

Overview of the Second NTCIR Workshop

Noriko Kando

National Institute of Informatics (NII), Japan

kando@nii.ac.jp

Abstract:

This paper introduces the Second NTCIR Workshop, an evaluation workshop, 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. It contains three tasks: Chinese Text Retrieval (CHTR), Japanese and English Information Retrieval (JEIR) and Text Summarization Challenge (TSC). Forty-five groups from eight countries have registered for one or more tasks. A brief history, tasks, participants, and test collections used in the workshop are described in this paper. To conclude, some thoughts on future directions are suggested.

1. Introduction

The Second NTCIR Workshop was co-sponsored by the National Institute of Informatics (NII, formerly the National Center for Science Information Systems, better known as NACSIS) and the Japan Society for the Promotion of Science as part of the "Research for the Future" Program (JSPS-RFTF96P00602). After the First NTCIR 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.

In the aspect of organization, the Second NTCIR Workshop is the first workshop hosting tasks organized by separate groups outside of the NII. They are the Chinese Text Retrieval Tasks and the Text Summarization Challenge. This venture added a variety of tasks to the NTCIR Workshop.

1.1 Purpose

The purposes of the NTCIR Workshop [1] 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.

The process of the Second NTCIR Workshop started in June 2000. We call the whole process 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.

1.2 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 worked together to organize the second NTCIR Workshop. The new challenging task of Text Summarization became feasible with this collaboration.

The international collaboration to organize Asian 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. For various reasons, the Korean IR evaluation HANTEC [4] was organized separately as a domestic venture, but both HANTEC and NTCIR have kept a close relationship with each other. Part of the search topics were exchanged and the results of the HANTEC was reported at the second NTCIR Workshop meeting.

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.

Traditional IR Testing

For the former, we have placed emphasis on retrieval with Japanese or 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].

Challenging Issues

There has been a strong push to investigate technology beyond the scope of traditional text retrieval. For example, the intersection of natural language processing (NLP) and IR, and a more realistic evaluation using more realistic types of documents for today's world, such as Web documents with both multilingual and multi-modal information. 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. NLP techniques help to make the information in the documents more usable, for example, by pinpointing the answer passages in the documents, summarization, and so on.

Moreover, each document genre has its own user group 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 investigated appropriate evaluation methods for each document genre.

In order to respond to the needs stated above, we have placed emphasis on CLIR and investigation of the intersection of NLP and IR to date. We employed real users of the document genre as the

topic authors and assessors. In the near future, we plan to move into more realistic evaluation using realistic types of documents for today's world, such as Web documents.

1.4 Evaluation Workshops

We call the NTCIR Workshop an "evaluation workshop" or just an "evaluation". An "evaluation workshop" provides to participants a set of data usable for experiments and unified procedures for evaluation of experiment results. Each participating research group conducts research and experiments with various approaches using the data provided. Each participant can participate in the workshop with their own purpose of experimentation.

The first, and one of the most successful examples of evaluation workshops in information retrieval, is the workshop series called the Text Retrieval Conference (TREC), which has been organized by the National Institute of Standards and Technology (NIST) in the United States since 1992 [6]. Large-scale test collections, which are comparable to the size of document collections searched in an operational setting, unified evaluation procedures, and various new techniques have been developed through TREC.

Benefits of Evaluation Workshops

The benefits of the IR evaluation workshops are, among others: (1) the development of large-scale test collections, (2) the facilitation of technology transfer, (3) the creation of a "Who's Who" environment, a forum of researchers, for exchanging research ideas, (4) the showcasing of state-of-the-art technology, (5) the accumulation of data usable for research, (6) the motivation towards research in specific topics, and (7) the creation of a model of unified evaluation procedures [7]. For example, the search effectiveness of the best systems against the same collection has more than doubled through the eight-year experience of the TREC Conferences. The impact of such evaluation workshops in enhancing and encouraging research of the topics is obvious.

