Evaluation of Information Access Technologies with Asian Languages at NTCIR Workshop

Noriko Kando

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan
kando@nii.ac.jp

Abstract

This paper introduces the *NTCIR Workshop*, a series of evaluation workshops that are designed to enhance research in information access technologies, such as information retrieval, cross-lingual information retrieval, text summarization, question answering and text mining, by providing infrastructure for large-scale evaluations. A brief history, the test collections, and tasks are described. To conclude, some thoughts on future directions are suggested.

1. Introduction

The *NTCIR* Workshop is a series of evaluation workshops designed to enhance research in information access (IA) technologies including information retrieval (IR), cross-lingual information retrieval (CLIR), question answering, automatic text summarization, text mining and so on by providing large-scale test collections and a forum for researchers.

The aims of the NTCIR project are:

1. to encourage research in information access technologies by providing large-scale test collections that are reusable for experiments;
2. to provide a forum for research groups interested in cross-system comparisons 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 main goal of the *NTCIR* project is to provide infrastructure for large-scale evaluations of IA technologies. The importance of such infrastructure in IA research has been widely recognized. Fundamental text processing procedures for IA, such as indexing includes language-dependent procedures. The *NTCIR* project therefore started in late 1997 with emphasis on, but not limited to, Japanese or other East Asian languages, and its series of workshops has attracted international participation.

In *NTCIR*, a workshop is held about once every one and a half years. Because we respect the interaction between participants, we consider the whole process from initial document release to the final meeting to be the "workshop". Each workshop selects several research areas called "*tasks*", or a "*challenges*" for the more challenging tasks. Each task has been organized by the researchers of the domain and a task may consist of more than one subtask.

1.1. Information Access

The term "information access" (IA) refers the whole process from when a user realizes his/her information needs, through the activity of searching for and finding relevant documents, and then utilizing information in them. We have looked at IA technologies to help users utilize the information in large-scale document collections. IR, summarization and question answering are part of a "family", aiming at the same target, although each of them has been investigated by rather different communities.

1.2 Focus of *NTCIR*

From the beginning of the project, we have looked at both traditional laboratory-type IR system testing and the evaluation of challenging technologies, as shown in Figure 1. For the former, we placed emphasis on text retrieval and CLIR with Japanese or other Asian languages and testing on various document genres.

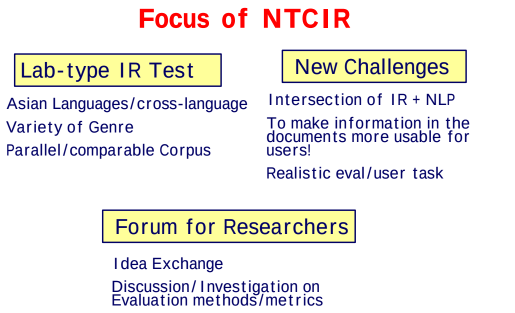
Focus of NTCIR
Lab-type IR Test
Asian Languages/cross-language
Variety of Genre
Parallel/comparable Corpus
New Challenges
Intersection of IR + NLP
To make information in the documents more usable for users!
Realistic eval/user task
Forum for Researchers
Idea Exchange
Discussion/Investigation on Evaluation methods/metrics

Figure 1. Focus of *NTCIR Workshops*

For the challenging issues, the target is to shift from document retrieval to technologies that utilize "information" in documents, and investigation of methodologies and metrics for more realistic and reliable evaluation. For the latter, we have paid attention to users' information-seeking tasks in the experiment design because they are deeply related to the appropriate types of documents, topics of the users' search requests and relevance judgment criteria even in the laboratory-type testing of the systems. These two directions have been supported by a forum of researchers who are interested in cross-system comparison and by their discussion

