Overview of the Fourth NTCIR Workshop

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

Abstract

This paper outlines the fourth 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 fourth NTCIR Workshop.

Keywords: evaluation, information access, information retrieval, text summarization, question answering, test collections, cross-lingual information retrieval, patent retrieval, Web retrieval.

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), automatic text summarization, question answering, 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.

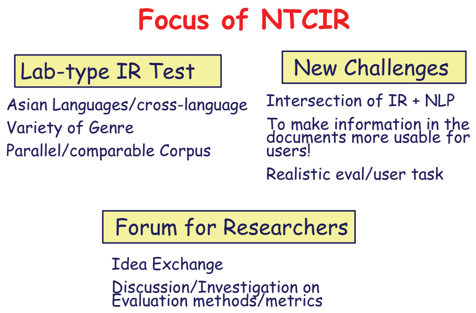

Focus of NTCIR
Lab-type IR Test
Asian Languages/cross-language
Variety of Genre
Parallel/comparable Corpus
New Challenges
Intersection of IR + NLP
To make information in the documents more usable for users!
Realistic eval/user task
Forum for Researchers
Idea Exchange
Discussion/Investigation on Evaluation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

Figure 1. Focus of *NTCIR Workshops*

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	NTCIR-4 WEB
			Navigational Retrieval	
			[Pilot] Geographical Information	
			[Pilot] (Search Results) Topical Classification	

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. These two directions have been supported by a forum of researchers who are interested in cross-system comparison and by their discussions.

For the *Fourth NTCIR Workshop* (NTCIR-4) [2], the process started from April 2003 and the meeting was held on 2-4 June 2004 [3], at National Institute of Informatics (NII) in Tokyo. It is co-sponsored by the NII and Japan’s MEXT Grant-in-Aid for Scientific Research on Informatics (#13224087)[1]and Research Center of

[1] “Digital Contents Research Group A02 of Japanese MEXT Grant-in-Aid for Scientific Research on Informatics (#13224087)”, Primary Investigator: Jun Adachi.

Information Resources at National Institute of Informatics (RCIR/NII). Question Answering Challenge's Subtask 3 was supported by NII Collaborative Research Grant Type B.

The Patent Retrieval task was organized in cooperation with the Japan Intellectual Property Association (JIPA) and NII, and the *CLIR* task was organized in cooperation with the National Taiwan University and the Korean Institute for Scientific and Technological Information (KISTI).

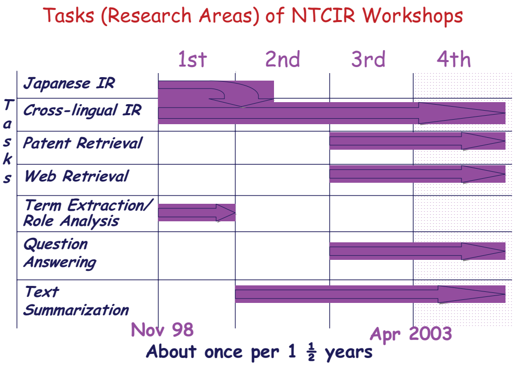
2. The Fourth NTCIR Workshop

2.1 Tasks

The *NTCIR-4* selected five areas of research as "tasks":

1. Cross-Lingual Information Retrieval Task (*CLIR*),
2. Patent Retrieval Task (*PATENT*),
3. Question Answering Challenge (*QAC*),
4. Text Summarization Challenge (*TSC*), and
5. WEB Task (*WEB*).

Since WEB was organized within somehow different management and run by its own schedule, this overview includes mainly CLIR, PATENT, QAC, and TSC.



Figirue 2. Tasks at *NTCIR Workshops*

As shown in **Table 1** and **Figure 2**, at the NTCIR-4, all of the tasks were some kind of continuation or enhancement from the previous NTCIR. Each of them increased the size of the test collections. *PATENT* proposed experiments within the different information seeking task of "invalidity search" task and challenging topic of "automatic patent map generation" as a feasibility task of a long-term research project which will last for two consecutive NTCIRs, it means for three years until NTCIR-5.

