CLIR System Evaluation at NTCIR Workshops

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

Abstract:

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

1. Introduction

The *NTCIR*[1] *Workshop* [1] is a series of evaluation workshops, which is designed to enhance research in information retrieval and related text processing techniques, such as summarization, extraction, by providing large-scale test collections and a forum for researchers.

1.1 Purpose

The purposes of the NTCIR Workshop are the following:

1. to encourage research in information retrieval (IR), and related text processing technology, including term recognition and summarization, by providing large-scale reusable test collections and a common evaluation setting that allows cross-system comparisons;
2. to provide a forum for research groups interested in comparing results and exchanging ideas or opinions in an informal atmosphere;
3. to investigate methods for constructing test collections or data sets usable for experiments, and methods for laboratory-type testing of IR and related technology.

We call the whole process from the data distribution to the final meeting the *"Workshop"* since we have placed emphasis on the interaction among participants, and the experience gained as all participants learn each other from each other's experience.

[1] NTCIR: NII-NACSIS Test Collections for Information Retrieval and Text Processing

1.2 Brief History

The *First NTCIR Workshop* started with the distribution of the training data set on 1 November 1998, and ended with the workshop meeting, which was held on 30 August - 1 September 1999 in Tokyo, Japan [2]. Many interesting papers with various approaches were presented at the meeting. The third day of the meeting was organized as the *NTCIR/IREX Joint Workshop*. The *IREX Workshop* [3], another evaluation workshop of information retrieval and information extraction (named entities) using Japanese newspaper articles, was held consecutively. IREX and NTCIR joined in 2000 and have worked together to organize the NTCIR Workshop. The new tasks of *Text Summarization* and *Question Answering* became feasible with this collaboration.

The international collaboration to organize Asian languages IR evaluation was proposed at the *4th International workshop on Information Retrieval with Asian Languages (IRAL'99)*, which was held in November 1999, in Taipei, Taiwan. According to the proposal, the *Chinese Text Retrieval Tasks* are organized by Hsin-Hsi Chen and Kuang-hua Chen, National Taiwan University at the second workshop and *Cross Language Retrieval* of Asian languages at the third workshop.

In the aspect of the organization, the first and second workshop were co-sponsored by the *National Institute of Informatics* (NII, formerly the National Center for Science Information Systems, NACSIS) and the *Japan society for the Promotion of Sciences* (JSPS) as part of the *"Research for the Future" Program* (JSPS-RFTF 96P00602). After the first workshop the NACSIS reorganized and changed its name to the NII, in April 2000. At the same time, the *Research Center for Information Resources* (RCIR), a permanent host of the NTCIR Project was launched by the NII. The third workshop will be sponsored by the RCIR at the NII.

From the second workshop [4], tasks are proposed and organized by separate groups outside of the NII. This venture added a variety of tasks to the NTCIR Workshop and as a result, attracted participants from various groups.

1.3 Focus of the NTCIR Workshop

From the beginning of the NTCIR project, we have focused on two directions of investigation, i.e., (1)

traditional laboratory-type text retrieval system testing, and (2) challenging issues.

1.3.1 Traditional IR Testing

For the former, we have placed emphasis on retrieval with Japanese and other Asian languages and cross-lingual information retrieval (CLIR). Indexing texts written in Japanese or other East Asian languages, such as Chinese, is quite different from indexing texts in English, French or other European languages since there is no explicit boundary (i.e., no space) between words in a sentence. CLIR is critical in the Internet environment, especially between languages with completely different origins and structure, such as English and Japanese.

Moreover, in scientific texts or everyday-life documents, for example Web documents, in East Asian languages, foreign language terms often appear in the native language texts both in their original spelling and in transliterated forms. To overcome the word mismatch that may be caused by such expression variance, cross-linguistic strategies are needed for even the monolingual retrieval of documents of this type [5].

1.3.2 Challenging Issues

Traditionally, IR has meant the technology that retrieves documents from a huge document collection and produces a ranked list of the retrieved documents in the order of the likelihood of relevance. However, retrieving documents that may contain relevant information is not all that the user may require, and the information in the documents is not always immediately usable. Research on the techniques helping to make the information in the documents more usable, for example, by pinpointing the answer passages in the documents, summarization, etc., and the appropriate evaluation methods are needed.

Each document genre has its own characteristic and usage pattern, and the criteria determining "successful search" may vary accordingly, although traditional IR research has looked at generalized systems which can handle any kind of document based on the generalized criteria of "successful search". For example, Web document retrieval has different characteristics from those of newspaper or patent retrieval, both with respect to the nature of the document itself and the way it is used. We have been interested in the appropriate evaluation methods for each document genre as well as generalized ones.