An IR test collection contains: (1) the document collection, (2) topics, and (3) relevant judgments (correct answers) for each topic. A topic is a written statement of a user's information needs. Relevance judgments are exhaustive lists of the relevant documents for each topic. In a large-scale test collection, it is impossible to judge every document for relevance. Instead, relevance judgments are conducted through *pooling* [8]—a certain number of top-ranked documents are collected from the search results of the various IR systems and are used to create a document pool of candidates for relevant documents. Human analysts assess the relevance of each document in the pool against the topic instead of judging all documents. The documents not included in the pool are not judged and are assumed to be irrelevant.

It is known that different IR systems can retrieve different relevant documents. We can then assume that if a sufficient number of diverse systems contribute results to a pool, it is likely that a large percentage of all relevant documents will be included. An evaluation workshop in which a wide variety of systems participated is one of the best opportunities for better pooling.

The success of the TREC has stimulated the construction of large-scale test collections in various languages as well as IR evaluation workshops, such as Amarylus [9], Topic Detection and Tracking (TDT) [10], Cross Language Evaluation Forum (CLEF) [11] and NTCIR. Each collection has its own strengths and each evaluation project has been motivated by its own specific needs and characteristics.

The evaluation workshop is also useful in constructing the data set, which is reusable for experiments in text processing technology, since the right answers can be set by consensus of the participants and the discussion among the participants itself stimulates the research. The Message Understanding Conference (MUC), the Summarization Conference (SUMMAC), the IREX-NE, and the Term Recognition Task at the First NTCIR Workshop are examples.

In the next section we outline this NTCIR Workshop. Section 3 describes the test collections used and Section 4 discusses some thoughts on future directions.

2. The Second NTCIR Workshop

This section outlines the Second NTCIR Workshop.

2.1 Tasks

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

Chinese Text Retrieval Tasks (CHTR): including subtasks of English-Chinese CLIR (ECIR) and Chinese monolingual IR (CHIR) using the test collection CIRB010, consisting of newspaper articles from five newspapers in Taiwan R.O.C.

Japanese-English IR Tasks (JEIR): using the test collections of NTCIR-1 and -2, including subtasks of 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 Challenge (TSC): text summarization of Japanese newspaper articles of various kinds, including intrinsic and extrinsic evaluations. The NTCIR-2 Summ collection is used.

The new challenging task is called "Challenge". Each task or challenge has been proposed and organized by a different research group 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 as a new "Challenge" has been thoroughly discussed in a discussion group.

2.2 Participants

Participants of the Second NTCIR Workshop

Below is a list of the active participants of the Second NTCIR Workshop. The term "active participant" means the participating research group that enrolled for one or more tasks set by the NTCIR Workshop organizers and submitted the results.

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)

As shown in the Table 1, 45 groups from eight countries registered for the Second NTCIR Workshop and 36 groups submitted results. Among them, 11 submitted results for CHTR, 25 for JEIR, and nine for TSC.

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. Among 36 groups, 20 are from universities, four are from non-profit national research institutes, and 12 are from companies. Table 2 shows the distribution of the attribute of each participating group across the tasks.

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

Some of the participating groups are joint groups of universities and companies or multi-nationals. The attribute and country of each participating group shown in these tables were counted according to the first entity of the group name that was originally registered to the NTCIR organizers. For example, "University of Cambridge/Toshiba/Microsoft" was seen as a university research group from the United Kingdom.

Table 3. Distribution of Participating Groups

	CHTR		JEIR.		TSC	
	enrl	sub	enrl	sub	enrl	sub
Canada	1	0	1	0	0	0
China	2	2	0	0	0	0
Hong Kong	2	1	0	0	0	0
Japan	3	3	21	18	12	9
Korea	0	0	1	1	0	0
Taiwan	2	2	2	2	1	0
UK	1	0	2	1	0	0
USA	5	3	4	3	2	0
total	16	11	31	25	15	9

The distribution of the countries and areas of the participants is shown in Table 3. Among them, four groups (three are from US and one are from Japan, but all the investigators are international visiting researchers) participated in JEIR without any Japanese language expertise. Many groups could not submit the results in the TSC because they could not obtain the document data. It is one of the problems we will have to resolve for the next workshop.