2.2 NTCIR-4 CL IR

As this is the second multilingual CLIR at NTCIR, the task design will be a continuation of the previous one. Only minor revisions were made to solve the major problems raised in the assessment on the NTCIR-3, as follows:

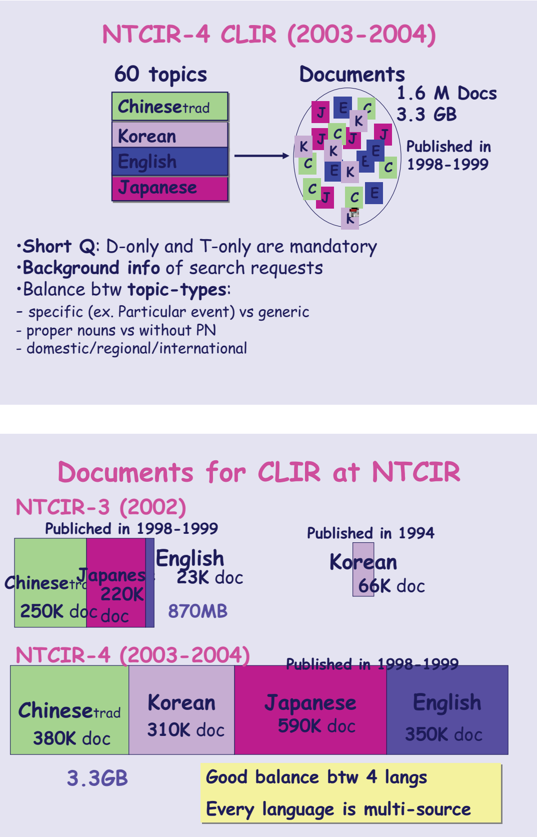
- Enlarge the English and Korean document collections to be comparable to the Chinese and Japanese ones, about 3 GB in total.
- A new sub-task of Pivot Language Bilingual CLIR
- Restrict the pairings of topic and document languages, so that comparisons will be more fruitful
- Make a T-only run mandatory as well as a D-only run
- Question-type topics were categorized according to the nature and types of the answers to obtain a good balance of topic sets

For CLIR, for every language, documents were collected from multiple sources of the same publication years in somewhere in East Asia and the collection size balance between different languages was much improved by increasing the document collection size as shown in Figure 3. The same Japanese documents were used in QAC and TSC.

The number of the topics was also increased from 50 to 60, so that the test collection can have sufficient number of topics (at least more than 50) after discarded some of the topics which did not have sufficient number of relevant topics on a particular language documents. The number of topics is one of the critical elements to make the evaluation reliable, stable and sensitive.

The new sub-task, pivot CLIR, uses English as a pivot language, then tests the effectiveness of the transitive CLIR. It is a practical approach to Multilingual CLIR in environments with less availability of direct translation resources but rich in those between each of the languages and English.

On the consideration of the real-world setting of Web search and other retrieval services, seach by very short queries is more practical and realistic, then we added TITLE-only run to the mandatory run which keeping DESC-only run as mandatory to make the comparison with previous results as reference to show the level of difficulty of the collection.



NTCIR-4 CLIR (2003-2004)

60 topics

Chinesetrad
Korean
English
Japanese

Documents

1.6 M Docs
3.3 GB

Published in 1998-1999

- **Short Q**: D-only and T-only are mandatory
- **Background info** of search requests
- Balance btw **topic-types**:
  - specific (ex. Particular event) vs generic
  - proper nouns vs without PN
  - domestic/regional/international

Documents for CLIR at NTCIR

NTCIR-3 (2002)

Published in 1998-1999

Chinesetrad 250K doc
Japanese 220K doc
English 23K doc
870MB

Published in 1994

Korean 66K doc

NTCIR-4 (2003-2004)

Published in 1998-1999

Chinesetrad 380K doc
Korean 310K doc
Japanese 590K doc
English 350K doc

3.3GB

Good balance btw 4 langs
Every language is multi-source

Figure 3. NTCIR-4 Test Collection for CLIR and Comparison of the Document Size with NTCIR-3

