Overview of the Sixth NTCIR Workshop

Noriko Kando
National Institute of Informatics
Tokyo 101-8430, Japan
Noriko.Kando@nii.ac.jp

Abstract

This paper outlines the sixth NTCIR Workshop, which is the latest in a series. It briefly describes the background, tasks, participants, and test collections of the workshop. The purpose of this paper is to serve as an introduction to the research described in detail in the rest of the proceedings of the sixth NTCIR Workshop.

Keywords: evaluation, information access, information retrieval, question answering, test collections, cross-lingual information retrieval, patent retrieval, opinion analysis, cross-lingual question answering, trend analysis, visualization, test collections.

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), 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 reusable for experiments;
2. to provide a forum for researchers interested in cross-system comparisons and exchanging research ideas in an informal atmosphere; and
3. to investigate methodologies to evaluate information access technologies.

The main goal of the *NTCIR* project is to provide infrastructure for large-scale evaluations of IA technologies, and then speed the research and technology transfer. 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*". 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.

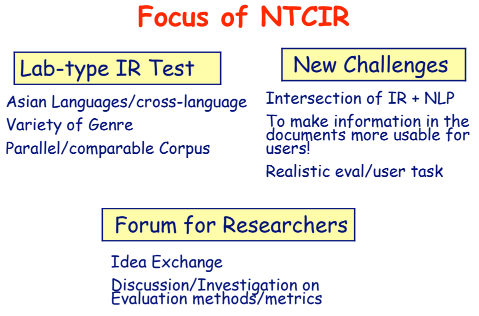
Focus of NTCIR
Lab-type IR Test
Asian Languages/cross-language
Variety of Genre
Parallel/comparable Corpus
New Challenges
Intersection of IR + NLP
To make information in the documents more usable for users!
Realistic eval/user task
Forum for Researchers
Idea Exchange
Discussion/Investigation on Evaluation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

Figure 1. Focus of *NTCIR Workshops*

Table 1. Tasks of the Past 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

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 discussions.

2. The Sixth NTCIR Workshop

2.1 Tasks

For the *Sixth NTCIR Workshop* (NTCIR-6) [2], the process started from the call-for-task-participation in April 2006 and the meeting is held on 15-18 May 2007 [3], at National Institute of Informatics (NII) in Tokyo.

It is sponsored by NII, and the Meeting was sponsored by Japan Society of Promotion of Sciences (JSPS) and NII.

The *NTCIR-6* selected six "tasks":

1. Cross-Lingual Information Retrieval (*CLIR*),
2. Cross-Lingual Question Answering (*CLQA*)
3. Multimodal Summarization of Trend *(MuST) (*as a pilot workshop)
4. Opinion Analysis Pilot *(OPNION)*
5. Patent Retrieval (*PATENT*)[1], and
6. Question Answering Challenge (*QAC*)

For *MuST,* the results of each participating group were presented in a separate workshop in March 2007[2] and selected papers will be presented at the NTCIR-6 Meeting. *OPINION* is a new task.

CLIR and *PATENT* used multiple past NTCIR test collections to obtain more substantial evaluation results and to see the variances across the collections. We see these series of experiments on multiple collections are to summarize the achievements obtained from the series of evaluation on *CLIR* and *PATENT* at NTCIR. *CLQA* and *MuST* are in the second cycle and continued the task design used in the previous one, i.e., in NTCIR-5, with minor change.

"Continuing for two cycles" is one of our basic policies for task selection and it is expected to effective to initiate a new task for the first, and enhance the researches in the second.

As shown in Figure 2, *QAC* started from NTCIR-3, and investigated factoid for two cycles, Evaluation of *Series of Questions* was started from NTCIR-3 but the design was drastically modified into the evaluation of *Information Access Dialog (IAD)* in NTCIR-4 and NTCIR-5. Cross-language started in NTCIR-5. In NTCIR-6, it started a new QA experimental design to cover every kinds of questions including complex questions like definition, how and why as well as factoid, and evaluate using a metric called *BE*, Basic Element, which was originally proposed as which will be used in n automatic evaluation of summarization.

