CLIR at NTCIR Workshop 3

Noriko Kando
National Institute of Informatics (NII), Tokyo
kando@nii.ac.jp

Abstract: This paper introduces NTCIR Workshop, a series of evaluation workshops, which are designed to enhance research in information retrieval and related text-processing techniques, such as summarization, question answering and extraction, by providing large-scale test collections and a forum for researchers. A brief history, tasks, participants, test collections, CLIR evaluation at the workshops, and brief overviews at the third NTCIR workshop are described in this paper. To conclude, some thoughts on future directions are suggested.

1. Introduction

The *NTCIR Workshop* [1] is a series of evaluation workshops designed to enhance research in information access (IA) technologies including information retrieval (IR), cross-lingual information retrieval (CLIR), information extraction (IE), automatic text summarization, question answering, etc.

The aims of the NTCIR project are;

1. to encourage research in information access technologies by providing large-scale test collections reusable for experiments and common evaluation infrastructures
2. to provide a forum of research groups interested in cross-system comparison and exchanging research ideas in an informal atmosphere, and
3. to investigate methodologies and metrics for evaluation of information access technologies and methods for constructing large-scale reusable test collections.

The importance of large-scale standard test collections in IA research has been widely recognized. Fundamental text processing procedures for IA like stemming and indexing are language-dependent. In particular, processing texts written in Japanese or other East Asian languages such as Chinese are quite different from those in English, French or other European languages since there are no explicit boundaries (i.e., no spaces) between words in a sentence. The NTCIR project therefore started in late 1997 with emphasis on Japanese or other East Asian languages, and its series of workshops has attracted international participation.

1.1. Information Access

A term "information access (IA)" includes a whole process to make information in the documents usable to the users. A traditional IR system returns a ranked list of retrieved documents which are likely containing relevant information to the user's information needs. This retrieves relevant documents from a vast document collection and makes these documents usable for the users, and is one of the most fundamental and core process of IA. It is however not the end of the story for the users. After obtaining a ranked list of retrieved documents, the user skims the documents, does relevance judgments, locates the relevant information, reads, analyses, summarizes, compares the contents with other documents, integrates, summarizes and does an information work such as decision making, problem solving, writing, etc., based on the information obtained from the retrieved documents. We have looked such IA technologies to support the users to utilize the information in the large-scale documents in document collections. The scope of the IA is more closely related to the theory and practice of the digital libraries than IR.

In the following, the next section provides definition of several terms used in this paper. Section 3 describes the *NTCIR* Workshop and test collections. Section 4 discusses the continuum of the system- to user-oriented evaluation and the context of evaluation design. Section 5 summarizes the discussion.

3. *NTCIR*

3.1 Brief History of the *NTCIR*

The *NTCIR Workshop* is periodical events which have been taken place once per about one and half years. It was co-sponsored by the Japan Society for Promotion of Science (JSPS) as part of the JSPS "*Research for Future" Program* (JSPS-RFTF 96P00602) and the National Center for Science Information Systems (NACSIS) since 1997. In April 2000, NACSIS was reorganized and changed its name to the National Institute of Informatics (NII). NTCIR was co-sponsored by the JSPS and the Research Center for Information Resources at NII (RCIR/NII,) in FY 2000, and by the RCIR/NII and *Japanese MEXT*[1] *Grant-in-Aid for Scientific Research on Informatics (#13224087)* in and after FY2001.The tasks, test collection constructed, participants of the previous workshops are summarized in Table 1.

For the First NTCIR Workshop, the process started with the distribution of the training data set on 1st November 1998, and ended with the workshop meeting, which was held from 30 August to 1st September 1999 in Tokyo, Japan [5]. The IREX [6], another evaluation workshop of IR and IE (named entities) using Japanese newspaper articles, and NTCIR joined forces in 2000 and have worked together to organize the NTCIR Workshop since then. The challenging tasks of Text Summarization and Question Answering became feasible with this collaboration.

An international collaboration to organize Asian languages IR evaluation was proposed at the 4th International Workshop on Information Retrieval with Asian Languages (IRAL'99). In accordance with the proposal, the Chinese Text Retrieval Tasks are organized by Hsin-Hsi Chen and Kuang-hua Chen, National Taiwan University, at the second workshop, and Cross Language Retrieval of Asian languages at the third workshop.

For the Second Workshop, the process was started from June 2000 and the meeting was held on 7-9 March 2001, NII, Tokyo [7]. The process of the Third NTCIR Workshop starts from August, 2001 and the meeting will be held on 8-10 October 2002, NII, Tokyo[8].

3.2 Focus of the *NTCIR*

Through the series of the NTCIR Workshops, we have looked at both traditional laboratory-typed IR system testing and evaluation of challenging technologies. For the laboratory-typed testing, we have placed emphasis on 1) information retrieval (IR) with Japanese or other Asian languages and 2) cross-lingual information retrieval. For the challenging issues, 3) shift from document retrieval to technologies to utilize "information" in documents, and 4) investigation for evaluation methodologies, including evaluation of automatic text summarization; multi-grade relevance judgments for IR; evaluation methods appropriate to the retrieval and processing of a particular document-genre and its usage of the

[1] MEXT: Ministry of Education, Culture, Sports, Science and Technology

user group and so on.