Table 1. Tasks of the NTCIR Workshops

	Period	Tasks	Subtasks	Test collections
1	Nov.1998- Sept.1999	Ad Hoc IR	J-JE	NTCIR-1
		CLIR	J-E	
		Term Extraction	Term Extraction/ Role Analysis	
2	June 2000- March 2001	Chinese Text Retrieval	Chinese IR: C-C	CIRB010
			CLIR: E-C	
		Japanese&English IR	Monolingual IR: J-J, E-E	NTCIR-1, -2
			CLIR: J-E, E-J, J-JE, E-JE	
		Text Summarization	Intrinsic - Extraction/Free generated	NTCIR-2Summ
			Extrinsic - IR task-based	
3	Oct. 2001- Oct. 2002	CLIR	Single Language IR:C-C,K-K,J-J	NTCIR-3CLIR
			Bilingual CLIR:x-J,x-C, x-K	
			Multilingual CLIR:x-CJE	
		Patent	Cross Genre w/ or w/o CLIR CCKE-J	NTCIR-3 PATENT
			[Optional] Alianment, RST Analysis of Claims	
		Question Answering	Subtask-1: Five Possible Answers	NTCIR-3QA
			Subtask-2: One Set of All the Answers	
			Subtask-3: Series of Questions	
		Text Summarization	Single Document Summarization	NTCIR-3 SUMM
			Multi-document Summarization	
		Web Retrieval	Survey Retrieval	NTCIR-3 WEB
			Target Retrieval	
			[Optional] Speech-Driven	
4	Apr. 2003 - June 2004	CLIR	Single Language IR:C-C,K-K,J-J	NTCIR-4CLIR
			Bilingual CLIR:x-J,x-C, x-K	
			Pivoted Bilingual CLIR	
			Multilingual CLIR:x-CKJE	
		Patent	"Invalidity Search"= Search Patents by a Patent	NTCIR-4 PATENT
			[Feasibility] Automatic Patent Map Creation	
		Question Answering	Subtask-1: Five Possible Answers	NTCIR-4 QA
			Subtask-2: One Set of All the Answers	
			Subtask-3: Series of Questions	
		Text Summarization	Multi-document Summarization	NTCIR-4 SUMM
		Web Retrieval	Informational Retrieval	NW100G-01, NTCIR-4 WEB
			Navigational Retrieval	
			[Pilot] Geographical Information	
			[Pilot] (Search Results) Topical Classification	
5	Aug. 2004- Dec. 2005	CLIR	Single Language IR:C-C,K-K,J-J	NTCIR-5CLIR
			Bilingual CLIR:x-J,x-C, x-K	
			Multilingual CLIR:x-CKJE	
		CLQA	Subtask on JE documents: JE, EJ, CE	NTCIR-5CLQA
			Subjtask on C documents: CC, EC	
		PATENT	Document Retrieval	NTCIR-4PATENT, NTCIR-5PATENT
			Passage Retrieval	
			Classification	
		Question Answering	Series of Questions (Information Access Dialog)	NW1000G-04, NTCIR-5WEB
		WEB	Navigational Retrieval Subtask	
			Query Term Expansion Subtask	

n-m: n=query language, m=document language(s), J:Japanese, E:English, C:Chinese, K:Korean, x:any of CJKE

2. History NTCIR Workshop

2.1. Tasks

As shown in Table 1 and Figure 2, has selected several areas of research as "tasks":

1. Cross-Lingual Information Retrieval (*CLIR*),
2. Term Extraction (TMREC),

3. Text Summarization Challenge (*TSC*),
4. Question Answering (*QAC*, *CLQA*),
5. Retrieval in Specialized Domain
    a) Patent Retrieval Task (*PATENT*),
    b) WEB Task (*WEB*),
6. Multimodal Summarization for Trend Information (*MuST*)

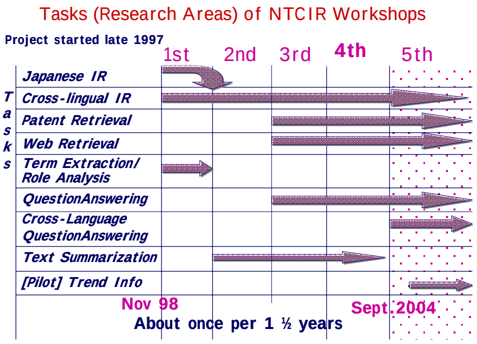

Figure 2. Tasks at *NTCIR Workshops*

2.2. Participants

As shown in Figures 3 and 4, the number of participants has gradually increased. Different tasks attracted different research groups. Many international participants enrolled in CLIR and CLQA. The PATNET task attracted participants from company research laboratories and "veteran" NTCIR participants. Classification Subtask of the PATENT attracted researchers on text categorization or machine learning, which is rather new to NTCIR.