OPINION was proposed by Hsin-Hsin Chen and Chin-Yew Lin at NTCIR-5 workshop, then initiated as a pilot task utilizing the multilingual comparable corpora constructed through past *CLIR* tasks at NTCIR-3 to -5, in which 110 topics were translated into four languages and each topic accompanying the sets of documents in each of the four languages in which each document was manually judged relevance. The document sets for 30 selected topics were used, and added opinion related tags upon them. The task focused on identification but the collection is usable for Cross-lingual opinion retrieval together with CLIR judgments. No participants were registered for OPINION's Application Oriented Subtask..

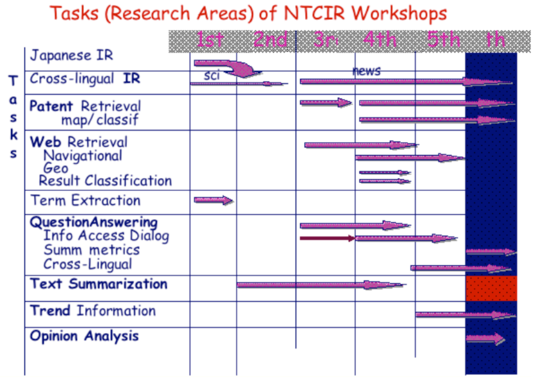

Figure 2. Tasks at *NTCIR Workshops*

2.2 Participants

Table 2 is a list of the active participating research groups in the *NTCIR-6*. A hundred and four groups from 15 different countries and areas registered, and eighty-five from twelve different countries and areas were remained as active participants to the final meeting.

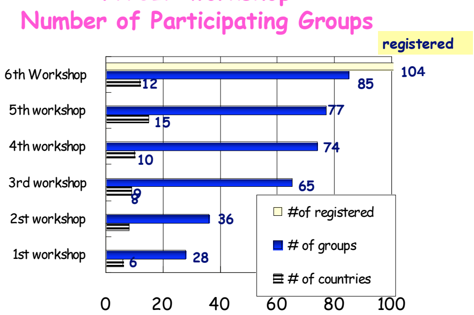

Figure 3. Number of Participating Groups

[1] Question Answering Challenge and Patent Retrieval tasks are partially supported by NII Joint Research Grant (Type-B) as well.

[2] *MuST* is supported by a Joint Research Grant of Nippon Telegraph and Telephone, Corp. and the University of Tokyo, and NII Joint Research Grant (Type-B).

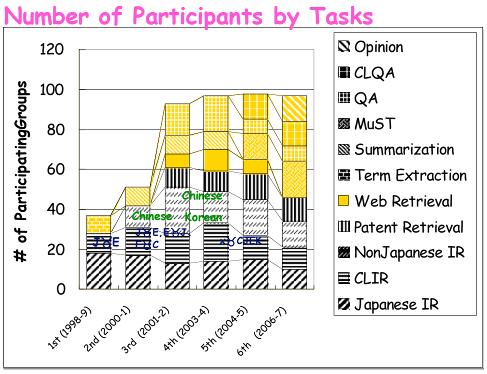

Figure 4. Number of Active Participating Groups, by Task

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, CLQA and OPINION*. The Retrieval subtasks at *PATENT* 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. And *OPINION* and *MuST* also contributed to enlarge the NTCIR community.

WEB was not conducted at NTCIR-6. It was not because of no interest. In opposite, we are interested in evaluating information access on WEB in various aspects. WEB has been incredibly changed in many ways both in quality and quantity of the information access in the real world, and technologies required to support htem. Not only evaluation of retrieval mechanisms as an abstraction of the real world, but also it has been attracted various interests on evaluating the technologies in the real world setting. Such tendencies include, for example, credibility, quality, balance or skewedness of the retrieval results, interaction and exploratory, social networking and connectivities, and users satisfaction with various tasks, etc. We then had one cycle break for better planning for the future tasks investigating various aspect on the WEB as well as those with other traditional document genres for various usage and information seeking tasks behind.

There are some large-scale research funding programs on the research on WEB under multiple government agencies in Japan, and each of them proposes to provide to the granted research groups with large-scale shared infrastructures of research and experiments, such as computing facilities like huge PC clusters or large-scale document collections. But none of them provides Evaluation infrastructures. Collaboration with these programs shall be expected to be mutually beneficial.