In the next section we outline the previous workshops. Section 3 describes the test collections used and Section 4 report the results. Section 5 introduces the tasks for the third workshop and discusses some thoughts on future directions.

2. The Previous NTCIR Workshops

This section outlines the previous NTCIR Workshops.

2.1 Tasks

Each participant has conducted one or more of the following tasks at the workshop.

2.1.1 NTCIR Workshop 1 (1998/1999)

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

The test collection NTCIR-1 was used in these three tasks. In the Ad Hoc Information Retrieval Task, the document collection containing Japanese, English and Japanese-English paired documents is retrieved by Japanese search topics. In Japan, document collections often naturally consist of such a mixture of Japanese and English. Therefore the Ad Hoc IR Task at the NTCIR Workshop 1 is substantially CLIR though some of the participating groups discarded the English part and did the task as Japanese monolingual IR.

2.1.2 NTCIR Workshop 2 (2000/2001)

- *Chinese Text Retrieval Task (CHTR):* including English-Chinese CLIR (ECIR; E>C) and Chinese monolingual IR (CHIR tasks, C>C) using the test collection CHIB01, consisting of newspaper articles from five newspapers in Taiwan R.O.C.
- *Japanese-English IR Task (JEIR):* using the test collection of NTCIR-1 and -2, including monolingual retrieval of Japanese and English (J>J, E>E) and CLIR of Japanese and English (J>E, E>J, J>JE, E>JE).
- *Text Summarization Task (TSC: Text Summarization Arrange):* text summarization of Japanese newspaper articles of various kinds. The NTCIR-2 Summ collection Collection was used.

Each task has been proposed and organized by a different research groups rather in an independent way, while keeping good contact and discussion with the NTCIR Project organizing group headed by the author. How to evaluate and what should be evaluated have been thoroughly discussed in a discussion group.

2.2 Participants

2.2.1 NTCIR Workshop 1.

Below is the list of active participating groups that submitted task results. Thirty-one groups, enrolled to participate in the first NTCIR Workshop. Of these groups, twenty-eight groups enrolled in IR tasks (23 in the Ad Hoc Task and 16 in the Cross-Lingual Task), and nine in the Term Recognition task. Twenty-eight groups from six countries submitted results. Two groups worked without any Japanese language expertise.

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

2.2.2 NTCIR Workshop 2

As shown in the Table 1, 45 groups from eight countries registered for the Second NTCIR Workshop and 36 groups submitted results. Among the above, four groups submitted results to both CHTR and JEIR, and three groups submitted results to both JEIR and TSC, and one group did all three tasks. Table 2 shows the distribution of the attribute of each participating group across the tasks.

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

Among them, four groups participated in JEIR without any Japanese language expertise. Many groups could not submit the results (more precisely could not conduct the task) in the TSC because they could not obtain the document data..

Table 1. Number of Participating Groups

Task	subtask	Enrolled	Submitted
CHTR	CHIR	14	10
	ECIR	13	7
	CHTR total	16	11
JEIR	J-J	22	17
	E-E	11	7
	monoLIR total	22	17
	J-E	16	12
	E-J	14	10
	J-JE	11	6
	E-JE	11	4
	J/E CLIR total	17	14
	JEIR total	31	25
TSC	A extrinsic		7
	B intrinsic		5
	TSC total	15	9
total		45	36

Table 2 Attribute of Participating Groups

	University	Natl.Instit.	Company
CHTR	7	2	2
JEIR	15	3	7
TSC	3	1	5
total	20	4	12

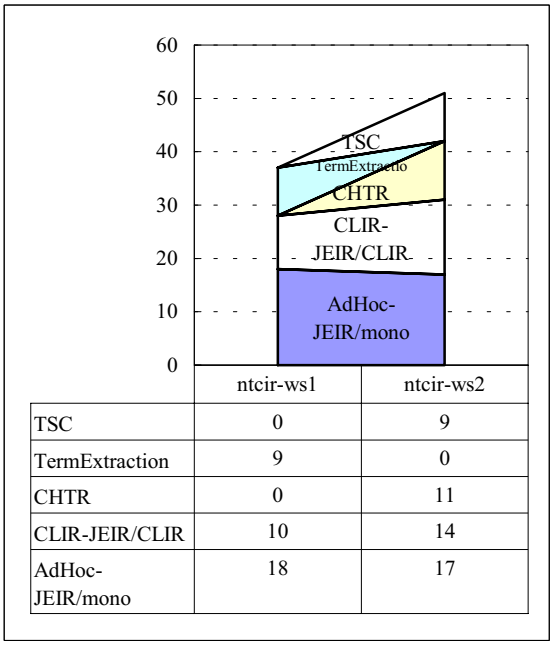
2.2.3 Comparison of the Workshops 1 & 2

Of the 18 participants of the Ad Hoc IR of Japanese and English documents at the first workshop: 10 groups participated in the equivalent tasks at the second workshop, i.e., JEIR monolingual IR tasks, or added participating tasks; one changed task to JEIR CLIR; one changed task to TSC; and six did not participate.