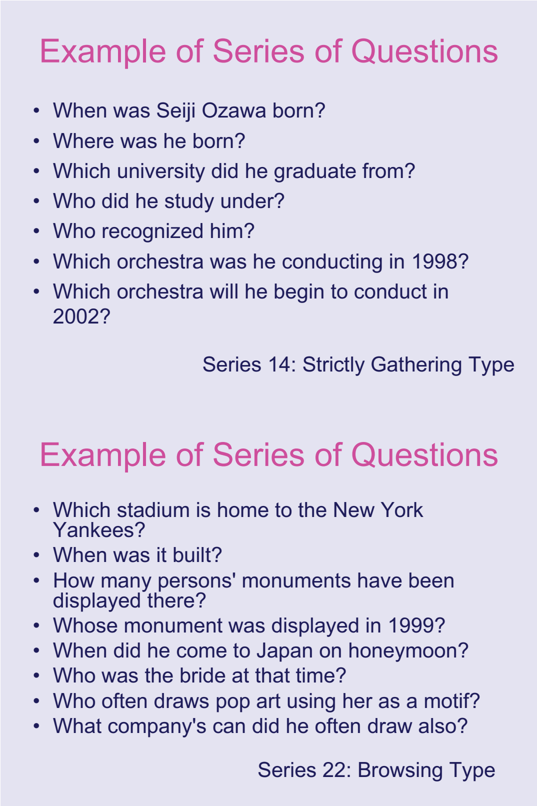
2.3 NTCIR-4 Question Answering (QAC) and Text Summarization (TSC)

QAC plans three subtasks as at NTCIR-3. Of the three, subtasks 1 and 2 will be done without major change. The only exceptions are the use of different question sets for subtasks 1 and 2, and an increase in the number of topics containing multiple answers and the number of answers in subtask 2 which tested the list questions. It was decided to avoid overestimates of the groups by ignoring the possibility of multiple answers and returning only the first priority answer to every question in subtask 2.

QAC subtask 3, answering a series of questions, is one of the major focuses of the *NTCIR-4 QAC*. The number of sequences was increased from the previous *QAC*, and the task design aimed at tackling the problems resembling the real-world interaction with users and information retrieval systems for information gathering for "Report Writing" on a topic, which we call "*Information Access Dialog (IAD)*" task. For the first, each of the topic authors set a topic for the report, then set a series of specific questions relating to the topic. Each of the questions ask a particular aspect of the topic.

Therefore each of the series of questions has a broder topic to lead the interaction towards information collection for report writing. But usually such topics were understood by the users, but not for the systems. Then, in the test, the systems were just provided a series of questions without the topics of the series.

In the real-world of the interaction between users and IR systems, the users sometimes strictly keeps on the target topics and search subtopics directly related to the topics. But the users sometimes shift their focus in the middle of interaction by inspired by the information retrieved and browse towards related topics of the information obtained in the retrieved documents. To reflect such situations, the series of questions used in the IAD task were classified into three types, Strict Information Gathering, Relaxed Information Gathering, and Browsing. Samples are shown in Figure 6.



Example of Series of Questions

- When was Seiji Ozawa born?
- Where was he born?
- Which university did he graduate from?
- Who did he study under?
- Who recognized him?
- Which orchestra was he conducting in 1998?
- Which orchestra will he begin to conduct in 2002?

Series 14: Strictly Gathering Type

Example of Series of Questions

- Which stadium is home to the New York Yankees?
- When was it built?
- How many persons' monuments have been displayed there?
- Whose monument was displayed in 1999?
- When did he come to Japan on honeymoon?
- Who was the bride at that time?
- Who often draws pop art using her as a motif?
- What company's can did he often draw also?

Series 22: Browsing Type

Figure 4. Sample of "Series of Questions" used in Information Access Dialog (IAD) Subtask in QAC

This IAD task design is also related to the TSC, in which a topic and a set of documents on the topic were provided to automatic summarizers, they were requested to produce a summary based on the given document set. For the evaluation, the content of the system produced summaries were assessed by assessors using a set of questions relating to the topic -- whether the summary contained sufficient content to answer each of the questions or not.

It can be said that those sets of questions used in QAC and TSC represent the attributes or aspects of the given topic. Such "topic - subtopics" relationship was also looked at in the CLIR, in which the types of the topics were investigated.

TSC included automatic evaluation of summaries and building a re-usable test collection for summarization. *CLIR* and *QAC* basically continued with minor changes in task design to remedy the major problems found in the third workshop.