Table 2. Active Participating Groups of the *Sixth* NTCIR Workshop

[CLIR]
Academia Sinica
Chinese Academy of Sciences (ISCAS)
Huazhong Normal Univ
Hummingbird
Institute for Infocomm Research
Justsystem Corporation
National Central Univ
NICT
National Taiwan Normal Univ
Newswatch , Co.
Osaka Kyoiku Univy
POSTECH
Queens College
Queensland Univ of Technology
Toshiba
Univ of Aizu
Univ of California; Berkeley
Univ of Montreal
Univ of Neuchatel
Univ of Nottingham
Yahoo! Japan

[CLQA]
Aoyama Gakuin Univ
Carnegie Mellon Univ
Chinese Academy of Sciences (ICT)
Academia Sinica
Mount Holyoke College
National Central Univ
National Cheng Kung Univ
Queens College
State Univ of New York at Albany
Tokyo Institute of Technology (Furui)
Toyohashi Univ of Technology (Akiba)
Yokohama National Univ

[MuST]

Hiroshima City Univ
Justsystem Corporation
Keio Univ (saito)
Mie Univ
NICT
NEC (Internet Systems Research Labs)
Ochanomizu Univ (2 groups)
Okayama Univ
Osaka Prefecture Univ (3 groups)
Ritsumeikan Univ
Tokyo Denki Univ
Tokyo Institute of Technology
"Tokyo Metropolitan Univ"
Univ of Tokyo (kato)
Yokohama National Univ

[OPINION]
Cornell Univ
Illinois Institute of Technology
Information and Communications Univ
Chinese Academy of Sciences (ISCAS)
National Chiao Tung Univ
National Institute of Informatics
NICT
NEC (Internet Systems Research Labs)
Chinese Univ of Hong Kong
Toyohashi Univ of Technology (seki)
Univ of Maryland
Univ of Sheffield

[PATENT]

Hiroshima City Univ
Hitachi; Ltd
Justsystem Corporation
Nagaoka Univ of Technology
NICT
National Taiwan Normal Univ
NTT DATA
NTT-CS
POSTECH
Toyohashi Univ of Technology (aono)
Univ of Sheffield
Univ of Tsukuba

[QAC]
Aoyama Gakuin Univ
Carnegie Mellon Univ
Hokkaido University (araki)
Chinese Academy of Sciences (ISCAS)
NTT-CS
Ritsumeikan Univ
Toyohashi Univ of Technology (akiba)
Yokohama National Univ