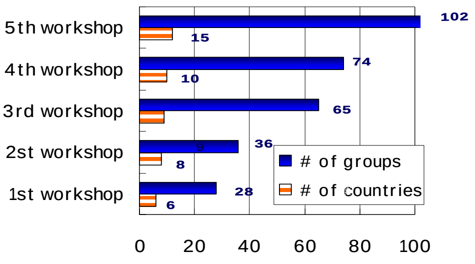

Figure 3. Number of Participating Groups

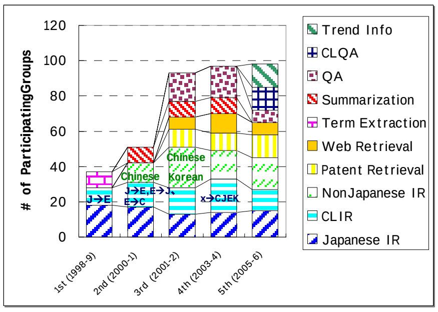

Figure 4. Number of Participants by Task

3. Test Collections

3.1. Documents

Table 2 shows the test collections constructed through the series of *NTCIR workshops*. In the *NTCIR* the term "*test collection*" is used for any kind of data set usable for system testing and experiments. One of our interests is to prepare realistic evaluation infrastructures and efforts include scaling up the document collection and increasing variety of document genres and languages.

Both patent and scientific document collections have *parallel corpora* of English and Japanese abstracts. For news document collections used in *NTCIR-3 through -5*, we prepared the corpora of news documents published in the years 1998-2001 in Chinese, Japanese, Korean and English. Each language sub-collection has multiple sources of the newspapers.

The Patent document collection contains fulltext of patent applications filed in the 10 years of 1993-2002 and it consists of about 3.5 Million documents. For WEB, the size of the document collection *NW1000GB-04* is 1.36 TB, which were mainly crawled in .jp domain.

The task (experiment) design and relevance judgment criteria were set according to the nature of the document collection and of the user community who use the type of document in their everyday life.

3.2. Topics and Questions

As shown in Figure 5, the structure of the topic in the NTCIR IR test collections is similar to that used in TREC [5] and CLEF [6]. They are defined as natural language statements of "users' search requests" rather than "queries", strings submitted to the system, so that both manual and automatic query construction can be tsted. Any field in a topic is usable for experiments as far as reported in the papers.

```
<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>
```

Figure 5. Sample Topic

3.3. Relevance Judgments and Evaluation

In IR-related tasks, relevance judgments were graded: highly relevant, relevant, partially relevant and irrelevant.