Comparison with the First NTCIR Workshop

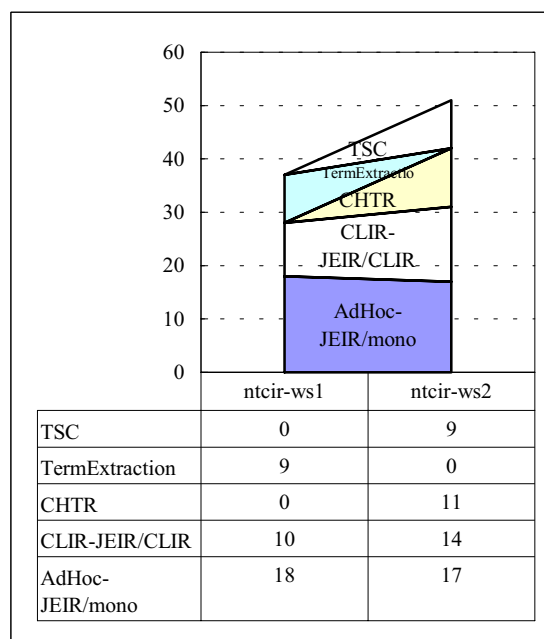
The active participants of the First NTCIR Workshop comprised 28 groups from six countries. Figure 1 shows the number of participants of each task in the first and second workshops. Comparison between these two workshops is not easy because

some of the participating groups changed names, joined with different groups, or separated into two.

When we counted using the first entity of a group we found that, of the 18 participants of the Ad Hoc IR of Japanese and English documents at the first workshop: 10 groups also 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.

**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 and did not participate in the previous workshop.

2.3 Procedures and Evaluation

The process of the second workshop was as follows:

- 1 June 2000:** Call for Participation to Tasks and distribution of the JEIR training data
- 10 August 2000:** distribution of the JEIR test data (new documents and 49 J/E topics)
- 30 August 2000:** distribution of the CHTR test data (new documents and 50 C/E topics)
- 8 September 2000:** TSC dry run
- 18 September 2000:** submission of JEIR results
- 20 October 2000:** submission of CHTR results
- 27 November - 1 December 2000:** TSC test
- 28 December 2000:** distribution of TSC evaluation results
- 10 January 2001:** distribution of CHTR & JEIR evaluation results
- 7-9 March 2001:** workshop meeting

CHTR and JEIR:

A participant could submit the results of more than one run for each task.

For CHTR and JEIR tasks, 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 and submit a system description form describing the detailed features of the system.

Human analysts assessed the relevance of retrieved documents to each topic. The relevance judgments (right answers) for the test topics were delivered to active participants who submitted search results. Based on these assessments, interpolated recall and precision at 11 points, average precision (non-interpolated) over all relevant documents, and precision at five, 10, 15, 20, 30, and 100 documents were calculated using TREC's evaluation program, which is available from the ftp site of Cornell University.

TSC:

Research in automatic text summarization has been carried out since the 1950s, but evaluation of

the technology is still a very challenging issue. In the Second NTCIR Workshop, we conducted both intrinsic and extrinsic evaluations. For the intrinsic evaluation, emphasis is placed on "round-table evaluation" and creating a reusable data set. Professional captionists created two types of summaries as the "right answer"; extract-type and abstract-type summaries. The former is created by selecting important sentences from the original articles and the latter is done by summarizing the original articles freely without worrying about sentence boundaries, trying to obtain the main ideas of the articles. Each submitted summary was then rated by these professional captionists, comparing it with those two "right answers" and the automatically created random summary of the article. Those evaluation results intended to serve as reference data for the round-table discussion at the workshop meeting, where all the participants share their experience and can have detailed discussions of the technology. For the extrinsic evaluation, we chose an IR task-based evaluation, which is similar to the method used at SUMMAC [12].

The overviews for each task were prepared separately and information on the participating groups and their systems can be found in the individual group reports in the rest of this volume, also available from the NTCIR web site.

3. Test Collections

Through the First and Second NTCIR Workshops and its ex-partner (now colleague of NTCIR) IREX, the following test collections or data sets, usable for laboratory-type testing of IR and related text processing technology, were constructed.