TSC and QAC used the same Japanese document collections used in CLIR. As the documents were collected from multiple sources, variance in expressions were naturally introduced and it made the summarization and question answering tasks more realistic and challenging than those tested using single source document collection in the past TSC and QAC.

2.4 Specialized Genre Related Tasks at NTCIR-4: Patent and WEB

Both PATENT and WEB tasks plan Main task(s) and Feasibility or Pilot studies for more challenging tasks as follows:

PATENT – Main: Invalidity Task:
To search patents to invalidate the query patents. The claims of the query patents are used as queries and they are segmented into components relating to the inventions or technologies comprising the investigation, then related patents are searched. A patent may be invalidated by one patent or by a combination of multiple patents. The search returns document IDs as well as relevant passages.

PATENT – Feasibility:
Long term research plan over NTCIR-4 and -5: Automatic Patent Map Creation, a kind of text mining – detect sets of technologies used in a set of patents, extract them, and make a table showing the relationship between technologies and patents, and evolution or trends among them.

WEB – Main: Informational Search and Navigation-Oriented Search, to find the most informative and reliable page.
WEB – Pilot: Geographically oriented and Topical classifications of the Search results.

2.5 Participants

Table 2 is a list of the active participating research groups in the *NTCIR-4*. A hundred and four groups registered, and seventy-four groups from ten different countries and areas submitted task results.

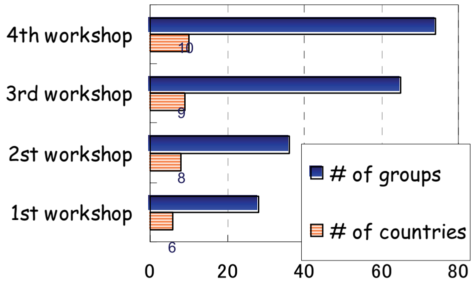


Figure 5. Number of Participating Groups

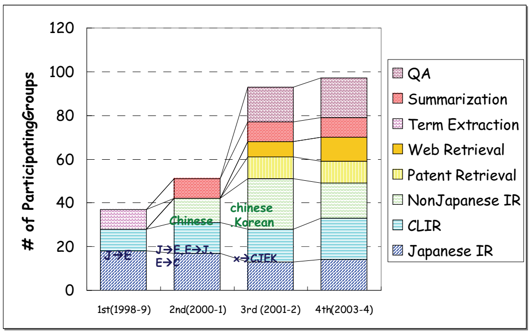


Figure 6. Number of Participating Groups, by Task

As shown in **Figures 5** and **6**, the number of participants has gradually increased. Different tasks attracted different research groups. Many international participants enrolled in *CLIR*. The *PATNET* task attracted participants from company research laboratories and "veteran" *NTCIR* participants. The *WEB* task had participants from various research communities such as machine learning and DBMS.

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. One of our interests is to prepare realistic evaluation infrastructures and efforts

Table 2. Active Participating Groups of the Third NTCIR Workshop

[CLIR]
Chinese Academy of Sciences (China PRC)
Clairvoyance Corporation and Justsystem (USA)
Communications Research Laboratory-1 (Japan)
Fu Jen Catholic University (Taiwan ROC)
Hong Kong Polytechnic University (Hong Kong, China PRC)
Hummingbird (Canada)
Institute of Inforcomm Research (Singapore)
Korea University (Korea)
Nara Institute of Science and Technology-1(Japan)
National Institute of Informatics-1 (Japan)
National Taiwan University (Taiwan ROC)
Oki Electric-1 (Japan)
PATOLIS (Japan)
Pohang University of Science and Technology (Korea)
Queens College City Univiersity of New York (USA)
Ricoh-1 (Japan)
Royal Melbourn Intitute of Technology (Australia)
Thomson Legal and Regulatory (USA)
Tianjin University (China PRC)
Toshiba (Japan)
University of Arizona (USA)
University of California Berkeley (USA)
University of Chicago (USA)
University of Neuchatel (Switzerland)
University of Tsukuba (Japan)
Yokohama National University (Japan)

