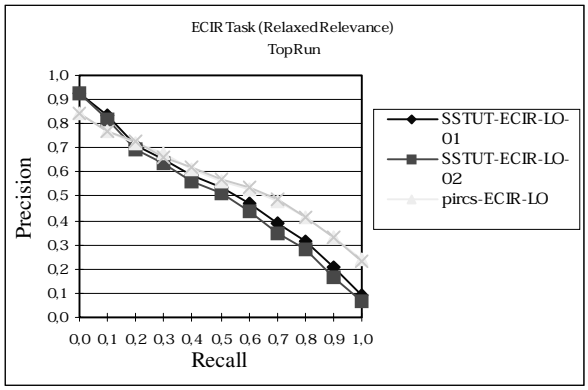

Figure 6. ECIR Task (Relaxed Relevance)

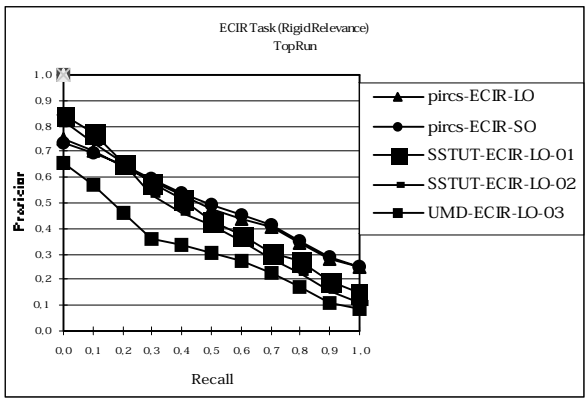

Figure 7. ECIR Task (Rigid Relevance)

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

Since only one group submits runs of all query types in the ECIR task, we will compare the performances of each query type based on its search results. 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.

The keywords in the <concepts> field for the CIRB topics provide significant information and make the performance using keywords better than other IR evaluation forum. We would like to explain the procedure for keywords preparation. We have executed a

pre-test for the CIRB010 test collection. Therefore, 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 using techniques of natural language processing. According to our analyses, the use of keywords in the <concepts> field as queries will produce the best performance when compared to the other fields. This process could be regarded as pseudo-relevance feedback using a natural language processing technique. This result emphasizes that the relevance feedback is an important technique to increase retrieval precision and recall. This also suggests that information retrieval should be a session-based interactive activity rather than a single and independent activity.

5. Towards Multilingual IR Evaluation

After applying the CIRB test collection in the CHTR task of NTCIR Workshop 2, NTCIR has become a joint-force effort. 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 author and colleagues will be in charge of the cross-language information retrieval task (CLIR).

The purposes of the CLIR task are shown as follows:

- Make the evaluation environment more realistic according to the cross-cultural and the cross-language nature of the Internet
- Provide a platform for research groups to test systems in a more complex environment
- Initiate a real cross-language multilingual text retrieval for Asian languages
- Investigate the behaviors of assessors
- Experiment with the new evaluation metrics
- Bridge research groups in Asian language retrieval

The CLIR task is a joint-effort of Japan, Korea, 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 the CLIR task in NTCIR workshop 3, the members of the executive committee met in Tokyo to decide the potential tracks, document set, topic set, criteria for relevance judgments, policy, schedule, and so forth. The following will describe the details of the 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 choose to join any one, any two, or all of the tracks. The document set consists of Chinese, English, Japanese, and Korean news articles. All but the Korean documents were published between 1998 and 1999. Table 7 shows the document set used in the CLIR task. The tag set is shown in Table 8.

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	CIRB011 (1998-1999): Chinese	132,173
	United Daily News (1998-1999): Chinese	249,508
	Taiwan News and China Times English News (1998-1999): English	10,204

Table 7. 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)

Table 8. Tags for Document Set

The topics are contributed by each country. Topics are designed to reflect the information needs of users and are represented in different level of detail. A

number of tags are used to denote information needs in topics. Table 9 shows the tags for topic.

Start Tag	End Tag	Notes
<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. A brief description of the information need, which is composed of one or two sentences.
<NARR>	</NARR>	The <NARR> has to be detailed, like the further interpretation of the request, the list of relevant or irrelevant items, the specific requirements or limitations, etc.
<CONC>	</CONC>	The keywords relevant to whole topic.

Table 9. Tags for Topic Set

The different run types based on the combination of variant fields of topic are allowed in CLIR. For example, participants could submit the 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 the TDNC run. However, the D run is a must-do run, that is, each participant has to submit a D run. In addition, each participant submits up to three runs for each language pair. Here a language pair means topic language and document language. For example, C-JE is a language pair, that is, the topic language is Chinese and the document languages are Japanese and English. The submitted runs have to be assigned a unique identifier. The format of the identifier is:

GroupId-TopicLanguage-DocLanguage-RunType-dd,

where *GroupId* is a group identifier named by the 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 two 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, respectively, 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, and so forth.