From the beginning, CLIR is one of the central interests of the *NTCIR*. It was because CLIR between English and own languages are critical for international information transfer in Asian countries, and it was challenging that CLIR between languages with completely different structures and origins such as English and Chinese, or English and Japanese. It was also partly because CLIR techniques are needed even for monolingual text retrieval [9]. For example, a part of a document is sometimes written in English (ex. A Japanese document often contains an English abstract or figure captions, but no Japanese abstract and caption). Technical terms or new terms can be represented in four different forms; i.e., English terms with original spelling, acronyms of the English terms using roman alphabets, transliterated forms of the English terms using Japanese characters, and Japanese terms. The variety in such term expression often causes the decline of the search effectiveness and CLIR techniques are effective to overcome the problem. Moreover, in these years interests towards other Asian cultures has been increased, and importance of the technological information in other Asian countries has been sharply increased in business and industrial sectors.

Table 1. Previous NTCIR Workshops

	period	tasks	subtasks	test collections	particip -ants*		coun tries
1	Nov.1998- Sept.1999	Ad Hoc IR	J-JE	NTCIR-1	18	28	6
		CLIR	J-E		10		
		Term Extraction	term extraction		9		
			role analysis				
2	June 2000- March 2001	Chinese Text Retrieva	C-C	CIRB010	11	36	8
			E-C				
		Japanese&English IR	monolingual IR: J-J, E-E	NTCIR-1, -2	25		
			CLIR J-E, E-J, J-JE, E-JE				
		Text Summarization	intrinsic - extraction	NTCIR-2Summ	9		
			intrinsic - abstract				
			extrinsic - IR task-based				
3	Aug. 2001- Oct. 2002	CLIR	single lang IR:C-C,K-K,J-J	NTCIR-3CLIR, CIRB020, KEIB010	23	63	9
			bilingual CLIR:x-J,x-C, x-K				
			mulilingual CLIR:x-CJE				
		Patent	cross genre	NTCIR-3Patent	11		
			CLIR CCKE-J				
			optional task				
		Question Answering	task1-basic	NTCIR-3QA	14		
			task2-right anwer				
			task3-serial questions				
		Automatic Text Summarization	single document	NTCIR-3Summ	7		
			multi-document				
		Web Retrieval	survey retrieval	NTCIR-3Web	9		
			target retrieval				
			optional task: output clustering, speech driven				

n-m: n=query language, m=document language(s), J:Japanese, E:English, C:Chinese, K:Korean, x:any
*: number of active participating groups that submitted task resutls

3.3 Test Collections

A test collection is a data set used in system testing or experiments. In the NTCIR project the term "test collections" are used for any kind of data sets usable for system testing and experiments however it often means IR test collections used in search experiments.

The test collections constructed through NTCIR Workshops are listed in Table 2.

Table 2. Test Collections constructed through NTCIR

collection	task	documents			topic		relevance judgment
		genre	size	lang	lang	#	
NTCIR-1	IR	sci. abstract	577MB	JE	J	83	3 grades
CIRB010	IR	nespaper 98-9	210MB	C	CE	50	4 grades
NTCIR-2	IR	sci. abstract	800MB	JE	JE	49	4 grades
NTCIR-2 SUMM	Summ	newspaer94,95,98	180 doc	J	J	-	-
NTCIR-2TAO	Summ	newspaer98	1000 doc	J	J	-	-
KEIB010	IR	newpaper94	74MB	K	CKJE	30	4 grades
CIRB011+020 , NTCIR-3CLIR	IR	newspaper98-9	870MB	CJE		50	4 grades
NTCIR-3PAT	IR	patent full'98-9	17GB	J	CKJE	31	3 grades
		+abstract'95-9	4GB	JE			
NTCIR-3 QA	QA	newspaper98-9	282MB	J	J	200+800	2 grades
NTCIR-3 SUMM	Summ	newspaper98-9	60 doc	J	J	-	-
NTCIR-3Web	IR	HTML	100GB	J(E)	J	110	5 grades

For example, an IR test collection used in search experiments consists of;

(1) *document collection*

(2) *a set of topics*: written statements of user's search request.

(3) *relevance judgments:* a list of relevant documents for each topic (right answers)

In the retrieval experiments, relevance judgments are most expensive procedure. However, once test collections are created, they can be independent from the settings of the original experiment and can be repeatedly used in the different experiments. Some of the NTCIR test collections contain additional data such as, tagged corpus in NTCIR-1(sentences in the selected documents are segmented manually) and segmented data in NTCIR-2 (every sentence in a whole documents set is automatically segmented into words and phrases beforehand).

3.3.1 Documents

Documents were collected from various domain or genres. Each task carefully selected the appropriate domain of document collection And the task (experiment) design and relevance judgment criteria are set according to each document collection and the supposed user community who use the type of documents in the everyday tasks.