Among 10 CLIR participants at the first workshop: six continued to participate in the

equivalent task, i.e., JEIR-CLIR; two groups changed the tasks to CHTR; and two changed to TSC.

Among nine participating groups in the Term Recognition Task at the first workshop: six changed tasks to JEIR; two changed to TSC; and two did not participate in the second workshop.



	ntcir-ws1	ntcir-ws2
TSC	0	9
TermExtraction	9	0
CHTR	0	11
CLIR-JEIR/CLIR	10	14
AdHoc-JEIR/mono	18	17

Fig. 1 Number of Participants of Each Task

Of the eight groups from the first workshop that did not participate in the second workshop, six are from Japanese universities, one is from a Japanese company and one is from a university in the UK.

Among the participants of CHTR, JEIR, and TSC at the second workshop, seven, 12, and four, respectively, are new to the NTCIR Workshop.

2.3 CLIR Evaluation

A participant could submit the results of more than one run for each task. Both automatic and manual query constructions were allowed. In the case of automatic construction in the JEIR task, the participants had to submit at least one set of results of the searches using only <Description> fields of the topics as *the mandatory run*. The intention of this is to enhance cross-system comparison. For optional automatic runs and manual runs, any field, or fields, of the topics could be used. In addition, each participant had to complete a system description form describing the detailed features of the system.

The relevance judgments were undertaken by pooling methods. The same number of runs were selected from each participating group and the same number of top ranked documents from each run for the topic were extracted and put into the document pool to be judged in order to retain the "fairness" and "equal opportunities" among each participating group. In order to increase the exhaustiveness of the relevance judgments, additional manual searches were conducted for those topics with more relevant documents than a certain threshold (50 in NTCIR-1 and 100 in NTCIR-2). A detailed description of the pooling procedure and the analysis of "fairness" are reported in Kuriyama et al. [6] in this volume.

Human analysts assessed the relevance of retrieved documents to each topic in multi-grades: three grades in the NTCIR-1 and IREX-IR, and four grades in the NTCIR-2 and CIRB010: highly relevant (S), relevant (A), partially relevant (B), irrelevant (C). Some documents will be more relevant than others: either because they contain more relevant information or because the information they contain is highly relevant, then we believe that multi-grade relevance judgments are more natural, or closer to the judgments made in real life [7-9]. However the majority of test collections have viewed relevance judgments as binary and this simplification is helpful for evaluators and system designers.

For NTCIR-1 and -2, two assessors judged the relevance to a topic separately and assigned one of the three or four degrees of relevance. After cross-checking, the primary assessors of the topic, who created the topic, made the final judgment. The *trec_eval* was run against two different lists of relevant documents produced by two different thresholds of relevance, i.e., *Level 1* (or "relevant level file" in NTCIR-1, *rigid relevance* in CIRB010), in which S and A-judgments were rated as "relevant", and *Level 2* (or "partial relevant level file" in NTCIR-1, *relaxed relevance* in CIRB010), in which S, A and B-judgments were rated as "relevant", even though the NTCIR-1 does not contain S.

2.3.1 Measure for Multigrade Judgments

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

Most of IR systems and experiments have assumed that the highly relevant items are useful to all users. However some user-oriented studies have suggested that partially relevant items may important for a specific users and they sould not be collapsed into relevant items, but should be analyzed separately [9]. More investigation is needed.

3. Test Collections

Table 3 shows the IR test collections construced through the First and Second NTCIR Workshops and its ex-partner (now colleague of NTCIR) IREX.

Addition to above, *NTCIR-2 Summ* contains ca.100 + ca. 2000 (*NTCIR-2 TAO Summ*) manually created summaries of various types of Japanese newspaper articles 1994, 1995 and 1998.