[PATENT]
Fujitsu Laboratories (Japan)
IBM Research (Japan)
Japan Patent Information Organization / Hitachi (Japan)
Nagaoka University of Technology (Japan)
NTT DATA (Japan)
Osaka Kyoiku University (Japan)
PATOLIS (Japan)
Ricoh-2 (Japan)
Tokyo Institute of Technology (Japan)
University of Tsukuba (Japan)

74 groups from 10 countries & areas

[QAC]
AIST/University of Nagoya/Univeristy of Tsukuba (Japan)
Communications Research Laboratory-1 (Japan)
Iwate Prefectural University (Japan)
Keio University (Japan)
Matsushita Electoric Industiral-1 (Japan)
Mie University (Japan)
Nagaoka University of Technology (Japan)
Nara Institute of Science and Technology-2 (Japan)
New York University (USA)/Communication Research Lobaratory-2 (Japan)
NTT Communication Science Laboratories-1 (Japan)
NTT DATA (Japan)
Oki Electric-2(Japan)
Pohang University of Science and Technology (Korea)
Ritsumeikan University (Japan)
Toshiba (Japan)
Toyohashi University of Technology-1 (Japan)
University of Tokyo-1 (Japan)
Yokohama National University (Japan)

[TSC]
Communications Research Laboratory-2 (Japan) / New York University (USA)
Graduate University for Advanced Studies (Japan)
Hokkaido University (Japan)
Pohang University of Science and Technology (Korea)
Ritsumeikan University (Japan)
Toyohashi University of Technology-1 (Japan)
University of Electro-Communications (Japan)
University of Tokyo-1 (Japan)
Yokohama National University (Japan)

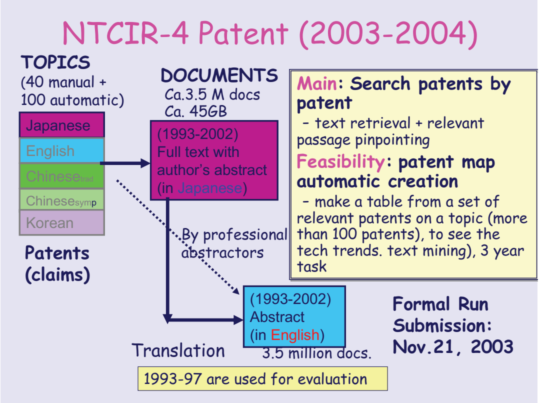
[WEB]
Hokkaido University (Japan)
Ibaraki University (Japan)
Matsushita Electoric Industiral-2 (Japan)
NEC (Japan)
NII-2/Univ. of Tokyo-2/KYA Group (Japan)
NTT Communication Science Laboratories-2 (Japan)
Osaka Kyoiku University (Japan)
Tokyo Metropolitan University (Japan)
Toyohashi University of Technology-1 (Japan)
Toyohashi University of Technology-2 (Japan)
University of Tsukuba/University of Nagoya

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.

Generally the size of documents collections were enlarged in NTCIR-4. For CLIR, we prepared the enlarged, well-balanced collections of Chinese, Korean, Japanese and English news article document collections -- the size of each language collection increased and consisted of the documents from multiple sources as shown in Figure 3.

The Patent document collection increased the size from 2 years of fulltext of patent applications to 10 years although only 5-years documents were used in NTCIR-4 as shown in Figure 7..

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 ev eryday life.



NTCIR-4 Patent (2003-2004)

TOPICS
(40 manual +
100 automatic)

Japanese
English
Chinesetrad
Chinesesymp
Korean

Patents
(claims)

DOCUMENTS
Ca.3.5 M docs
Ca. 45GB

(1993-2002)
Full text with
author's abstract
(in Japanese)

By professional
abstractors

Translation

(1993-2002)
Abstract
(in English)
3.5 million docs.