Fig 1 shows an sample record in NTCIR-1 JE. More than half of the documents in the NTCIR-1 JE Collection are English-Japanese paired. Documents are plain text with SGML-like tags in the NTCIR collections. A record may contain document ID, title, a list of author(s), name and date of the conference, abstract, keyword(s) that were assigned by the author(s) of the document, and the name of the host society.

```
<DOC>
<DOCNO>ctg_xxx_19990110_0001</DOCNO>
<LANG>EN</LANG>
<HEADLINE> Asia Urged to Move Faster inShoring Up Shaky Banks </HEADLINE>
<DATE>1999-01-10</DATE>
<TEXT>
<P>HONG KONG, Jan 10 (AFP) - Bank for International Settlements (BIS) general manager Andrew Crockett has urged Asian
economies tomove faster in reforming their shaky banking sectors, reports said Sunday. Speaking ahead of Monday's meeting at the
BIS office here of international central bankers including US Federal Reserve chairman Alan Greenspan, Crockett said he was
encouraged by regional banking reforms but "there is still some way to go." Asian banks shake off their burden of bad debt if they
were to be able to finance recovery in the crisis-hit region, he said according to the Sunday Morning Post. Crockett added that more
stable currency exchange rates and lower interest rates had paved the way for recovery. "Therefore I believe in the financial area, the
crisis has in a sense been contained and that now it is possible to look forward to real economic recovery," he was quoted as saying by
the Sunday Hong Kong Standard.</P>
<P>"It would not surprise me, given the interest I know certain governors have, if the subject of hedge funds was discussed during the
meeting," Crockett said. </P>
<P>He reiterated comments by BIS officials here that the central bankers would stay tight-lipped about their meeting, the first to be
held at the Hong Kong office of the Swiss-based institution since it opened last July. </P>
</TEXT>
</DOC>
```

Fig. 1 Sample Document (NTCIR-3 CLIR)

3.3.2 Topics

A sample topic record used in the CLIR at the NTCIR Workshop 3 is shown in Fig. 2. Topics are defined as statements of "user's requests" rather than "queries", which are the strings actually submitted to the system, since we wish to allow both manual and automatic query construction from the topics.

The topics contain SGML-like tags. A topic consists of the title of the topic, a description (question), a detailed narrative, and a list of concepts and field(s). The title is a very short description of the topic and can be used as a very short query that resembles those often submitted by users of Internet search engines. Each narrative may contain a detailed explanation of the topic, term definitions, background knowledge, the purpose of the search, criteria for judgment of relevance, etc.

```
<TOPIC>
<NUM>013</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>NBA labor dispute</TITLE>
<DESC>
To retrieve the labor dispute between the two parties of the US National Basketball Association at the end of 1998 and the agreement that they reached.
</DESC>
<NARR>
The content of the related documents should include the causes of NBA labor dispute, the relations between the players and the management, main
controversial issues of both sides, compromises after negotiation and content of the new agreement, etc. The document will be regarded as irrelevant if
it only touched upon the influences of closing the court on each game of the season.
</NARR>
<CONC>
NBA (National Basketball Association), union, team, league, labor dispute, league and union, negotiation, to sign an agreement, salary, lockout, Stern,
Bird Regulation.
</CONC>
</TOPIC>
```

Fig. 2. A Sample Topic (CLIR at NTCIR WS 3)

3.3.3 Relevance Judgments (Right Answers)

The relevance judgments were conducted using multi-grades. In relevance judgment files contained not only the relevance of each document in the pool, but also contained extracted phrases or passages showing the reason the analyst assessed the document as "relevant". These statements were used to confirm the judgments, and also in the hope of future use in experiments related to extracting answer passages.

In addition, we proposed new measures, *weighted R precision* and *weighted average precision*, for IR system testing with ranked output based on multi-grade relevance judgments [10]. Intuitively, the highly relevant documents are more important for users than the partially relevant, and the documents retrieved in the higher ranks in the ranked list are more important. Therefore, the systems producing search results in which higher relevance documents are in higher ranks in the ranked list, should be rated as better. Based on the review of existing IR system evaluation measures, it was decided that both of the proposed measures be single number, and can be averaged over a number of topics.

Most IR systems and experiments have assumed that the highly relevant items are useful to all users. However, some user-oriented studies have suggested that partially relevant items may be important for specific users and they should not be collapsed into relevant or irrelevant items, but should be analyzed separately [11]. More investigation is required.

3.3.4 Linguistic analysis (additional data)

NTCIR-1 contains a "Tagged Corpus". This contains detailed hand-tagged part-of-speech (POS) tags for 2,000 Japanese documents selected from NTCIR-1. Spelling errors are manually collected. Because of the absence of explicit boundaries between words in Japanese sentences, we set three levels of lexical boundaries (i.e., word boundaries, and strong and weak morpheme boundaries).

In NTCIR-2, the segmented data of the whole J (Japanese document) collection are provided. They are segmented into three levels of lexical boundaries using a commercially available morphological analyzer called HAPPINESS. An analysis of the effect of segmentation is reported in Yoshioka et al. [12].

3.3. 5 Robustness of the System Evaluation using the Test Collections

The test collections NTCIR-1 and -2 have been tested for the following aspects, to enable their use as a reliable tool for IR system testing:

- exhaustiveness of the document pool
- inter-analyst consistency and its effect on system evaluation
- topic-by-topic evaluation.