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 CLIR task was launched on 2001-09-30 and was closed on 2002-10-10.

6. Relevance Judgment Tool

The research of human behaviors has become more and more important in recent years. The paradigm, shifting from system-oriented to human-oriented research is clearly in keeping with the postmodernism increasingly accepted by many researchers. This paradigm also has impacts on research into information retrieval. Many researchers have paid much more attention to users' behaviors than ever before.

From the viewpoint of evaluating IR systems, the assessors' behaviors are also important. Since, the stage of relevance judgment is a crucial and controversial issue in IR evaluation, we would like to further investigate the whole process of relevance judgment. As a result, we have designed a tool, ReleAssessor, for relevance judgment and the storing of each event in relevance judgment. Figure 8 shows a snapshot of the opening window of this tool. This window consists of four parts: Name of Assessor; Setup of Relevance Scores; Task Type (pre-search evaluation or relevance judgment); and Setup of Paths for Pool file, Topic file, and Document file.

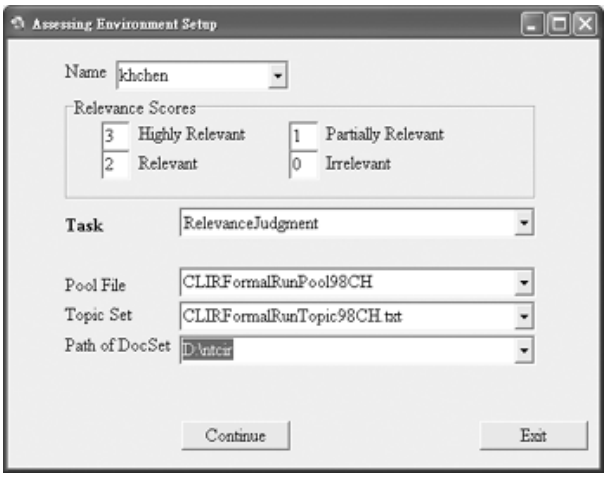

Figure 8. The Startup Snapshot of Evaluation Tool

After setting up these parameters, clicking on the "Continue" button will launch the window for topic selection as shown in Figure 9. The assessor could start judging any topic or continue judging the same topic against different documents.

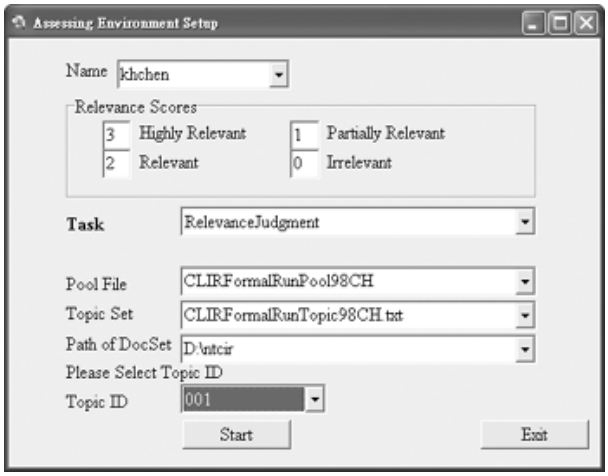

Figure 9. Select Topic for Judgment

Figure 10. Working Window for Relevance Judgment

Once the topic for relevance judgment has been selected, the main working window will appear as shown in Figure 10. This window consists of four parts: the basic information of this relevance judgment; the topic (shown in the upper-right sub-window); the document (shown in the bottom sub-window); and the given score (shown in the upper part of the document sub-window). The basic information includes Assessor Name, Topic ID, PoolFile, TopicFile, Document Rank, DOCNO, and Path of Document. The two buttons are for "Exit" and for going to the "Main Menu"; another two buttons are for "Give Comments" and "Consult Previous Judgment".

If the assessor would like to give comments or identify supporting texts for scoring from the document, s/he could click the "Give Comments" button. Two sub-windows will occupy the upper-right of the main windows: one for comments and the other for supporting texts as shown in Figure 11. The assessor could copy the words into the document and paste them into the "Supporting Texts" sub-window or give the comments in the "MEMO" sub-window where this document is scored.