1993-97 are used for evaluation

Main: Search patents by patent
- text retrieval + relevant passage pinpointing

Feasibility: patent map automatic creation
- make a table from a set of relevant patents on a topic (more than 100 patents), to see the tech trends. text mining), 3 year task

Formal Run
Submission:
Nov.21, 2003

Figure 7. Patent Test Collection for NTCIR-4

3.2 Topics and Questions

The structure of the topic in the IR test collections is similar to that used in TREC [5] and

Table 3. Test collections constructed by *NTCIR*

NTCIR Test Collections; IR and QA

Collection	Task	Documents						Task data		
								Topic/		Relevance
		Genre	Filename	Lang.	Year	# of docs	Size	Lang.	#	judge
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 full	kkh *3	J	1998-1999	697,262	18GB	C_tC_sKJE	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-	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			96,683				
			Xinhua +			208,167				
NTCIR-4 PATENT	IR	patent full	Publication of unexamined patent application	J	1993-2002	ca. 3,500,000	ca.45GB	CtCsKJE	Main: 34, Add: 69	3 grades
		Abstract	Patent Abstracts of Japan (PAJ) +	E	1993-2002	ca. 3,500,000	ca.10GB			
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*		

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)

NTCIR Text Summarization

Collection	Task	Documents					Summaries		
		Genre	Filename	Lang	Year	# of doc	Types	Analysts	total#
NTCIR-2 SUMM	Single doc	News	Mainichi	J	1994.1995.1998	180 doc	7	3	3780
NTCIR-2 TAO	Single doc	News	Mainichi	J	1998	1000 doc	2	1	2000
NTCIR-3 SUMM	Single doc	News	Mainichi	J	1998-1999	60 docs	7	3	1260
	Multi doc		Mainichi	J		50 sets	2	3	300
NTCIR-4 SUMM	Multi doc	News	Mainichi Yomiuri	J	1998-1999	30 sets	2	1	60*