The results have been reported and published on various occasions [13–16]. In terms of exhaustiveness, pooling the top 100 documents from each run worked well for topics with fewer than 100 relevant documents. For topics with more than 100 relevant documents, although the top 100 pooling covered only 51.9% of the total relevant documents, coverage was higher than 90% if combined with additional interactive searches. Therefore, we conducted additional interactive searches for the topics with more than 50 relevant documents in the first workshop, and those with more than 100 relevant documents in the second workshop.

When the pool size was larger than 2500 for a specific topic, the number of documents collected from each submitted run was reduced to 90 or 80. This was done to keep the pool size practical and manageable for assessors to keep consistency in the pool. Even though the numbers of documents collected in the pool were different for each topic, the number of documents collected from each run is exactly the same for a specific topic.

A strong correlation was found to exist between the system rankings produced using different relevance judgments and different pooling methods, regardless of the inconsistency of the relevance assessments among analysts and regardless of the different pooling methods used [13–15,17]. It served as an additional support to the analysis reported by Voorhees [18].

3.4 NTCIR Workshop 1 (Nov. 1998 -- Sept. 1999)

The first NTCIR Workshop [5] hosted three tasks below;

1. *Ad Hoc Information Retrieval Task*: to investigate the retrieval performance of systems that search a static set of documents using new search topics (J>JE).
2. *Cross-Lingual Information Retrieval Task*: an ad hoc task in which the documents are in English and the topics are in Japanese (J>E).
3. *Automatic Term Recognition and Role Analysis Task*: (1) to extract terms from titles and abstracts of documents, and (2) to identify the terms representing the "object", "method", and "main operation" of the main topic of each document.

In the Ad Hoc Information Retrieval Task, the document collection containing Japanese, English and Japanese-English paired documents is retrieved by Japanese search topics. In Japan, document collections often naturally consist of such a mixture of Japanese and English. Therefore, the Ad Hoc IR Task at the NTCIR Workshop 1 is substantially CLIR, although some of the participating groups discarded the English section and performed the task as a Japanese monolingual IR.

Communications Research Laboratory (Japan)
Fuji Xerox (Japan)
Fujitsu Laboratories (Japan)
Central Research Laboratory, Hitachi Co. (Japan)
JUSTSYSTEM Corp. (Japan)
Kanagawa Univ. (2) (Japan)
KAIST/KORTERM (Korea)
Manchester Metropolitan Univ. (UK)
Matsushita Electric Industrial (Japan)
NACSIS (Japan)
National Taiwan Univ. (Taiwan ROC)
NEC (2) (Japan)
NTT (Japan)
RMIT & CSIRO (Australia)
Tokyo Univ. of Technology (Japan)
Toshiba (Japan)
Toyohashi Univ. of Technology (Japan)
Univ. of California Berkeley (US)
Univ. of Lib. and Inf. Science (Tsukuba, Japan),
Univ. of Maryland (US)
Univ. of Tokushima (Japan)
Univ. of Tokyo (Japan)
Univ. of Tsukuba (Japan)
Yokohama National Univ. (Japan)
Waseda Univ. (Japan).

Table 3. Active participants for the first NTCIR Workshop

3.5 NTCIR Workshop 2 (June 2000 -- March 2001)

The second workshop [7] also hosted three tasks, and each task was proposed and organized different research group on the topic.

1. *Chinese Text Retrieval Task (CHTR):* including English-Chinese CLIR (ECIR; E>C) and Chinese monolingual IR (CHIR tasks, C>C) using the test collection CIRB010, consisting of newspaper articles from five newspapers in Taiwan R.O.C.
2. *Japanese-English IR Task (JEIR):* using the test collection of NTCIR-1 and -2, including monolingual retrieval of Japanese and English (J>J, E>E), and CLIR of Japanese and English (J>E, E>J, J>JE, E>JE).
3. *Text Summarization Task (TSC: Text Summarization Arrange):* text summarization of Japanese newspaper articles of various kinds. The NTCIR-2 Summ Collection was used.

Each task had been proposed and organized by a different research group in a relatively independent manner, while maintaining good contact and discussion with the NTCIR Project organizing group, headed by the author. Evaluation, and what should be evaluated, have been thoroughly discussed in a discussion group.

ATT Labs & Duke Univ. (US)
Communications Research Laboratory (Japan),
Fuji Xerox (Japan)
Fujitsu Laboratories (Japan)
Fujitsu R&D Center (China PRC)
Central Research Laboratory, Hitachi Co. (Japan)
Hong Kong Polytechnic (Hong Kong, China PRC)
Institute of Software, Chinese Academy of Sciences (China PRC)
Johns Hopkins Univ. (US)
JUSTSYSTEM Corp. (Japan)
Kanagawa Univ. (Japan)
Korea Advanced Institute of Science and Technology (KAIST/KORTERM) (Korea)
Matsushita Electric Industrial (Japan)
National TsinHua Univ. (Taiwan, ROC)
NEC Media Research Laboratories (Japan)
National Institute of Informatics (Japan)
NTT-CS & NAIST (Japan)
OASIS, Aizu Univ. (Japan)
Osaka Kyoiku Univ. (Japan)
Queen College-City Univ. of New York (US)
Ricoh Co. (2) (Japan)
Surugadai Univ. (Japan)
Trans EZ Co. (Taiwan ROC)
Toyohashi Univ. of Technology (2) (Japan)
Univ. of California Berkeley (US)
Univ. of Cambridge/Toshiba/Microsoft (UK)
Univ. of Electro-Communications (2) (Japan)
Univ. of Library and Information Science (Japan)
Univ. of Maryland (US)
Univ. of Tokyo (2) (Japan),
Yokohama National Univ. (Japan)
Waseda Univ. (Japan).