Table 3. Test collections constructed by *NTCIR*

Class	Collection	Task	Documents						Task data		
									Topic/ Question		Relevance judge
			Genre	Filename	Lang.	Year	# of docs	Size	Lang.	#	
CLIR on Sciemtific	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	-		-	-
	NTCIR-2	IR	Sci. abstract	ntc2-j ++	J	1986-1999**	400,248	600MB	JE	49	4 grades
				ntc2-e ++	E		134,978	200MB			
CLIR on News	CIRB010	IR	News	CIRB010	C_t	1998-1999	132,173	132MB	C_tE	50	4 grades
	NTCIR-3 CLIR	IR	News	KEIB010	K	1994	66,146	74MB	CKJE	30	4 grades
			News	CIRB011	C_t	1998-1999	132,173	870MB	CKJE	50	4 grades
				CIRB020 ++			249,508				
				Mainichi +&	J		220,078				
				EIRB010	E		10,204				
				Mainichi Daily++			12,723				
	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-5 CLIR	IR	News	CIRB040r ++	Ct	2000-2001	901,446		CtKJ E	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				
				DailyYomiuri +&			17,741				
				Xinhua +&			198,624				
	NTCIR-6 CLIR	IR	News	CIRB040r ++	Ct	2000-2001	901,446		CtKJ E	50 (selected from NTCIR-3,4)	4 grades
				Hankookilbo ++	K		85,250				
				Chosenilbo ++			135,124				
				Mainichi +&	J		199,681				
				Yomiuri +&			658,719				
CLQA	NTCIR-5 CLQA	QA	News	CIRB040r ++	C	2000-2001	901,446		CJE	smpl:300 , test:200 *6	3 grades *7
				Yomiuri +&	J		658,719				
	NTCIR-6 CLQA	QA	News	DailyYomiuri +&	E	1998-1999	17,741		CJE	J-E/J-J/E-J: 200, C-E/C-C/E-C/E-E: 150	3 grades *7
				CIRB020 ++	Ct		249,203				
				Mainichi +&	J		220,078				
				EIRB010	E		10,204				
				Mainichi Daily++			12,723				
				Korea Times ++			19,599				
				Hong Kong Standard ++			96,683				
OPINION	NTCIR-6 OPINION	IE/analysis	News	CIRB020 ++	Ct	1998-2001	249,203	32, 30, 28 doc sets for C, J, E, respectively, were selected from those describe	CtKJ E	32 (selected from NTCIR-3,-4,-5 CLIR)	2 types, 3 metrics
				CIRB040r ++			901,446				
				Mainichi +&	J		419,759				
				Yomiuri +&			1,032,277				
				EIRB010	E		10,204				
				Mainichi Daily++			242,878				
				Korea Times ++			50,129				
				Hong Kong Standard ++			96,683				
				Xinhua +&			409,971				
Patent	NTCIR-3 PATENT	IR	Patent full	kkh *3 ++	J	1998-1999	697,262	18GB	CC_s K JE	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-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-5 PATENT	IR	patent full	Publication of unexamined patent application++	J	1993-2002	3,496,252	ca.45GB	JE	34+1189 in NTCR-5, added 349+1681in NTCR-6	3 grades
			Abstract	Patent Abstracts of Japan (PAJ) ++	E	1993-2002	3,496,252	ca.10GB			
	NTCIR-6 PATENT	IR	patent full	Patent grant data published from USPTO++	E	1993-2000	981,948		E	3221	3 grades
QA	NTCIR-3 QA	QA	News	Mainichi +&	J	1998-1999	220,078	260MB	J*	1200	exact answer
	NTCIR-4 QA	QA	News	Mainichi +&	J	1998-1999	220,078	ca.776MB	J*	197	exact answer
				Yomiuri +&			373,558			199	
										251	
	NTCIR-5 QA	QA	News	Mainichi +&	J	2000-2001	199,681	260MB	J*	50 series (360 Q)	graded
	NTCIR-6 QA	QA	News	Mainichi +&	J	1998-2001	419,759	535MB	J	100 Q (anykind of Q)	graded (3 types, 4 levels)
WEB	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 WEB	IR	Web (html/text)	NW100G-01++	multiple*4	crawled in 2001	11,038,720	100GB	J*		3 grades
	NTCIR-5 WEB	IR	Web (html/text)	NW1000G-04++	multiple*4	crawled in 2004	98,870,352	1.36TB	J*	269 + 847	3 grades
OTHERs	available for future task		QA site on Web	Yahoo! Q&A corpus (Chiebukuro) ++	J	Apr.2004 to Oct.2005					
			News	Singpore Press ++	Cs	1998-2001					

J:Japanese, E:English, C:Chinese (C_t:Traditional Chinese, C_s: Simplified Chinese), K:Korean;
All the topics/questions and relevance judgements/answers are available for research purpose for free

"++" indicates the document collections available from NII for research purpose
"+&" indicates the document collections available for task participants for free, and available
for research purpose use from research purpose other than NTCIR participation from other partywith fee

* 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

3. Test Collections

3.1 Documents

Table 3 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.

In NTCIR-6, most of the tasks except PATENT use the document collections in the some tasks in the past NTCIRs. Fulltexts of patent grant data from the United States Patent and Trades Office (USPTO) were newly added and used in the PATENT. The task participants were provided all the document data usable for evaluation for free of charge. Some of the documents are available for task participants only but the number of the documents collections usable for research purpose by non-participants are increasing.

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 this type of document in their everyday life.

3.2 Topics and Questions

The structure of the topic in the IR test collections is similar to that used in TREC [5] and CLEF [6]. These topics are defined as natural language statements of "users' requests" rather than "queries", strings submitted to the system, so that both manual and automatic query construction can be done. In NTCIR, *Mandatory Runs* are defined for each IR-related task, and every participant must submit at least one mandatory run using the specified topic field only. The purpose of this is to enhance cross-system comparisons by basing them on common conditions, and to judge the effectiveness of the additional information.

3.3 Relevance Judgments and Evaluation

In IR-related tasks, relevance judgments were graded using a scale similar to previous *NTCIR* workshops: highly relevant, relevant, partially relevant and irrelevant. The metrics suitable for such graded relevance judgments were proposed and tested in the past NTCIR submissions. In NTCIR-6, some of the such metrics were officially used in CLIR.

For Question answering, QAC at NTCIR-5 introduced a graded scoring for answer sets with Information Access Dialog (IAD) evaluation. Not only the experimental design simulated interaction like IAD, such graded scoring of the answers are natural in assessing the correctness of the answers. With increasing the complexity of the answers in QAC at NTCIR-6, QAC set four levels in answer correctness.