3.1 Documents

More than half of the documents in the NTCIR-1 JE Collection are English-Japanese paired. NTCIR-2 contains author abstracts of conference papers and extended summaries of grant reports. About one-third of the documents are Japanese- and English-paired, but the correspondence between English and Japanese is unknown during the workshop. A sample document record of the JE Collection in the NTCIR-1 is shown in Fig. 2. Documents are plain text with SGML-like tags in the NTCIR collections and the IREX-IR. A record may contain document ID, title, a list of author(s), name and date of the conference, abstract, keyword(s) that were assigned by the author(s) of the document, and the name of the host society.

A sample Document record used in the CLIR at the NTCIR Workshop 3 is shown in Fig. 3. All the document collection in four languages are coded in the same set of mandatory tags and some optional tags. A document record in the CIRB010 is coded by XML, but the elements are similar.

```
<REC>
<ACCN>gakkai-0000011144</ACCN>
<TITL TYPE="kanji">電子原稿・電子出版・電子図書館-「SGML 実験誌」の作成実験を通して</TITL>
<TITE TYPE="alpha">Electronic manuscripts, electronic publishing, and electronic library </TITE>
<AUPK TYPE="kanji">根岸 正光</AUPK>
<AUPE TYPE="alpha">Negishi, Masamitsu</AUPE>
<CONF TYPE="kanji">研究発表会(情報学基礎)</CONF>
<CNFE TYPE="alpha">The Special Interest Group Notes of IPSJ</CNFE>
<CNFD>1991. 11. 19</CNFD>
<ABST TYPE="kanji"><ABST.P>電子出版というキーワードを中心に、文献の執筆、編集、印刷、流通の過程の電子化について、その現状を整理して今後の動向を検討する。とくに、電子出版に関する国際規格である SGML (Standard Generalized Markup Language)に対するわが国での動きに注目し、学術情報センターにおける「SGML 実験誌」およびその全文 CD-ROM 版の作成実験を通じて得られた知見を報告する。また電子図書館について、その諸形態を展望する。出版文化に依拠するこの種の社会システムの場合、技術的な問題というのは、その技術の社会的な受容・浸透の問題であり、この観点から標準化の重要性を論じる。</ABST.P></ABST>
<ABSE TYPE="alpha"><ABSE.P>Current situation on electronic processing in preparation, editing, printing, and distribution of documents is summarized and its future trend is discussed, with focus on the concept: "Electronic publishing: Movements in the country concerning an international standard for electronic publishing. Standard Generalized Markup Language (SGML) is assumed to be important, and the results from an experiment at NACSIS to publish an "SGML Experimental Journal" and to make its full-text CD-ROM version are reported. Various forms of "Electronic Library" are also investigated. The author puts emphasis on standardization, as technological problems for those social systems based on the cultural settings of publication of the country, are the problems of acceptance and penetration of the technology in the society.</ABSE.P></ABSE>
<KYWD TYPE="kanji">電子出版 // 電子図書館 // 電子原稿 // SGML // 学術情報センター // 全文データベース</KYWD>
<KYWE TYPE="alpha">Electronic publishing // Electronic library // Electronic manuscripts // SGML // NACSIS // Full text databases</KYWE>
<SOCN TYPE="kanji">情報処理学会</SOCN>
<SOCE TYPE="alpha">Information Processing Society of Japan</SOCE>
</REC>
```

Fig. 2 Sample Document (NTCIR-1, JE)

```
<DOC>
<DOCNO>chinatimes_focus_0005660</DOCNO>
<LANG>CH</LANG>
<DATE>05071999</DATE>
<HEADLINE>解決高鐵融資 尋求第三管道</HEADLINE>
<TEXT>
<P>【記者羅兩莎台北報導】據負責台灣高速鐵路聯合貸款的主・銀行表示，高鐵融資問題目前仍・在銀行團、交通部高鐵局以及台灣高鐵公司「三方合約」・容的訂定。在銀行團和交通部一直未能就相關・見達成共識之下，三大主・銀行原則決定，將尋求行政院經建會等第三管道與交通部協調，以儘早解決銀行團和交通部之間對融資問題的・見。
</P>
<P>高鐵案將向國・銀行融資二千八百多億元，這項聯貸案確定由交銀、台銀和中國國際商業銀行共同主・。不過，由於高鐵是國・首宗ＢＯＴ案，潛在風險究竟有多高，銀行無從評估。三大主・銀行與交通部和台灣高鐵公司訂定貸款合約時，重點亦著重在風險控制以及債權確保。</P>
<P>據主・銀行主管表示，銀行當然希望債權確保不會有問題，譬如，在三方合約中訂定，由政府出面保證萬一將來台灣高鐵公司蓋不下去時，政府可以出面買下，負責把工程完成等。</P>
</TEXT>
</DOC>
```

Fig 3. Sample Document (NTCIR-3 clir CH)