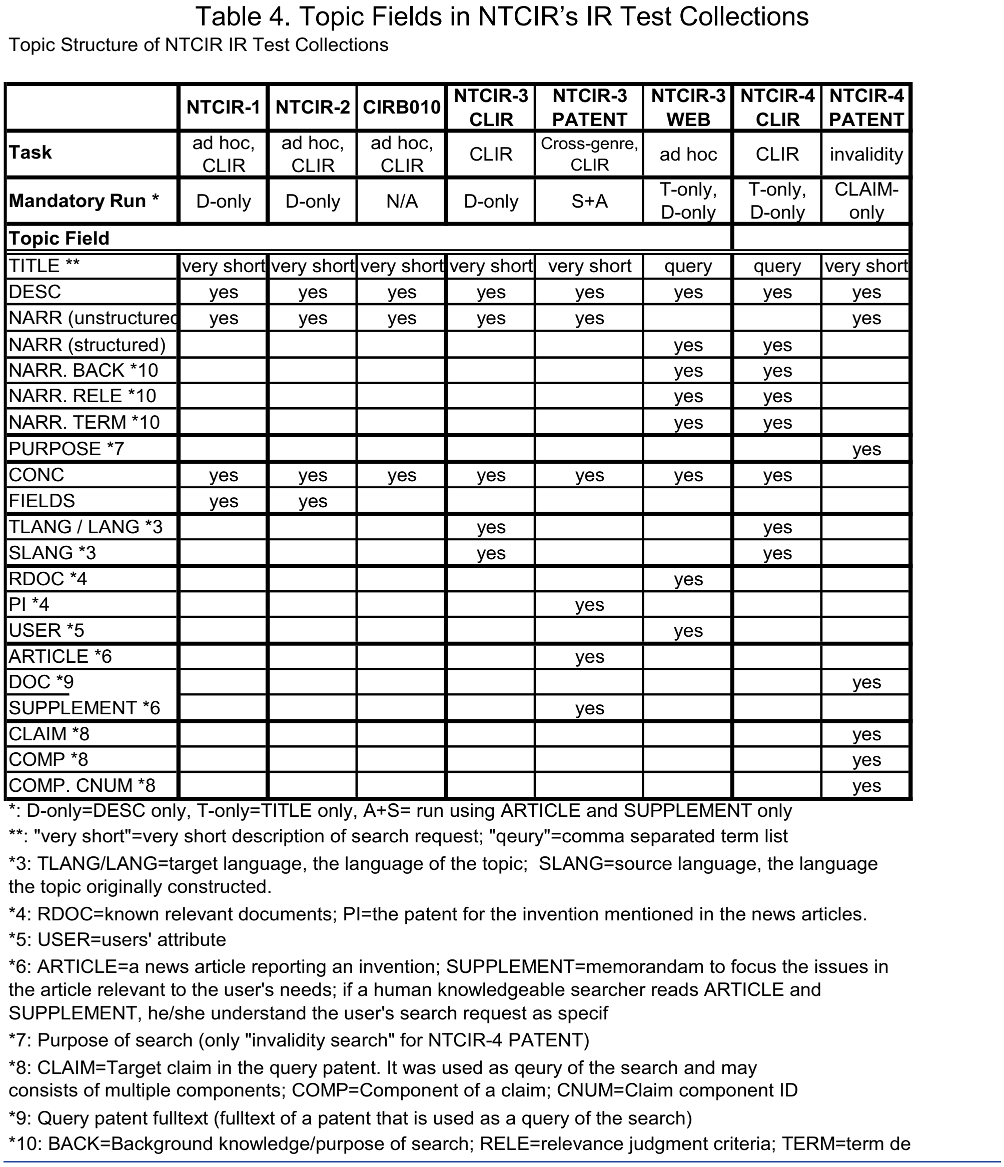


Table 4. Topic Fields in NTCIR's IR Test Collections

Topic Structure of NTCIR IR Test Collections

	NTCIR-1	**NTCIR-2**	**CIRB010**	**NTCIR-3 CLIR**	**NTCIR-3 PATENT**	**NTCIR-3 WEB**	**NTCIR-4 CLIR**	**NTCIR-4 PATENT**
Task	ad hoc, CLIR	ad hoc, CLIR	ad hoc, CLIR	CLIR	Cross-genre, CLIR	ad hoc	CLIR	invalidity
Mandatory Run *	D-only	D-only	N/A	D-only	S+A	T-only, D-only	T-only, D-only	CLAIM-only
Topic Field								
TITLE **	very short	very short	very short	very short	very short	query	query	very short
DESC	yes	yes	yes	yes	yes	yes	yes	yes
NARR (unstructured	yes	yes	yes	yes	yes			yes
NARR (structured)						yes	yes	
NARR. BACK *10						yes	yes	
NARR. RELE *10						yes	yes	
NARR. TERM *10						yes	yes	
PURPOSE *7								yes
CONC	yes	yes	yes	yes	yes	yes	yes	
FIELDS	yes	yes						
TLANG / LANG *3				yes			yes	
SLANG *3				yes			yes	
RDOC *4						yes		
PI *4					yes			
USER *5						yes		
ARTICLE *6					yes			
DOC *9								yes
SUPPLEMENT *6					yes			
CLAIM *8								yes
COMP *8								yes
COMP. CNUM *8								yes

*: D-only=DESC only, T-only=TITLE only, A+S= run using ARTICLE and SUPPLEMENT only

**: "very short"=very short description of search request; "qeury"=comma separated term list

*3: TLANG/LANG=target language, the language of the topic; SLANG=source language, the language the topic originally constructed.

*4: RDOC=known relevant documents; PI=the patent for the invention mentioned in the news articles.

*5: USER=users' attribute

*6: ARTICLE=a news article reporting an invention; SUPPLEMENT=memorandam to focus the issues in the article relevant to the user's needs; if a human knowledgeable searcher reads ARTICLE and SUPPLEMENT, he/she understand the user's search request as specif

*7: Purpose of search (only "invalidity search" for NTCIR-4 PATENT)

*8: CLAIM=Target claim in the query patent. It was used as qeury of the search and may consists of multiple components; COMP=Component of a claim; CNUM=Claim component ID

*9: Query patent fulltext (fulltext of a patent that is used as a query of the search)

*10: BACK=Background knowledge/purpose of search; RELE=relevance judgment criteria; TERM=term de

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. Mandatory runs were originally designated “<DESC> only” because <DESC> is the basic description of the users’ search requests, but from NTCIR-4, CLIR was designated both “<TITLE> only” and “<DESC> only”. It was partially because short queries like <TITLE> only runs are more realistic and partially because that to test the effectiveness of the disambiguation mechanisms, which is one of the critical components in CLIR, shorter queries is more preferable. Any combination of topic fields may be used in experiments for research purposes.