The unit of the judgments is also a issue: For example, whether document- or passage-levels in information retrieval, or whether document- or sentence- levels in OPINION.

3.4 Research Purpose Use by Non-Participating Researchers

One of the major importance of the test collections as an infrastructure of the research and testing of information access technologies is reusability and the fact that they exist in "*ready-to-use*" status for experiment.

In NTCIR, all the data constructed through NTCIR activities, i.e. topics / questions and relevance judgment / answers, and other annotations are available for research purpose by the researchers who did not participated in the NTCIR's tasks.

With continuous efforts of various individuals including local organizers and kind understanding and cooperation from related parties, the number of document collections which are usable for research purpose by non-participating researchers is gradually increased. That's effort resulted in the acute increase of the number of research purpose uses of the test collection in 2006 – 502 new applications were received and approved in the fiscal year for the research purpose usage of the past NTCIR test collections.

4. 4. Tasks

4.1 Cross-Lingual Information Retrieval **[7]**

CLIR consists of Stages 1 and 2, and each of them had three subtasks as below, and then evaluated using the metrics for graded relevance judgments:

- *Multilingual CLIR* (MLIR),
- *Bilingual CLIR* (BLIR), and
- *Single Language IR* (SLIR).

Stage one conducted an ad-hoc retrieval evaluation, but used the NTCIR-5 CLIR document collections except English, which were published in the years of 2000-2001, and use the topics selected from NTCIR-3 and -4, in which the runs were done against the documents published in 1998-1999. The past topics and documents were reused, but we used them as unseen one and not as a routing task.

In stage 2, each participant requested to submit the runs on the past NTCIR CLIR collections of NTCIR-3, -4, and -5. The purposes of the Stage 2 are 1) to see the technological improvement from the time of NTCIR-3, -4, and -5 by using the current state-of-the-art retrieval mechanisms on the same test collections, then see the improvement in each participating group if they participated in the past, and 2) to evaluate the search effectiveness across multiple collections to obtain further stability and reliability of the test.

The structure of subtasks and languages are basically the same as those in *NTCIR-4 and -5 CLI*R except English. We used the metrics for graded relevance judgments like normalized DCG [8], Q-Measure [9], generalized average precision [10] in addition to the metrics available by TREC_Eval package like mean average precision (MAP) based on the rigid and relaxed relevance judgments (AP_g, AP_x, respectively).

The major findings are:

- Segmentation on Chinese and Korean: In past NTCIR, n-gram based indexing was well used for Chinese and Korean while morphological analysers were well used for Japanese. Different segmentation strategies including semi-supervised learning were tested in NTCIR-6.
- For translation,
  - Statistical language model based translation by RALI [11]
  - Tackled the OOV problem using WEB or wikipedia in various ways.
  - Cognate matching on C-J
  - Pivot approach using English as pivot
- Retrieval model
  - Investigated combination strategies of different models and fusions
  - Optimal parameter estimation
- Enhancement and further analysis of Psuedo Relevant Judgments
- Document re-ranking was proposed by I2R in NTCIR-4 and known to be effective. Enhanced and variant methods were tested in NTCIR-6

4.2 Cross-Lingual Question Answering **[12]**

This is the second cycle of CLQA, Cross-Language Question Answering on E-J and E-C including monolingual. This task focused on factoid QA while QAC targeting various types of Questions including complex questions like definition, why and how.

- Subtasks: J-E, J-J E-J, and
- Subtasks: C-E, C-C, E-C, E-E

To use the NTCIR-5 Collection as training, the documents from different years were used in the test. Judgments were done on the pairs of [Answer, DOCNO] and evaluated as Right, correct but Unsupported, Wrong using the metrics of Accuracy of Top 1 answers, Mean Reciprocal Rank (MMR), and Top5.

We believed that the factoid questions are still important as many search requests are related to named entities, and unsolved problems especially on the language pairs using different alphabets like Chinese or Japanese and English.

As results, the effectiveness of each systems seemed improved but the module by module analysis showed that there are still problems in retrieval and translations. That suggested that further experiments with various top ranked CLIR systems shall be interesting. “Does the best CLIR system work better for CLQA?” or “Does the best CLIR system provides best retrieval document sets for any CLQA systems?” or so. So far the effectiveness of the Information Retrieval has been evaluated as intrinsic evaluation, but it can be possible to consider an extrinsic evaluation in which we will analyse the effectiveness of the retrieval in terms of the performance for the external task upon the retrieval such as question answering.