CIRB010; 132,173 (200MB) Chinese articles from five Taiwan newspapers of 1998 and 1999. Fifty Chinese topics and their English translations, four-grade relevance judgments

NTCIR-1; JE, J, and E collections. JE collections containing 339,483 (577MB) Japanese and English documents. Author abstracts of conference papers hosted by 65 academic societies in Japan. More than half are English-Japanese paired. The J (332,918 documents, 312MB) and E (187,080 documents, 218MB) collections are constructed by extracting the Japanese or English parts of the documents, respectively, from the JE Collection of 83 Japanese topics, three-grade relevance judgments. It contains a tagged corpus

NTCIR-2; J and E collections. 403,248 Japanese documents (600MB) and 134,978 English documents (200MB), 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. Forty-nine Japanese topics and their English translations, four-grade relevance judgments. It contains the segmented data of Japanese Collections.

NTCIR-2 Summ; ca.100 + ca. 2000 (NTCIR-2 TAO Summ) manually created summaries of various types of Japanese articles from Mainichi Newspaper of 1994, 1995 and 1998.

IREX-IR; 221,853 Japanese newspaper articles (221MB) from Mainichi Newspaper of 1994 and 1995, 30 Japanese topics, three-grade relevance judgments

IREX-NE; Named entity extraction from Japanese newspaper articles.

```

The rest of the section outlines the test
collections usable for IR experiments.<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 A Sample of a Document Record

3.1 Documents

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 document record in the CIRB010 is coded by XML, but the elements are similar.

3.2 Topics

A sample topic record of the NTCIR-1 is shown in Fig. 3. Topics are defined as statements of "user's requests" rather than "queries", which are the strings actually submitted to the system, since we wish to allow both manual and automatic query construction from the topics. The NTCIR-1 contains 30 training topics and 53 test topics. Among them, 21 training topics and 39 test topics are usable for cross-lingual retrieval. Among the 83, 20 topics were translated into Korean and were used with the Korean HANTEC Collection [4]

The topics contain SGML-like tags. A topic in NTCIR-1, NTCIR-2 and CIRB010 consists of the title of the topic, a description (question), a detailed narrative, and a list of concepts and field(s) (see Fig. 3). 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.

Query types, combination of topic fields, used in tasks are described in Section 3.6.