Table 4. Active participants for the second NTCIR Workshop

3.6 NTCIR Workshop 3 (Sept. 2001 -- Oct. 2002)

The third NTCIR Workshop started with the document data distribution in September 2001 and the workshop meeting will be held in October 2002. We selected five areas of research as tasks; (1) Cross-language information retrieval of Asian languages (CLIR), (2) Patent retrieval (PATENT), (3) Question answering (QAC), (4) Automatic text summarization (TSC2), and (5) Web retrieval (WEB). The updated information is available at http://research.nii.ac.jp/ntcir/workshop/.

3.6.1 Cross-Language Retrieval Task (CLIR)

After the second NTCIR workshop [1, 2], researchers from Japan, Korea, and Taiwan have discussed a much more complicated cross-language information retrieval (CLIR) evaluation task, which is closer to the realistic application of

IR environment and is a real challenge to IR researchers. We propose a CLIR task in the third NTCIR workshop and organize an executive committee to fulfill this task. The CLIR Task Executive Committee consists of 9 researchers from Japan, Korea, and Taiwan. These members meet 3 times in Japan to discuss the details of CLIR Task, to make the schedule, and to arrange the agenda. Topic creation and relevance judgments on each language documents are done by each country group. Evaluation and report writing were done by Kuang-hua Chen and pooling was done by Kazuko Kuriyama.

Documents and topics are in four languages (Chinese, Korean, Japanese and English). Fifty topics for the collections of 1998–1999 (Topic98) and 30 topics for the collection of 1994 (Topic94). Both topic sets contain four languages (Chinese, Korean, English and Japanese). Context of the experimental design is "report writing".

Multilingual CLIR (MLIR): Search the document collection of more than one language by one of four languages of topics, except the Korean documents because of the time range difference (Xtopic98>CEJ).

Bilingual CLIR (BLIR): Search of any two different languages as language and documents, except the searching of English documents (Xtopic98>C, Xtopic94>K, Xtopic98>J).

Single Language IR (SLIR): Monolingual Search of Chinese, Korean, or Japanese. (Ctopic98>C, Ktopic94>K, Jtopic98>J).

When we think of the "layers of CLIR technologies"[19], the CLIR of newspaper articles closely related to the "pragmatic layer (social, cultural convention, etc) " and cultural/social differences among the countries is the issues we should attack in both topic creation and retrieval. For the scientific information transfer, CLIR between English and own language is the one of the biggest interests in East Asian countries. In these years, interests towards social/cultural aspects in East Asia is increasing especially in younger generation. Also technological information transfer among Asia is one of the critical issues in business and industrial sector. According to these changes in the social needs, the CLIR task has changed from English-Japanese Scientific documents to multilingual newspapers and patent documents.

For the next NTCIR Workshop, Korean newspaper articles published in 1998-99 in both English and Korean language will be added, then multilingual CLIR of four languages of Chinese, Korean, English which is published in Asia and Japanese will be feasible.

After executing relevance judgment, we find the number of relevant documents for some topics is small or is zero. For example, the topics created by Japanese member have few relevant Chinese documents. As a result, the members of Executive Committee of CLIR Task have discussed how to screen out the unsuitable topics for each combination of target languages based on a basic idea of keeping as many topics as possible. Here, the source language is query language and the target language(s) is/are document language(s). We adopt the so-called “3-in-S+A” criterion. The “3-in-S+A” means that a qualified topic must have at least 3 relevant documents with ‘S’ or ‘A’ score. Based on this criterion, we identify various topic sets for each combination of target language document set no matter what source (query) languages are used.

We construct a **NTCIR-3 Formal Test Collection** for the Evaluation of NTCIR-3 CLIR Task which we describe in the following.For Chinese, Japanese, and English Document Set (note that these are 1998-1999 news articles) and the accompanying 1998-1999 Topic Set with 50 topics as we send to you in the beginning of CLIR task, we create the following sub-test collection based on “3-in-S+A” criterion. In the FORMAL Test Collection, each target document set (C, J, E, CJ, CE, JE, CJE, and K) has different set of topics.

(1) NTCIR-3 Formal Chinese Test Collection

It contains 381,681 Chinese documents and 42 topics in source language of Chinese, Japanese, Korean, and English. The IDs of topics in 1998-1999 Topic Set used in this collection are;

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 45, 46, 47, 48, 49, and 50.

(2) NTCIR-3 Formal Japanese Test Collection

It contains 220,078 Japanese documents and 42 topics in source language of Chinese, Japanese, Korean, and English. The IDs of topics in 1998-1999 Topic Set used in this collection are;

2, 4, 5, 7, 8, 10, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, and 50.