Table 3 IR Test Collections

collection	documents				topic	rel judgment
		rec#	size	genre		
CIRB010	C	132K	200MB	newspaper '98-99	C:50 E:50	4 grades
NTCIR-1	JE	340K	577MB	scientific abstract	J:83	3 grades
	J	333K	312MB		83	
	E	187K	218MB		J:60	
NTCIR-2	J	403K	600MB	scientific abstract	49	4 grades
	E	135K	200MB		49	
IREX-IR	J	222K	221MB	newspaper '94-95	50	3 grades

3.2 Topics

A sample topic record which will be used in the CLIR at the NTCIR Workshop 3 is shown in Fig. 4. Topics are defined as statements of "user's requests" rather than "queries", which are the strings actually submitted to the system, since we wish to allow both manual and automatic query construction from the topics. Among the 83 topics of the NTCIR-1, 20 topics were translated into Korean and were used with the Korean HANTEC Collection [11]

The topics contain SGML-like tags. A topic in NTCIR-1, NTCIR-2 and CIRB010 contains similar tag set though tags are longer than above (ex. <DESCRIPTION>), and consists of the title of the topic, a description (question), a detailed narrative, and a list of concepts and field(s). The title is a very short description of the topic and can be used as a very short query that resembles those often submitted by end-users of Internet search engines. Each narrative may contain a detailed explanation of the topic, term definitions, background knowledge, the purpose of the search, criteria for judgment of relevance, etc.

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

Fig. 4 A Sample Topic (CLIR at NTCIR WS 3)

3.3 Relevance Judgments (Right Answers)

The relevance judgments were conducted using multi-grades as stated in the section 2.3. In NTCIR-1 and -2, relevance judgment files contain not only the relevance of each document in the pool, but also contain extracted phrases or passages showing the reason the analyst assessed the document as "relevant". These statements were used to confirm the judgments and also hoped future use in experiments of the extracting answer passages or so.

3.4 Linguistic Analysis

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

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

3.5 Robustness of the System Evaluation using the Test Collections

The test collections NTCIR-1 and -2 have been tested for the following aspects so that they can be used as a reliable tool for IR system testing:

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

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

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

It was found a strong correlation between the system rankings produced using different relevance judgments and different pooling methods, regardless

of the inconsistency of the relevance assessments among analysts and regardless of the different pooling methods [6,13-15]. It served as an additional support to the analysis reported by Voorhees [17].

4. Evaluation Results of CLIR at NTCIR WS 2

4.1 English-Chinese CLIR (ECIR)

The 17 search results of ECIR task are submitted from 7 participating groups. According to the task overview report [18], query expansion is a good method to increase system performance. In general, the probabilistic model shows better performance. For ECIR task, select-all approach seems to be better than other select-X approaches in dictionary look-up, if no further techniques are adopted. PIRCS used MT approach and it out performed. For ECIR task, word-based indexing approach is better.

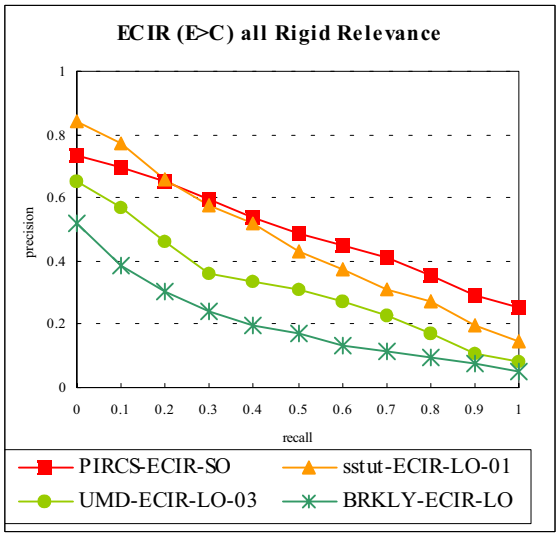


Fig. 5 Top ECIR Runs (All runs) Rigid Relevance

4.2 CLIR of English and Japanese [19]

There were 95 submitted runs for CLIR of Japanese and English from 14 groups. For J-E, E-J, J-JE, E-JE, 40 runs from 12 group, 30 runs from 10, 14 runs from 6, and 11 runs from 4 were submitted respectively.

Most of groups used query translation approach but LISIF group used an approach combined query translation and query translation. The top 1000 documents in the initial search were translated and further processing was done on them. Three groups used corpus based approach but generally the performances were less effective compared with other approach though some of them participated in the NTCIR Workshop 1 and the relative performance was better. New approaches including flexible pseudo-relevance feedback, segmented LSI were proposed.