Though the problem of factoid QA has not completely solved, as a further extension of the investigation, Complex Cross-Language QA is proposed for the next cycle of NTCIR.

4.3 Multimodal Summarization for Trend Information **[13]**

Multimodal Summarization for Trend Information (MuST) 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 in 2005 and eighteen in 2006..This is an interesting mixture of different communities, IR, NLP, Web intelligence, Fuzzy, etc. The results was presented in the separate workshop held in March, 2006 and 2007, and selected papers will be presented at the NTCIR-6 Meeting.

The major achievements are, that we had established an infrastructure for the investigation of the technologies for trend analysis. It includes, Community of researcher, a set of tags, a set of tagged corpus usable for the experiments, a platform system for visualization. The platform software for visualization has been prepared as a software can be used by researchers and is expected to serve as one of the common basis for further evaluation of the trend analysis including usability or user-involved evaluation for exploratory.

MuST is proposed as an ordinary task at main NTCIR activities for the next cycle.

4.4 Opinion Analysis Pilot Task **[14]**

The roadmap for the task is shown in Figure 5. Though it was originally proposed to use Chinese monolingual corpus, but extended into multilingual ones. It means that we started from very complicated task from the beginning.

This task used past NTCIR-CLIR multilingual comparable corpus and added annotation related to opinion identification. Therefore the test collections are usable for automatic identification of Opinionated sentences and documents, as well as the task of the cross-language opinion retrieval on particular topics in conjunction with the CLIR relevance judgements.

As it was a pilot task, just delivering annotated sample documents were planned for the first. But later, many participants requested a kind of training data as most of them worked on machine learning based strategies, then later, some additional annotated corpus were delivered as a training set. In such a way, there were some unexpected changes in planning and schedule, but most of the participants overcame such difficulties and submitted results. Three metrics, LWK, DEK, and YS were used across all dataset.

We welcome any further discussion, leads, and advices for the further can not be delivered

Multimodal Summarization for Trend Information (MuST) 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 in 2005 and eighteen in 2006..This is an interesting mixture of different communities, IR, NLP, Web intelligence, Fuzzy, etc. The results was presented in the separate workshop held in March, 2006 and 2007, and selected papers will be presented at the NTCIR-6 Meeting.

4.5 Patent Retrieval **[15][16]**

This is the fourth attempt on Patent in the series of the NTCIR Workshop. It consists of three subtasks

- Japanese Retrieval Subtask
- English Retrieval Subtask, and
- Classification Subtask.

One of the major challenges for this Patent task is investigating the possibility of the evaluation without human judgments. Document Retrieval tested the invalidity search as in NTCIR-4 and -5 was performed, but the numbers of search topics were increased from manually judges 34 topics in NTCIR-4, additional automatic judgments using Patent examiners citation for 1189 topics in NTCIR-5.

The basic idea here was, the evaluation stability, reliability and sensitivity are related to the number of topics (or more precisely to the number of evaluation points where calculating the evaluation scores). However, the relevance judgments are very labour-intensive, time consuming and expensive tasks especially for the documents genres requires high professional assessors like patents. At the same time, we can find a kind of quasi-relevant judgments data like patent Office examiners' citations. Then we have investigated the methods and scope of the reliability of the using large number of topics with less-intensive judgments. In NTCIR-5, We found that Examiners' judgments were rather precision oriented, and they are generally usable for evaluation but there were possibility to fail to find the best systems.

In NTCIR-6, we then added extra topics which utilize the two different types of existing quasi-judgments; 1) 349 topics with recall oriented judgments selected from the "search reports" serving as a search before the examination to find slightly larger number of related patents by out-sourced search intermediaries, and 2) 1681 new topics using Examiners citations. Participants ran against all the available test collection on patents for the invalidity task and compared the results across multiple test collections with different type of judgments to search for the more reliable evaluations and for the evaluation with less-labor intensive judgments.

The results showed that such a kind of quasi-judgments can be usable to depict the differences among systems, but not fully usable to differentiate the performance between the different mechanisms on the same system.

Classifying patent applications has promise to improve the quality of the patent map generation task, which was tested in NTCIR-4 as a feasibility task. Additionally, the document classification can automatically be evaluated using the patent classification system. In our case, a multidimensional classification system called "F-term (File Forming Term)" was used.