(3) NTCIR-3 Formal English Test Collection

It contains 22,927 English documents and 32 topics in source language of Chinese, Japanese, Korean, and English. The IDs of topics in 1998-1999 Topic Set used in this collection are;

2, 4, 5, 7, 9, 12, 13, 14, 18, 19, 20, 21, 23, 24, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 42, 43, 45, 46, and 50.

(4) NTCIR-3 Formal CJ Test Collection

It contains 601,759 Chinese and Japanese documents and 50 topics in source language of Chinese, Japanese, Korean, and English. All of the 50 topics in 1998-1999 Topic Set are used in this collection.

(5) NTCIR-3 Formal CE Test Collection

It contains 404,608 Chinese and English documents and 46 topics in source language of Chinese, Japanese, Korean, and English. The IDs of topics in 1998-1999 Topic Set used in this collection are;

1 ,2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32 ,33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 45, 46, 47, 48, 49, and 50.

(6) NTCIR-3 Formal JE Test Collection

It contains 243,005 Japanese and English documents and 45 topics in source languages of Chinese, Japanese, Korean, and English. The topics in 1998-1999 Topic Set used in this collection are;

2, 4, 5, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, and 50.

(7) NTCIR 3 Formal CJE Test Collection

It contains 624,686 Chinese, Japanese, and English documents and 50 topics in source language of Chinese, Japanese, Korean, and English. All of the 50 topics in 1998-1999 Topic Set are used in this collection 2. For Korean Document Set (note that these documents are 1994 news articles) and the accompanying 1994 Topic Set with 30 topics as we send to you in the beginning of CLIR task, we create one sub-test collection also based on “3-in-S+A” criterion.

(8) NTCIR-3 Formal Korean Test Collection

It contains 66,146 Korean documents and 30 topics in Chinese, Japanese, Korean, and English. All of the 30 topics in

1994 Topic Set are used in this collection. To sum up, we will have a **NTCIR-3 Formal Test Collection** with 8 sub-test collections for all of the possible combination of target language documents except that Korean documents could not be combined with Chinese, Japanese, and English documents.

3.6.2 Patent Retrieval Task (PATENT)

Context of the experimental design is "search for technological trend survey". Regarding "Cross Genre Retrieval", we assumes that someone send a newspaper article clip to a patent intermediary and ask to retrieve the related patents. Search using ordinary topic fields such as <DESC>, <NARR>, etc. are accepted as well as non-mandatory runs. Topic creation and relevance judgments were conducted by professional patent intermediaries who are members of the Information Retrieval Committee at the Japan Intellectual Property Association. The task design was also done with close collaboration with these professionals.

Main Task

- *Cross-language Cross-Genre Retrieval*: retrieve patents in response to newspaper articles associated with technology and commercial products. Thirty query articles with a short description of the search request. Topics are available in Japanese, English, Chinese (simplified, traditional), and Korean.
- *Monolingual Associative Retrieval:* retrieve patents associated with a Japanese patent as input. Thirty query patents with a short description of search requests.

Optional task: Any research reports are invited on patent processing using the above data, including, but not limited to: generating patent maps, paraphrasing claims, aligning claims and examples, summarization for patents, clustering patents.

document:

- Japanese patents: 1998–1999 (ca. 17GB, ca 700K docs)
- JAPIO patent abstracts: 1995–1999 (ca. 1750K docs)
- Patent Abstracts of Japan (English translations for JAPIO patent abstracts): 1995–1999 (ca. 1750K)
- -- Newspaper articles (included in topics)

3.6.3 Question Answering Challenge (QAC)

Task 1: System extracts five answers from the documents in some order. One hundred questions. The system is required to return support information for each answer to the questions. We assume the support information is a paragraph, hundred-character passage or document that includes the answer.

Task 2: System extracts only one answer from the documents. One hundred questions. Support information is required.

Task 3: evaluation of a series of questions. The related questions are given for 30 of the questions of Task 2.

3.6.4 Text Summarization Challenge (tsc2)

Task A (single-document summarization): Given the texts to be summarized and summarization lengths, the participants submit summaries for each text in plain text format.

Task B (multi-document summarization): Given a set of texts, the participants produce summaries of it in plain text format. The information, which was used to produce the document set, such as queries, as well as summarization lengths, is given to the participants.

3.6.5 Web Retrieval Task (web)

Survey retrieval is a search for survey and aims to retrieve many relevant documents as possible. *Target retrieval* is a search aiming a few highly relevant documents to get a quick answer for the search request represented as a topic. "Topic retrieval" is a search in response to a search request and "similarity retrieval" is a search by given relevant document(s). In the relevance judgments, one-hop linked documents were also included in the consideration. A topic contain several extra fields specialized to Web retrieval such as a) known relevant documents, b) information on topic author, who is basically relevance assessors.