Table 3. Test collections constructed by *NTCIR*

Collection	Task	Documents						Task data		
								Topic/ Question		Relevance judge
		Genre	Filename	Lang.	Year	# of docs	Size	Lang	#	
NTCIR-1	IR	Sci. abstract	ntc1-je	JE	1988-1997	339,483	577MB	J	83	3 grades
			ntc1-j	J		332,918	312MB			
			ntc1-e	E		187,080	218MB		60	
	TE*5		ntc1-tmrc	J		2,000	-		-	-
CIRB010	IR	News	CIRB010	C_t	1998-1999	132,173	132MB	C_tE	50	4 grades
NTCIR-2	IR	Sci. abstract	ntc2-j	J	1986-1999**	400,248	600MB	JE	49	4 grades
			ntc2-e	E		134,978	200MB			
NTCIR-3 CLIR	IR	News	KEIB010	K	1994	66,146	74MB	C_tKJE	30	4 grades
		News	CIRB011	C_t	1998-1999	132,173	870MB	C_tKJE	50	4 grades
			CIRB020			249,508				
			Mainichi	J		220,078				
			EIRB010	E		10,204				
			Mainichi Daily			12,723				
NTCIR-3 PATENT	IR	Patent	kkh *3	J	1998-1999	697,262	18GB	C_tC_s KJE	31	3 grades
		Abstract	jsh *3	J	1995-1999	1,706,154	1,883MB			
		Abstract	paj *3	E	1995-1999	1,701,339	2,711MB			
NTCIR-3 QA	QA	News	Mainichi	J	1998-1999	220,078	282MB	J*	1200	exact answer
NTCIR-3 WEB	IR	Web (html/text)	NW100G-01	multiple*4	crawled in 2001	11,038,720	100GB	J*	47	4 grades + relative
			NW10G-01			1,445,466	10GB			
NTCIR-4 CLIR	IR	News	CIRB011	Ct	1998-1999	132,173	ca.3GB	CtKJE	60	4 grades
			CIRB020			249,203				
			Hankookilbo +	K		149,921				
			Chosenilbo +			104,517				
			Mainichi	J		220,078				
			Yomiuri +			373,558				
			EIRB010	E		10,204				
			Mainichi Daily			12,723				
			Korea Times +			19,599				
			Hong Kong Standard +			96,683				
			Xinhua +			208,167				
NTCIR-4 PATENT	IR	patent full	Publication of unexamined patent application +	J	1993-1997	ca. 1,700,000	ca.27GB	E	Main: 34, Add: 69	3 grades
		Abstract	Patent Abstracts of Japan (PAJ) +	E	1993-1997	ca. 1,700,000	ca.5GB			
NTCIR-4 QA	QA	News	Mainichi	J	1998-1999	220,078	ca.776MB	J*	197	exact answer
			Yomiuri +			373,558			199	
									251	
NTCIR-4 WEB	IR	Web (html/text)	NW100G-01	multiple*4	crawled in 2001	11,038,720	100GB	J*		3 grades
NTCIR-5 CLIR	IR	News	CIRB040r	Ct	2000-2001	901,446		CtKJE	50	4 grades
			Hankookilbo	K		85,250				
			Chosenilbo			135,124				
			Mainichi	J		199,681				
			Yomiuri			658,719				
			Mainichi Daily	E		12,155				
			Korea Times			30,530				
			Daily Yomiuri			17,741				
			Xinhua			198,624				
NTCIR-5 CLQA	QA	News	CIRB040r	J	2000-2001	901,446		CJE	smpl:300 + tes 200 *6	3 grades *7
			Yomiuri			658,719				
			Daily Yomiuri			17,741				
NTCIR-5 PATENT	IR	patent full	Publication of unexamined patent application	J	1993-2002	ca. 3,500,000	ca.45GB	JE	34+1189	3 grades
		Abstract	Patent Abstracts of Japan (PAJ) +	E	1993-2002	ca. 3,500,000	ca.10GB			
NTCIR-5 QA	QA	News	Mainichi	J	2000-2001	199,681		J*	50 series (360 Q)	graded
NTCIR-5 WEB	IR	Web (html/text)	NW1000G-04	multiple*4	crawled in 2004		1.36TB	J*	269 + 847	3 grades

J:Japanese, E:English, C:Chinese (C_t:Traditional Chinese, C_s: Simplified Chinese), K:Korean;
"+" indicates the document collection was newly added for NTCIR-4
* English translation is available
** gakkai subfiles: 1997-1999, kaken subfiles: 1986-1997
*3: kkh : Publication of unexamined patent application, jsh: Japanese abstract, paj: English translation of jsh
*4: almost Japanese or English (some in other languages)
*5: Term extraction/ role analysis
*6: 300+200 questions for C documents, and 300+200 questions for JE documents
*7: Right, Unsupported, Wrong

4. Areas of Researches

4.1. Cross-Lingual Information Retrieval

4.1.1. History

Cross-lingual information retrieval (CLIR) is a search that a user submits a query in a language and the system retrieves documents which may represented in other languages as well as the documents in the query language.