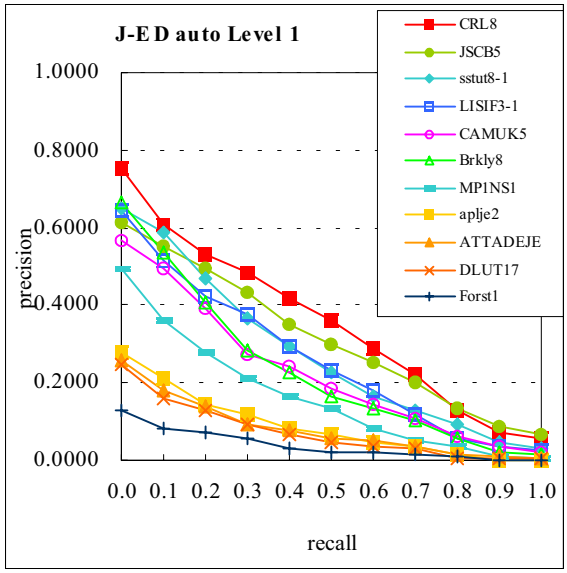


Fig 6 J-E Runs (D) Level 1

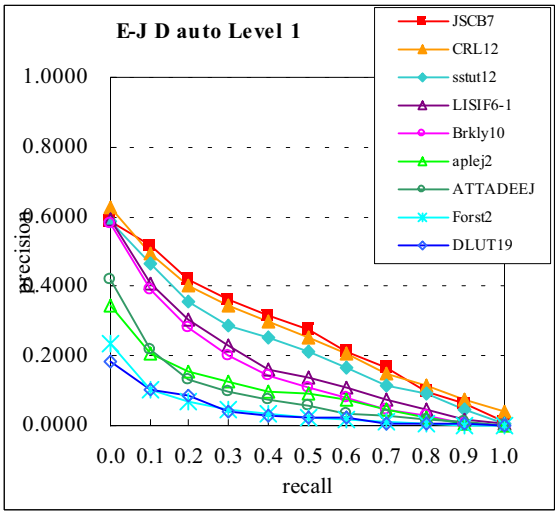


Fig 7 E-J Runs (D) Level 1

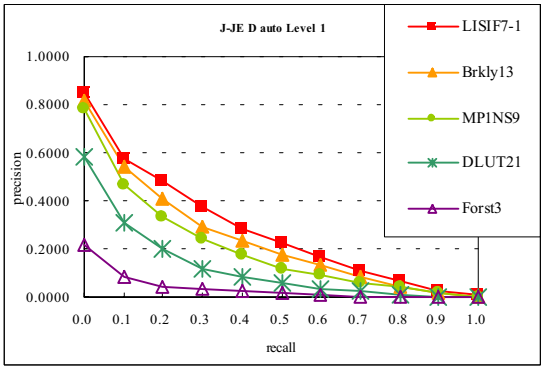


Fig 8 J-JE Rus (D) Level 1

4.3 Issues for the Next CLIR Evaluation

In the round table dicussion at the NTCIR Workshop 3 and the Program committee meeting,

and after Workshop meeting, some issues were raised to conduct more appropriate and valid evaluation at the next workshop.

CHTR and JEIR at the second workshop were organized rather an independent way but we aimed to follow the consistent or at least compatible procedures each other. However regretably we could find unintended incompatibility between CHTR and JEIR including categories of query types and pooling methods. The CLIR task at the NTCIR Workshop 3 will be organized by the organizers of CHTR and JEIR, and HANTEC group. The organizers had face-to-face meetings and decided detailed procedures included topic creation, topic format, document format, query types and mandatory runs. Pooling will be done once, so there will never be inconsistency. For query type, the mandatory run is the one using <DESCRIPTION> only and we are also keen to the difference between search using <CONCEPT> or without it. For the details, please consult http://research.nii.ac.jp/ntcir/workshop/clir/CFPinNTCIR3CLIRr.htm

The other issue is reuse of training set and experiment design using paired corpus. At the NTCIR Workshop 3, bigger and higher quality paired corpus of English and Japanese will be provided in the Patent Retrieval Task, but we plan to allow to use 1995-1997 parallel corpus for training and dictionary development and the test will be done using full patent documents of 1998-1999 and parallel corpus of 1998-1999 are not allowed to use.

Documents sets were also problematic. At the Second workshop, text summarization task used Mainichi Newspaper corpus of 1994, 1995 and 1998 and asked the participants obtained the data from the newspaper company since they sell the corpus for research purpose use. As a results some of the participating groups did not obtain the data and could not conduct the task. For the next workshop, the NII will provide all the data for participants though the Mainichi Newspaper documents allowed only limited years of use; two years for Japanese participants, and up to 7 years for participants from outside Japan.