A. Survey Retrieval (both recall and precision are evaluated)
  A1. Topic Retrieval
  A2. Similarity Retrieval
B. Target Retrieval (precision-oriented)
C. Optional Task
  C1.Search Results Classification
  C2. Speech-Driven Retrieval
  C3. Other

3.6.6 Features of the NTCIR Workshop 3 Tasks

For the next workshop, we planed some new ventures, including:

(1) Multilingual CLIR (CLIR)
(2) Search by Document (Patent, Web)
(3) Passage Retrieval or submit "evidential passages", passages to show the reason the documents are supposed to be relevant (Patent, QA, Web)
(4) Optional Task (Patent, Web)
(5) Multi-grade Relevance Judgments (CLIR, Patent, Web)
(6) Various Relevance Judgments (Web)
(7) Precision-Oriented Evaluation (QA, Web).
(8) Various types of relevance judgments

For (1), it is our first trial of the CLEF [20] model in Asia. We would like to invite any other language groups who wish to join us by providing document data and relevance judgments or by providing query translation. For (3), we suppose that identifying the most relevant passage in the retrieved documents is required when retrieving longer documents such as Web documents or patents. The primary evaluation will be done from the document base, but we will use the submitted passages as secondary information for further analysis.

(4). For patent and Web tasks, we invite any research groups who are interested in the research using the document collection provided in the tasks for any research projects. Those document collections are new to our research community and many interesting characteristics are included. We also expect that this venture will explore the new tasks possible for future workshops.

For (5), we have used multi-grade relevance judgment so far since it is more natural to the users than binary judgments although most of the standard metrics used for search effectiveness are calculated based on the binary relevance judgments. We uses "cumulated gain" [21] and proposes new metrics, "weighted average precision" [10] for that purpose. We will continue this line of investigation and will add "top relevant" for the Web task, as well as standard metrics can be calculated by *trec_eval.*

The some results will be reported at the Workshop.

4. Discussion

The multilingual CLIR of Asian languages has just started from this year. So far, the needs for CLIR in East Asia was focused on the CLIR between own language and English, or other "international language". Mutual interests towards each other culture are increasing quite recently, especially among younger generation. Information transfer in the technological domain is also very acute. For example, Japanese technological information is a great interest for Korean industry and in other side, Korea is the second largest patent import, just next to the Unite State, to Japan. Regardless of the close interaction through the long history, each language is quite different. Even Taiwan ROC, Japan, People's Republic of Chine use "Chinese Characters" in the own languages, the pronounce and sentence structures are completely different between Chinese and Japanese and the Characters are simplified and modified in own way in PRC and Japan, so that we can not understand each other using own languages. Also there are no umbrella organization like the Europe Union in Europe. However the needs for CLIR among these languages are increasing in rather informal, or a grass-root movement like way.

In responding to such movement, some search engine provide simple CLIR functionalities and an operational information providers specialized to patent plan to start CLIR of East Asian languages from the next year. Therefore, it was several years behind from the situation in Europe, now we are in the stage to initiate the multilingual CLIR of east Asian languages.

The multilingual CLIR of East Asian languages have to tackle to new problems, like character codes (there are 4 standards for Japanese character codes, 2 in Korea and the character codes used in the simplified Chinese and the traditional Chinese are different and can not be 100% certain convert), less available resources and research staff who can understand the contents of other languages, transliteration of proper names in English documents, and so on. The structure of languages are quite different each other. In summary, the barriers for cross language information access in East Asia are various in every layer of the CLIR technologies shown below, and this bring new challenges to CLIR research and CLIR evaluation design and organization as well.

pragmatic layer: cultural & social aspects, convention
semantic layer: concept mapping
lexical layer: language identify, indexing
symbol layer: character codes
physical layer: network

For the future direction, Korean newspaper articles and English newspaper articles publish in Korea in 1998-1999, the same year of Chinese, Japanese and Taiwan-English and Japan-English. So the real multilingual CLIR of Chinese, Korea, Japanese, and English's in each country will be feasible for the next NTCIR.

Moreover, we may think of the possibility of the following CLIR, which are rather aiming to the application for the operational CLIR systems. In the following, focuses placed on the directions towards implications in operational or real life CLIR systems.

Cumulating the experiences
Switching language CLIR
Task/genre oriented CLIR
Pragmatic layer of CLIR technologies and identifying the differences
Towards CL information access

Cumulating the experiences: In the Internet environment, we have to cope with wide varieties of languages and combination of languages in operational setting. IR contains lots of language dependent process and the effective IR technologies can vary according to the languages or the combination of query and target languages. Therefore it is critical to cumulate the experiences on each language or combination of languages by reviewing and summarizing the findings of the CLIR research on them like building blocks of understanding on the effectiveness of CLIR technologies on each languages and language combination. Evaluation campaigns or workshops which wide variety of CLIR systems tested various approached on the same collection on the common infrastructures for cross-system comparison are expected to contribute to such activities to cumulate and summarize the experiences.

Switching (Pivot) language CLIR: In the Internet environment which wide varieties of languages are included, the shortness of the resource for translation knowledge is one of the critical problems for CLIR. In the real world environment, we can often find parallel or quasi-parallel document collections of own languages and English in non-English speaking countries. It has been proposed by many CLIR researchers to utilize such parallel document collections and connecting them by using English as switching or key language to obtain translation knowledge, but the actual testing of such direction of researches are seldom evaluated. It was partly because relevance judgments on such multilingual document collections are difficult to be done by an individual research group. International collaboration of evaluation activities can contribute to this direction.