As shown in Table 4, emphasis has been shifted towards the topic structure to allow more realistic experiments and to gauge the effect of background information on the topic. For example, the narrative <NARR>, longer natural language explanation for each topic, can be structured using tags indicating subfields in <NARR>, such as <BACK> for “Background/Purpose of Search”, <RELE> as “Relevance Judgment Criteria”, or <TERM> for “Term Definition”. Most *NTCIR* collections contain a list of concepts <CONC>, but they are not heavily used by participants. The topics in the *PATENT* collections are various according to the information seeking tasks each of the tasks set upped.

For TSC, both the documents themselves and the topics of each of the document sets were given to the participants. These topics are very simple expression typically a few terms, but this can be seen as users’ initial search requests and the set of documents were produced as retrieval results for the requests.

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. For the *Patent Retrieval* task, professional patent intermediaries conducted judgments on the pooled documents consisting of the documents listed in the higher ranks in each submitted run, together with intensive interactive search and judgments using several commercial patent retrieval systems and the system provided by the task organizers. Such integration of the two different strategies to collect relevant documents was found to improve the completeness of the relevance judgments for a large-scale document collection with longer documents.

For the *QAC*, exact answers were used for evaluations. They were prepared before the runs by assessors, then also all the submitted answers were reviewed and revised answer sets were released after. For the evaluation, the mean reciprocal rank (MRR) is used for subtask 1, in which the participating systems were requested to return five possible answers with no penalty for wrong answers, and the modified mean F-measure is used for subtask 2, in which the participating systems returned one set of all the answers with penalties given for wrong answers. For subtask 3, a series of questions are used for either of in the user’s information seeking tasks of “information gathering” in which a user supposes to raise a series of questions on a particular topic, and “browsing”, in which users questions are keep drifting through the interaction with systems.

For *Text Summarization*, two types of summaries, short and long, were produced by analysts as gold standards, and then each sentence in those summaries are related to the sentences in the source documents to be summarized. The analysts asked to check all the possible relations between sentences in the human created summaries and the source documents. Using these greedily annotated human produced summaries, the effectiveness of system produced summaries can be automatically evaluated as an extract in the aspects of “number of sentences should be extracted”, “precision”, and “coverage” as the intrinsic evaluation of extraction. For intrinsic evaluation for abstract, content and readability were tested using a set of quality questions. For extrinsic evaluation of abstracts, system produced summaries were evaluated by question answering.

6. 4. Discussion

A brief overview of the *fourth 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].

To enhance the research in the fourth workshop, special attentions were paid (1) to provide longer time period for experiments, and (2) to enhance the document collections. It was because that, in the *NTCIR-3*, lots had to be done by the participants for the new tasks and new task components. As results, participants could only implement some of their research ideas, but generally such new task components had not been fully investigated nor analyzed because of tight schedule of the workshop.

Moreover, in the *NTCIR-3*, for some of the new components like “passage-level relevance judgments” *PATENT*, *QAC* and *WEB* tasks and “Search Results Classification” at the *WEB* task, none of the participants fully accomplished. Then, in the *NTCIR-4,* in order to obtain sufficient time to think of the task and original and unique idea for the experiment strategies, we released the document collection as early as possible and omitted the dry runs for the tasks in which the *NTCIR-3 collection* were usable for training. For much more challenging issue, we set the "*feasibility study*" subtask, in which the investigation is performed through the two consecutive workshops, i.e. for three years. In such ways, we expected that each participant could spend sufficient time for experimentation and implementation.

Technologies keep improving and evolving, and the society kept changing. Evaluation must adapt to technological evolution and the change in social needs. We are working towards this goal, and any suggestions and leads are always welcome.

References

1. NTCIR Project: http://research.nii.ac.jp/ntcir/
2. NTCIR Workshop 4 (2003-2004) : http://research.nii.ac.jp/ntcir-ws4/work-en.html
3. NTCIR Workshop 4 Meeting (2-4 June 2004) : http://research.nii.ac.jp/ntcir/ntcir-ws4/
4. Kando, N. and Ishikawa, H. (eds): NTCIR Workshop 4: Proceedings of the Fourth NTCIR Workshop Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization, Tokyo Japan, April 2003 - June 2004., NII, Tokyo. ISBN: 4-86049-030-4 (This volume) (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/)
5. TREC: http://trec.nist.gov/
6. CLEF: http://clef.iei.pi.cnr.it/