5. NTCIR Workshop 3

The third NTCIR Workshop will start from September 2001 and the workshop meeting will be held in October 2002. We picked five areas of research as tasks. The updated information will be found at http://research.nii.ac.jp/ntcir/workshop/.

5.1 Tasks

Below is a brief summary of the tasks envisaged for the Workshop. A participant will conduct one or more of the tasks or subtasks below. Participation in only one subtask (for example Japanese monolingual IR (J-J) in the CLIR Task) is available:

5.1.1 Cross Language Retrieval Task (clir)

Documents and topics are in four languages (Chinese, Korean, Japanese and English). 50 topics for the collections of 1998-1999 (Topic98) and 30 topics for the collection of 1994.(Topic94) Both topic sets contain four languages (Chinese, Korean, English and Japanese).

(a) *Multilingual CLIR (MLIR)*: Search document collection more than one languages by one of four languages of topics. Excepting Korean documents because of time range difference. (Xtopic98>CEJ)

(b) *Bilingual CLIR (BLIR)*: Search of any two different languages as language and documents, excepting search of English documents (Xtopic98>C, Xtopic94>K, Xtopic98>J)

(c) *Single Lanugage IR (SLIR)*: Monolingual Search of Chinese, Korea, or Japanese.(Ctopic98>C, Ktopic94>K, Jtopic98>J)

DOCUMENT: newspapers publish in Asia:

- Chinese: *CIRB010*, *United Daily News* (1998-1999)
- Korean: *Korea Economic Daily* (1994)
- Japanese: *Mainichi Newspaper* (1998-1999)
- English: *Taiwan News, China English News*, *Mainichi Daily News* (1998-1999)

5.1.2 Patent Retrieval Task (patent)

(a) Main Task

- *Cross-language Cross-DB Retrieval*: retrieve patents in response to J/E/C newspaper articles associated with technology and commercial products. 30 query articles with short description of search request.
- *Monolingual Associative Retrieval*: retrieve patents associated with an input Japanese patent. 30 query patents with short description of search requests.

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

DOCUMENT:

- Japanese patents: 1998-1999 (ca. 17GB, 700K docs)
- Japio patent abstracts: 1995-1999 (ca.1750K docs)

- Patent Abstracts of Japan (English translations for Japio patent abstracts): 1995-1999 (ca. 1750K)
- Patolis test collection (34 topics and relevance assessment on the Patent 1998)
- Newspaper articles (Japanese/ English/ Traditional Chinese)

5.1.3 Question Answering Task (qac)

(a) *Task 1*: System extracts five answers from the documents in some order. 100 questions. System is required to return support information for each answer of the questions. We assume the support informationas a paragraph, 100 letter passage or document which includes the answer.
(b) *Task 2*: System extracts only one answer from the documents. 100 questions. Support information is required.
(c) *Task 3*: evaluation of a series of questions. The related questions are given for the 30 of questions of Task 2.

DOCUMENT: Japanese newspaper articles (Mainichi Newspaper 1998-1999)

5.1.4 Automatic Text Summarization Task (tsc2)

(a) *Task A (single document summarization):* Given the texts to be summarized and summarization lengthes, the participants submit summaries for each text in plain text format.
(b) *Task B (multi-document summarization):* Given a set of texts, the participants produce summaries of it in plain text format. The information which was used to produce the document set, such as queries, as well as summarization lengthes are given to the participants.

DOCUMENT: Japanese newspaper articles (Mainichi Newspaper 1998-1999)*

5.1.5 Web Retrieval Task

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

DOCUMENT: Web documents mainly collected from jp domain (ca.100GB & ca.10GB) Available at the "Open-Lab" in the NII

5.2 Workshop Schedule

2001-09-30 Application Due
2001-10-01 Document release (newspaper)
2001-10/2002-01 Dry Run and Round-Table Discussion (varied with on each task)
2001-12 Open Lab start
2001-12/2002-03 Formal Run (varied with each task)
2002-07-01 Evaluation Results Delivery
2002-08-20 Paper for Working Note Due
2002-10-08/10 NCIR Workshop 3 Meeting
Days 1-2: Closed session (task participants only)
Day 3: Open session
2002-12-01 Paper for Final Proceedings Due

5.3 Features of the NTCIR Workshop 3 Tasks

For the next workshop, we plan some new ventures including below;.