Pragmatic layer of CLIR technologies which cope with social and cultural aspects of languages are ones of the most challenging issues of real world CLIR. CLIR research has so far placed emphasis on technologies to provide access to the relevant information across the different languages. Identifying the differences of viewpoints expressed in different languages or in documents produced in different cultural or social backgrounds is also critical to improve the realistic global information transfer across the languages.

Towards CL information access: In order to widen the scope of the CLIR in the whole process of the information access, we can think of the technologies to make information in the document more usable for users, for example, cross language summarization, cross language question answering, cross language text mining, and so on. The technologies to enhance the interaction between systems and users or to support the query construction on CLIR systems are also included in this direction.

References:

[1] NTCIR Project: http://research.nii.ac.jp/ntcir/

[2] TREC. http://trec.nist.gov/

[3] Smeaton, A.F. and Harman, D. "The TREC (IR) experiments and their impact on Europe", *Journal of Information Science*, No. 23, pp 169-174, 1997.

[4] Sparck Jones, K., Rijsbergen, C.J. *Report on the need for and provision of an 'ideal' information retrieval test collection*, Computer laboratory, Univ. Cambridge., 1975 (BLRDD Report)

[5] *NTCIR Workshop 1: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, Tokyo, Japan, 30 Aug.–1 Sept., 1999. ISBN4-924600-77-6. (http://research.nii.ac.jp/ntcir/workshop/ OnlineProceedings/).

[6] IREX URL:http://cs.nyu.edu/cs/projects/proteus/irex/

[7] *NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, Japan, June 2000–March 2001.ISBN4-924600-96-2. (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings/)

[8] *NTCIR Workshop 3 Meeting: Working Note of the Third NTCIR Workshop Meeting*, Tokyo, Japan, Oct.8-10, 2002. 6 vols.

[9] Kando, N.: "Cross-linguistic scholarly information transfer and database services in Japan". Presented in the panel on Multilingual Database in the Annual Meeting of the American Society for Information Science, Washington DC. , USA, November 1997.

[10] Kando, N., Kuriyama, K., Yoshioka, M. "Evaluation based on multi-grade relevance judgments". *IPSJ SIG Notes*, Vol.2001-FI-63, pp.105-112, July 2001. (in Japanese w/English abstract)

[11] Spink, A., Greisdorf, H. "Regions and levels: Measuring and mapping users' relevance judgments". *Journal of the American Society for Information Sciences,* Vol.52, No.2, pp.161-173, 2001

[12] Yoshioka, M., Kuriyama, K., Kando, N.: "Analysis on the usage of Japanese segmented texts in the NTCIR Workshop 2." In *NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, June 2000–March 2001（ISBN：4-924600-96-2).

[13] Kando, N, Nozue, T., Kuriyama, K., Oyama, K., "NTCIR-1: Its policy and practice", *IPSJ SIG Notes*, Vol.99, No.20, pp.33-40, 1999. (in Japanese w/English abstract)

[14] Kuriyama, K., Nozue, T., Kando, N., Oyama, K.: "Pooling for a large scale test collection: Analysis of the search results for the pre-test of the NTCIR-1 Workshop", *IPSJ SIG Notes*, Vol.99-FI-54, pp.25-32 May, 1999 [in Japanese].

[15] Kuriyama, K., Kando, K. "Construction of a large scale test collection: Analysis of the training topics of the NTCIR-1", *IPSJ SIG Notes*, Vol.99-FI-55, pp.41-48, July 1999. (in Japanese w/English abstract)

[16] Kando, N., Eguchi, K., Kuriyama, K., "Construction of a large scale test collection: Analysis of the test topics of the NTCIR-1", In *Proceedings of IPSJ Annual Meeting* [in Japanese]. pp.3-107 -- 3-108, 30 Sept.–3 Oct. 1999.

[17] Kuriyama, K., Yoshioka, M., Kando, N., "Effect of cross-lingual pooling". In *NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, June 2000–March 2001（ISBN：4-924600-96-2)

[18] Voorhees, E.M., "Variations in relevance judgments and the measurement of retrieval effectiveness", In *Proceedings of 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*., Melbourne, Australia, August 1998, pp.315-323.

[19] Kando, N. "Towards real multilingual information discovery and access ". Presented at ACM Digital Libraries and ACM-SIGIR Joint Workshop on Multilingual Information Discovery and Access. Panel on the Evaluation of the Cross-Language Information Retrieval. Berkeley, CA, USA, August 15, 1999. (http://www.clis2.umd.edu/ conferences/midas/papers/kando2.ppt)

[20] CLEF: Cross-Language Evaluation Forum, http://www.iei.pi.cnr.it/DELOS/CLEF

[21] Jarvelin、 K.、 Kekalainen、 J.: IR evaluation methods for retrieving highly relevant documents. Proceedings of ACM-SIGIR 2000、 p. 41-48、 2000

[22] Nozue, T., Kando, N. “Primary considerations in the concept of relevance: Relevance judgement of NTCIR”. *IPSJ SIG Notes,* 99-FI-53, Vol.99, No.20, March 1999, p. 49-56. (in Japanese w/English abstract)