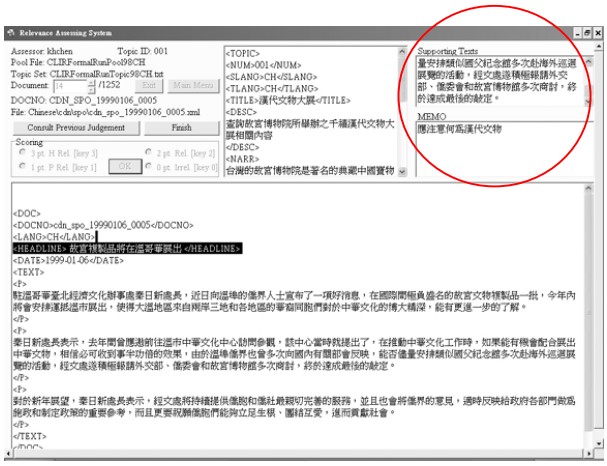

Figure 11. Comments and Supporting Texts

Figure 12. Consult Previous Judgment

If the assessor is unsure of how to score the current document, s/he could click the "Consult Previous Judgment" button, launching a window of the history of the relevance judgments as shown in Figure 12. The "Previous Judgment" window shows the rank of the document, the score of the document, and the title of the document. Further clicking on any one record will trigger the "Supporting Texts" and "MEMO" and show the appropriate information.

In order to store the log information for analyses, a set of XML-style tags is defined for documentation of various events. Table 10 shows the tags used in the log file.

Tags	Notes
<LOG> </LOG>	The top-level tag for log file
<EVENT> </EVENT>	The top-level tag for each event
<TYPE> </TYPE>	Identify event types
<EVTID> </EVTID>	Identifier of event
<TIME> </TIME>	Event time
<TOPICNO> </TOPICNO>	Topic ID involved in this event
<DOCNO> </DOCNO>	Document ID involved in this event
<MEMO> </MEMO>	Memo in judging relevance
<SUPPORT> </SUPPROT>	Supporting texts in judging relevance
<SCORE> </SCORE>	Score for relevance judgment

Table 10. The Tags for Log

The DTD for Log file is defined as Figure 13 shows.

```
<?xml version=1.0?>
<!-- top level LOG -->
<!ELEMENT LOG (EVENT*)>
<!ELEMENT EVENT (EVTID, TYPE, TIME,
   TOPICNO?, DOCNO?, MEMO?, SUPPROT?,
   SCORE?)>
<!ELEMENT EVTID (#PCDATA)>
<!ELEMENT TYPE (#PCDATA)>
<!ELEMENT DATE (#PCDATA)>
<!ELEMENT TIME (#PCDATA)>
<!ELEMENT TOPICNO (#PCDATA)>
<!ELEMENT DOCNO (#PCDATA)>
<!ELEMENT MEMO (#PCDATA)>
<!ELEMENT SUPPORT (#PCDATA)>
<!ELEMENT SCORE (#PCDATA)>
<!-- end of DTD -->
```

Figure 13. The DTD for Log File

7. Conclusions and Future Researches

We have developed a test collection for Chinese text retrieval called Chinese Information Retrieval Benchmark (CIRB010) and cooperated with NII Japan to hold a task of Chinese Text Retrieval against multilingual queries in NTICR Workshop 2 using this test collection. Pre-task evaluation has been investigated for assessors' consistency. Post-task evaluation has also been executed to identify useful techniques for Chinese text retrieval.

Many fruitful experiences have promoted learning and some effective IR techniques for Chinese text retrieval have been identified in the workshop.

- Most participating groups apply the inverted file approach for index structure.
- Many participating groups adopt the *tf/idf*-based approaches.
- "Short Query" and "Very Short Query" show much better performance.
- Query expansion is a good method to increase system performance.
- In general, the probabilistic model shows better performance.
- For the CHIR task, a stop-word list is a good resource for enhancing system performance.
- For the ECIR task, the select-all approach seems to be better than other select-X approaches, if no other corpus-based techniques are adopted.
- For the ECIR task, the MT approach is much better than a dictionary-based approach.
- For the ECIR task, a word-based indexing approach is better.

We would like to say that these findings are drawn from our evaluation results of the CHTR task using the "CIRB010" test collection. They cannot directly apply to other test collections, since each test collection has its characteristics and each language also has its characteristics. We have to carry out more detailed analyses using other test collections to reach more substantial conclusions.

In order to provide real cross-language multilingual text retrieval tasks, a multilingual test collection including Chinese, English, Japanese, and Korean documents has been developed by Japan, Korea, and Taiwan and was announced in October 2002 after the NTCIR workshop 3. This more complicated environment has been designed for examination of cross-language information retrieval and new measurement will been applied for performance evaluation.

A relevance judgment tool, ReleAssessor, has been designed to help assessors to judge relevance and to store the events of relevance judgment. The log file created by ReleAssessor will be used to analyze the behavior of the assessor. This will help us to understand the impacts of assessors upon relevance judgment and system performance. The author is planning to analyze these log files in the near future.

Acknowledgment

The author would like to thank colleagues of CIRB and NTCIR for their contribution to the research on test collections. Many thanks go to the China Times, Commercial Times, China Times Express, Central Daily News, China Daily News, and United Daily News, for their kindly providing news articles. The work described in this paper was partially sponsored by the National Science Council of the Republic of China under the NSC grant number NSC88-2213-E-002-035.