(1) Multilingual CLIR (CLIR)
(2) Search by Document (Patent, Web)
(3) Passage Retrieval or submit "evidential passages", passages to show the reason why the documents are supposed to be relevant (Patent, QA, Web)
(4) Optional Task (Patent,Web)
(5) Multigrade Relevance Judgments (CLIR, Patent, Web)
(6) Precision Oriented Evaluation (QA, Web)

For (1), it was our first trial of the CLEF model in the Asia. Also we would like to invite any other language groups who wish to join us by providing document data and relevance judgments or by providing query tranlsation.

For (3), we suppose that idintifying most relevant passage in the retrieved documents are needed when retrieving longer documents like Web documents or patents. The primary evaluation will be done document base but we will use the submitted passages as a secondary information for further analysis.

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

For (5), we have used multigrade relevance judgment so far and proposed new measures, Weighted Average Precision and Weighted R Precision for the purpose. We will continue this line

of investigation and will add "top relevant" for Web Task as well as evaluation by trec_eval.

5.4 Future Directions

In the future, we desire the enhancement of the investigation in the following directions:

Evaluation of CLIR systems
Evaluation of retrieval of new document genres and more realistic evaluation
Evaluation of technology to make information in the documents immediately usable.

One of the problems of CLIR is the availability of resources that can be used for translation. Enhancement of the processes of creating and sharing the resources is important. In the NTCIR Workshops, some groups automatically constructed a bilingual lexicon from a quasi-paired document collection. Such paired documents can be easily found in non-English speaking countries and on the Web. Studying the algorithms to construct such resources and sharing them is one practical way to enrich the applicability of CLIR. International collaboration is needed to construct multilingual test collections and to organize the evaluation of CLIR, since creating topics and relevance judgments are language- and cultural-dependent, and must be done by native speakers. Cross-lingual summarization and qustion answering are also considered for the future workshops.

REFERENCES

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

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

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

[4] *NTCIR Workshop 2 : Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, June 2000- March 2001（ISBN : 4-924600-96-2)

[5] Kando, N.: Cross-Linguistic Scholarly Information Transfer and Database Services in Japan. Annual Meeting of the ASIS, Washington DC. Nov. 1, 1997

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

[7] Spink, A., Bateman, J. From highly relevant to not relevant: Examining different regions of relevance. *Information Processing and Management,* Vol.34, No.5, pp.599-622, 1998

[8] Dunlop, M.D.Reflections on Mira, *Journal of the Americal Society for Information Sciences,* Vol.51, No.14, pp.1269-1274, 2000

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

[10] Kando, N., Kuriyama, K., Yoshioka, M. Evaluation based on multi-grade relevance judgements. *IPSJ SIG Notes,* Vol. 2001-FI-63, pp. 105-112, July 2001.

[11] Sung, H.M. "HANTEC Collection". Presented at the panel on IR Evaluation in the 4th IRAL, Hong Kong, 30 Sept.-3 Oct. 2000.

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

[13] Kando, N, Nozue, T., Kuriyama, K., Oyama, K.: NTCIR-1: Its Policy and Practice, *IPSJ SIG Notes*, Vol.99, No.20, pp. 33-40, 1999 [in Japanese].

[14] Kuriyama, K., Nozue, T., Kando, N., Oyama, K.: Pooling for a Large Scale Test Collection: Analysis of the Search Results for the Pre-test of the NTCIR-1 Workshop, *IPSJ SIG Notes*, Vol.99-FI-54, pp.25-32 May, 1999 [in Japanese].

[15] Kuriyama, K., Kando, K.: Construction of a Large Scale Test Collection: Analysis of the Training Topics of the NTCIR-1, *IPSJ SIG Notes*, Vol.99-FI-55, pp.41-48, July 1999 [in Japanese].

[16] Kando, N., Eguchi, K., Kuriyama, K.: Construction of a Large Scale Test Collection: Analysis of the Test Topics of the NTCIR-1, In *Proceedings of IPSJ Annual Meeting* [in Japanese]. pp.3-107 -- 3-108, 30 Sept -3 Oct. 1999.

[17] Voorhees, E.M.: Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness, In *Proceedings of 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. pp. 315-323, Melbourne, Australia, August. 1998

[18] Chen, K.H., Chen, H.H.: The Chinese Text Retrieval Tasks of NTCIR Workshop II. In *NTCIR Workshop 2 : Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, June 2000- March 2001（ISBN : 4-924600-96-2)

[19] Kando, N., Kuriyama, K., Yoshioka, M.: Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop. In *NTCIR Workshop 2 : Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, June 2000- March 2001（ISBN : 4-924600-96-2)