In Asian context, CLIR is quite different from that in European languages and really challenging as the alphabets and language structures are completely different from English or other European languages and each other. The initial stage of the CLIR is search between English and the own language. Then we started from English-Japanese CLIR at NTCIR-1 and -2 using English and Japanese quasi-parallel scientific/scholarly paper abstracts and English-Chinese CLIR at NTCIR-2 using Chinese news documents.

Among East Asian countries, there are long (more than 2000 years) historical relationships but less interaction from 1950's to early 1990's because of the very sad history in early 20th Century. Interests and human exchanges have increased acutely in these years in both commercial/industrial exchanges and cultural/social/ordinary life interests. CLIR across East Asian languages are now really needed both in business and private. Commercial CLIR systems are gradually started their service. Then we moved to CLIR across East Asian languages, tested CLIR on (Traditional) Chinese, Japanese and English documents and monolingual search on Korea at NTCIR-3, and CLIR among these four languages at NTCIR-4 and -5 using comparable corpora of news documents published in 1998-2001 in Chinese, Korean, Japanese and English. Topics were created in each countries in order to reflect each other's point of views, and manually translated into other three languages.

4.1.2. Lessons Learned

Segmentation and indexing are discussed in the first stage, n-gram, word-based, hybrid, etc. Translating technical terms and transliteration, identify named entities and special treatment for them, out-of-vocabulary (OOV) and using various resource including WEB to find the translation for them were tested and showed improvement. A kind of cognate matching was test between Chinese and Japanese on Patent documents for highly technical terms and showed effective.

Pseudo relevance feedback (PFR) is known to be effective to improve the search effectiveness on average of the multiple topics, but its effectiveness differs on different topics. Many groups proposed and tested different query expansion techniques, including PRF, Web-based, statistical thesaurus, bounce-and-throw, document re-ranking, and filter to select the topics to apply Web-based expansion. Investigation of the effects of the document length were thoroughly tested using rather uniformed length of abstracts, moderately various length of news documents and quite widely differences in Patent documents.

4.2. Term Extraction (TMREC)

How to identify the "term" is one of the first and essential problems to processing East Asian language texts. In NTCIR-1, we initiated such discussion and released a Japanese corpus with two-level term annotation.

4.3. Text Summarization Challenge (*TSC*)

Single document summarization on Japanese news documents were tested at NTCIR-2 and -3, and multi-document summarization were at NTCIR-3 and -4. Three different assessors created human created model summaries for each of the document sets for NTCIR-2 and NTCIR-3 in the two levels of summarization ratio, 10% and 20%. For NTCIR-4, unfortunately we have prepared only one human created model summary par topic, but each sentences in the model summaries were greedily tagged the correspondence between all the sentences in the input documents with the level of contribution, then using such detailed annotation automatic evaluation of the system produced summaries were enable using the NTCIR-4 Summ test sets.

4.4. Question Answering (*QAC*, *CLQA*)

Factoid types of question answering were tested at NTCIR-3 through -5 using Japanese news documents. Testing on the returning top 5 answers and a set of the multiple answers at NTCIR-3 and -4. A series of questions were tested at NTCIR-3 through -5. Especially from NTCIR-4, it was resembling the Information Access Dialog (IAD), where we set an information seeking task that a user has a broad topic to write a report in mind and submits a series of questions related to the broad topic to obtain sufficient information for the reports. Among the topics, some are "Information Gathering" type series, in which the topic of each question in a series was strictly concentrated one topic, while others can be a "Browsing" type series, in which the topics of the questions in a series can be shifted and drifted according to the interest of the user and affected from the answers for the previous question.

More complicated types of questions like "why" or "definitions" will be tested at NTCIR-6.