References

AMARYLLIS (1996). *Project Amaryllis*. INIST-CNRS, Meudon, France.

Bawden, D. (1990). *User-oriented Evaluation of Information Systems and Services*. Aldershot: Gower.

Borlund, P., & Ingwersen, P. (1997). The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. *Journal of Documentation, 53*(3), 225-250.

Bush, V. (1945). As We May Think. *Atlantic Monthly, 176(1)*, 101-108. Retrieved May 15, 2002, from <http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>

Chen, K.-H., & Chen, H.-H. (2001). The Chinese Text Retrieval Tasks of NTCIR Workshop 2. In N. Kando (Ed.), *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization (NTCIR 2)*, (pp. 51-72). Tokyo: NII.

CIRB (2000). Chinese Information Retrieval Benchmark Homepage. Retrieved May 15, 2002, from <http://lips.lis.ntu.edu.tw/cirb/index.htm>

CLEF (2000). *Cross-Language Evaluation Forum Homepage*. Retrieved May 15, 2002, from <http://clef.iei.pi.cnr.it:2002/>

Cleverdon, C. W. (1967). The Cranfield Tests on Index Language Devices. *Aslib Proceedings*,*19*(6), 173-194.

Fox, E. A. (1983). *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. Retrieved May 15, 2002, from <http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR83-561>

HANTEC (1999). *HANTEC Homepage*. Retrieved May 15, 2002, from <http://hantec. kisti.re.kr/>

Harman, D. K. (1996). Panel: Building and Using Test Collections. *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* Switzerland, 335-337.

Hersh, W. (1994). OHSUMED: An Interactive Evaluation and New Large Test Collection for Research. *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Ireland, 192-201.

IREX (1999). *Information Retrieval and Extraction Exercise Homepage*. Retrieved May 15, 2002, from <http://cs.nyu.edu/cs/projects/proteus/irex/index-e.html>

Kageura, K. et al. (1997). NACSIS Corpus Project for IR and Terminological Research. *Proceedings of Natural Language Processing Pacific Rim Symposium '97*, Thailand, 493-496.

Kando, N. et al. (1999). Overview of IR Tasks at the First NTCIR Workshop. *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, Japan, 1-46.

Kitani, T. et al. (1998). Lessons form BMIR-J2: A Test Collection for Japanese IR Systems. *Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval*, Australia, 345-346.

NTCIR (2002). *NACSIS Test Collection for IR Systems Homepage*. Retrieved May 15, 2002, from <http://research.nii.ac.jp/ntcir/>

Salton, G. (1972). A New Comparison between Conventional Indexing (MEDLARS) and Automatic Text Processing (SMART). *Journal of the American Society for Information Science, 23*(1), 75-84.

Salton, G. (1992). The State of Retrieval System Evaluation. *Information Processing & Management*, *28*(4), 441-449.

Shaw, W. M. et al. (1991). The Cystic Fibrosis Database: Content and Research Opportunities. *Library and Information Science Research, 13*, 347-366.

Shaw, W. M. et al. (1997). Performance Standards and Evaluations in IR Test Collections: Vector-Space and Other Retrieval Models. *Information Processing and Management, 33*(1), 15-36. Retrieved May 15, 2002, from <http://ruby.ils.unc.edu/~howep/perform/hypergeom.html>

Smeaton, A. F. & Harman, D. K. (1997). The TREC Experiments and Their Impact on Europe. *Journal of Information Science, 23*(2), 169-174.

Sparck Jones, K. (1981). The Cranfield Tests. In K. Sparck Jones (Ed.), *Information Retrieval Experiment* (pp. 256-284). London: Butterworths.

Sparck Jones, K. (1995). Reflections on TREC. *Information Processing and Managements*, *31*(3), 291-314.

Sparck Jones, K. and van Rijsbergen, C. J. (1976). Information Retrieval Test Collections. *Journal of Documentation, 32*, 63-73.

TREC (1992). *Text REtrieval Conference Homepage*. Retrieved May 15, 2002, from <http://trec.nist.gov/>

Voorhees, E. M., & Harman, D. K. (1996). Overview of the Fifth Text REtrieval Conference (TREC-5). In E. M. Voorhees & D. K. Harman (Eds.), *The Fifth Text REtrieval Conference (TREC-5).* Retrieved May 15, 2002, from <http://trec.nist.gov/pubs/trec5/papers/overview.ps.gz> .

Voorhees, E. M. & Harman, D. K. (1998). Overview of the Fifth Text REtrieval Conference (TREC-7). In E. M. Voorhees & D. K. Harman (Eds.), *The Fifth Text REtrieval Conference (TREC-7).* Retrieved May 15, 2002, from <http://trec.nist.gov/pubs/trec7/papers/overview.ps.gz>