4.6 Question Answering Challenge **[17]**

QAC at this NTCIR-6 started a new evaluation design to cover every types of questions including complex questions like definition, how and why as well as factoids and evaluated using a metrics called Basic Element (BE), which was originally proposed by University of Southern California ISI group to evaluate summarization of English, and Japanese version was prepared by Junichi Fukumoto with a collaboration with ISI.

Many issues relating to further evaluation were found and suggested for further discussion.

5. Discussion

A brief overview of the *Sixth 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 this volume [4].

The test collections used in the tasks of the NTCIR-6 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.

References

1. NTCIR Project: http://research.nii.ac.jp/ntcir/
2. NTCIR Workshop 6 (2006-2007) : http://research.nii.ac.jp/ntcir/ntcir-ws6/meeting/
3. NTCIR Workshop 6 Meeting (15-18 May 2007) http://research.nii.ac.jp/ntcir/ntcir-ws6/meeting/index.html
4. Noriko Kando, David K. Evans, (eds): *NTCIR Workshop 6: Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo Japan, May 15-18, 2007, NII, Tokyo 2007. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/
5. TREC: http://trec.nist.gov/
6. CLEF: http://clef.iei.pi.cnr.it/
7. Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen: Overview of CLIR Task at the Sixth NTCIR Workshop. In *Proceedings of the Sixth NTCIR Workshop Meeting*, Tokyo Japan, May 15-18, 2007, NII, Tokyo 2007
8. Kalervo Javelin, Jaana Kekalaine. IR evaluation methods for retrieving hightly relevant documents. In Proceedings of ACM-SIGIR 2000, pp. 41-48.
9. Tetsuya Sakai, M. Koyama, A. Kumano. Toshiba BRIDGE at NTCIR-4 CLIR: Monolingual / bilingual IR and flexible feedback. In Proceedings of NTCIR-4 Workshp, 2004.
10. Kazuaki Kishida. Property of Average Precision and Its Gerenalization: An Examination of Evaluation Indicator for Information Retrieval. NII Technical Reports, NII-2005-014E, 2005. http://research.nii.ac.jp/TechReports/05-014E.html
11. Lixin Shi, Jian Yun-Nie. Using unigram and bigram language models for monolingual and cross-language IR. In Proceedings of NTCIR-6 Workshop, 2007.
12. Yutaka Sasaki, Chuan-Jie Lin, Kuang-hua Chen, Hsin-Hsi Chen.: Overview of the NTCIR-6 Cross-Lingual Question Answering (CLQA) Task. In *Proceedings of the Sixth NTCIR Workshop Meeting*, Tokyo Japan, May 15-18, 2007, NII, Tokyo 2007.
13. Tsuneaki Kato, Mitsunori Matsushita, Noriko Kando: Expansin of Multimodal Summarization for Trend Information – Report on the First and Second Cycles of the MuST Workshop. In *Proceedings of the Sixth NTCIR Workshop Meeting*, Tokyo Japan, May 15-18, 2007, NII, Tokyo 2007
14. Yohei Seki, David Kirk Evans, Lun-Wei Ku, Hsin-His Chen, Noriko Kando, Chin-Yew Lin.: Overview of Opeinion Analysis Pilot Task at the NTCIR-6, In *Proceedings of the Sixth NTCIR Workshop Meeting*, Tokyo Japan, May 15-18, 2007, NII, Tokyo 2007
15. Atsushi Fujii, Makoto Iwayama, Noriko Kando: Overview of Patent Retrieval Task at NTCIR-6. In *Proceedings of the Sixth NTCIR Workshop Meeting*, Tokyo Japan, May 15-18, 2007, NII, Tokyo 2007.
16. Makoto Iwayama, Atsushi Fujii and Noriko Kando. Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task
In *Proceedings of the Sixth NTCIR Workshop Meeting*, Tokyo Japan, May 15-18, 2007, NII, Tokyo 2007.
17. Jun'ichi Fukumoto, Tsuneaki Kato, Fumito Masui, Tatsunori Mori.: An Overview of the 4th Question Answerig Challenge (QAC-4) at NTCIR Workshop 6, In *Proceedings of the Sixth NTCIR Workshop Meeting*, Tokyo Japan, May 15-18, 2007, NII, Tokyo 2007.

.