Cross Language Question Answering (CLQA) has started from NTCIR-5. The participation showed a good balance -- a number of QAC "veterans" groups and other strong NLP-oriented groups participated as well as new comers challenged it. There are two sets of subtasks:

- Subtasks on JE documents: JE, EJ, CE, and
- Subtasks on C documents: CC, EC.

They mainly focused named entities since identify named entities and find appropriate translation are one of the critical problems and usable technologies in the real world in the Asian context where all the character sets are different each other.

4.5. Patent Retrieval Task (*PATENT*)

Patent is a very interesting corpus. Retrieval, analysis and cross-lingual access are seriously needed in the real world. At the same time, it contains interesting characteristics for text processing research in various aspects. A patent document has structure, highly technical terms, and hierarchical classification codes. The document length is quite variable according to each document. it can be used as a technological document as well as a legal document. It was not used in NTCIR, but through the patent-family relationship, real multilingual parallel corpus can be created.

In NTCIR, we have concerned the usage of the documents in the real world and designed the tasks according to the real world information seeking tasks of the users of the document genre.

At NTCIR-3, patent retrieval for technology survey was examined. In this task, the patents were treated as technological documents. Use newspaper articles as queries, then searched related patents from 2-year fulltext collections of Japanese patent applications filed in 1998-1999. The queries were translated in other 4 languages, traditional Chinese, simplified Chinese, Korean, and English. And English and Japanese parallel abstracts filed in 1995-1999 which were prepared by professional abstractors were also included in the collection. For NTCIR-4 and after, the patent retrieval for invalidate the patent applications were tested.

Other analytical tasks using patent such as automatic patent map creation, classification were also tested.

4.6. WEB Task (*WEB*)

WEB is really critical problems in IR in various aspects, size, units of the documents, usages, various genres, etc.

In NTCIR, the WEB document collections of 10GB, 100GB, 1.36TB were used and various tasks related to different information seeking tasks of the users were tested in NTCIR-3 through -5.

4.7. Multimodal Summarization for Trend Information (*MuST*)

Multimodal Summarization for Trend Information (MuST) (Kato, Matsushita, Kando, 2005) is organized as a pilot workshop of the NTCIR. It investigates the task to extract numeric expressions from a set of documents, summarize, and visualize so that the users easily understand the tendencies among the set of documents. The examples of the topics are the stock market price, amount of import/export of a particular products, etc. Thirteen groups participated. This is an interesting mixture of different communities, IR, NLP, Web intelligence, Fuzzy, etc. The results will be presented in the separate workshop held in March, 2006.

5. 5. Discussion

A brief overview of the *NTCIR Workshop* is reported here. The details of the achievements from each task and those of each participant are reported in the reports from each task in this issue, the papers in the proceedings (Kando and Takaku, 2005).

The test collections used in the tasks of the NTCIR and the archives of the system produced submission raw data will be available for research purpose. We expect that many of the research groups involved in the larger NTCIR community will work collaboratively to investigate the system mechanisms and to analyze the further results, and then learn each other from each other's experience.

Evaluation must adapt to technological evolution and the change in social needs. We are working towards this goal, and suggestions are always welcome.

6. References

Noriko Kando, Masao Takaku, (eds): *NTCIR Workshop 5: Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo Japan, December 6-9, 2005, NII, Tokyo 2005 http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings5/

Tsuneaki Kato, Mitsunori Matsushita, Noriko Kando: MuST: A Workshop on Multimodal Summarization for Trend Information In*: Proceedings of the Fifth NTCIR Workshop Meeting*, Tokyo Japan, December 6-9, 2005, NII, Tokyo 2005.

7. Related URL

NTCIR Project: http://research.nii.ac.jp/ntcir/

NTCIR Workshop 6 (2005-2006) : http://research.nii.ac.jp/ntcir-ws6/work-en.html

NTCIR Workshop 6 Meeting (15-18 May 2007) : http://research.nii.ac.jp/ntcir/ntcir-ws6/