```

<TOPIC q=0005>
<TITLE>
特徴次元リダクション
</TITLE>
<DESCRIPTION>
クラスタリングにおける特徴次元リダクション
</DESCRIPTION>
<NARRATIVE>
オブジェクトのクラスタリングを行なうとき、オブジェク
トを特徴ベクトルで表現することが望まれる。アプリケー
ションによっては、オブジェクトの次元は数千、数万とな
ることがある。このような場合、事前に次元を落とすこと
が必要になる。正解文書は、特徴次元リダクションの方法
について、理論面から、または実験によって、提案、比較
などを行なっているもの。画像処理などの実験の操作の一
部として特徴次元リダクションを用いているだけでは要求
を満たさない。
</NARRATIVE>
<CONCEPT>
特徴選択, 主成分分析, 情報の粒度, 幾何クラスタリング
</CONCEPT>
<FIELD>
1.電子・情報・制御
</FIELD>
</TOPIC>

```

Fig. 3 A Sample Topic

Multi-grade Judgments

The relevance judgments were undertaken by pooling methods. Assessors and topic authors are always the users of the document genre. The relevance judgments were conducted using multi-grades: three grades in the NTCIR-1 and IREX-IR, and four grades in the NTCIR-2 and CIRB010. 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 [13-17]. However the majority of test collections have viewed relevance judgments as binary and this simplification is helpful for evaluators and system designers. Most of IR evaluation metrics are constructed based on the binary judgments.

It was announced to use TREC's evaluation program for the formal evaluation of the IR tasks at the First and Second NTCIR Workshops. To run TREC's evaluation program to calculate mean average precision, recall-level precision, document level precision, we set two thresholds for the level of relevance.

For NTCIR-1 and -2, the assessors are researchers in each subject domain since they contain scientific documents; 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 also created the topic, made the final judgment. The TREC's evaluation program 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), in which "highly relevant (S)" and "relevant (A)" are rated as "relevant", and *Level 2* (or "partial relevant level file" in NTCIR-1), in which S, A and "partially relevant (B)" were rated as "relevant", even though the NTCIR-1 does not contain "highly relevant (S)".

Judgments by Different Users

Relevance judgments in the CIRB010 were conducted according to the method originally proposed by Chiang and her supervisor Kuang-hua Chen, who was one of the organizers of the Chinese IR Task at the Second NTCIR Workshop [18]. Three different groups of users, i.e., information specialists including librarians, subject specialists, and ordinary people, conducted judgments separately and assigned to each document one of four different degrees of relevance: very relevant (3), relevant (2), partially relevant (1) and irrelevant (0). Then, three relevance judgments assigned by each assessor were averaged out to between 0 and 1 using the formula below:

The so-called *rigid relevance* means the final relevance should be between 0.6667 and 1. This is equivalent to each assessor assigning "relevant (2)" or higher to the document, and corresponds to *Level 1* in NTCIR-2. The so-called *relaxed relevance* means that the final relevance should be between 0.3333 and 1. That is to say, it is equivalent to each assessor assigning "partially relevant (1)" or higher to the document, and corresponds to *Level 2* in NTCIR-2. The TREC's evaluation program was run against these two levels of relevance.

The reason three different groups of users were employed as assessors is because the genre of newspaper articles is used by various kinds of users. The idea of averaging out the assessments by different user groups is new compared to the traditional approach of test collection building, in which the topic author should be the most qualified assessor. A similar idea was mentioned by Dr Andrei Broder, Vice President for Research and Chief Scientist at Alta Vista, in his invited talk at the TREC-9 Conference held on 13-16 November 2000. He proposed the need to average out the relevance judgment of 15 to 20 users in the evaluation of Web search engines, since the users of the systems are heterogeneous and systems cannot know the user's profile during the search.

It was not clearly mentioned in the task organizers' report that the attributions of the assessors actually judged for the CHTR of the Second NTCIR Workshop though originally proposed as above. Even though three different assessors were selected from similar user groups, to see the consistency among them is one of the approach to assess the topical relevance as Dunlop pointed out [14] and similar to those done for Cystic Fibrosis [19].

Additional Information

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.

In the NTCIR-1, situation-oriented relevance judgments were conducted based on the statement of "purpose of search" or "background" in <NARRATIVE> in each topic, as well as topic-oriented relevance judgments, which are more common in ordinary IR systems laboratory testing. However, only topic-oriented judgments are used in the formal evaluation of this Workshop.

Rank-Degree Sensitive Evaluation Metric on Multi-grade Relevance Judgments

It is obvious that we need a metric that is sensitive to the degree of relevance of the documents and their rank in the ranked list of the retrieved documents. Intuitively, highly relevant documents are more important for users than partially 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 the more relevant documents are given higher ranks in the ranked list should be rated as better.

Multi-grade relevance judgments are used in several test collections such as Cystic Fibrosis [19] and OHUMED [20], although specific evaluation metrics for them were not produced for the collection. We examined the several rank-degree sensitive metrics had been proposed so far, including Average Search Length [21], Relative Relevance and Ranked Half-Life [22], and Cumulated Gains [23], and tested Weighted Mean Average Precision [25] on the JEIR results.

Most of IR systems and experiments have assumed that only the most highly relevant items are useful to all users. However some user-oriented studies have suggested that partially relevant items may important for a specific user group and partially relevant items should not be collapsed into relevant items, but should be analyzed separately [13]. More analysis and investigation shall be needed.

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 also 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. [24]

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 of these studies have been reported and published on various occasions [26-29]. As a result, 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 [26-28, 30]. It served as an additional support to the analysis reported by Voorhees [31].

3.6 Differences between CHTR and JEIR

CHTR and JEIR were organized rather an independent way but we aimed to follow the consistent or at least compatible procedures each other. It was mainly for the ease of the participants who participated both tasks and to avoid confusion. However regrettably we could find unintended incompatibility between CHTR and JEIR including names of topic fields, labels of relevance degree, pooling methods and query types (combination of topic fields used in a search). It was probably the emphasis had been placed different aspects by each task organizer and we have to understand the difference in order to examine the evaluation results properly.

Among them, inconsistency of the names of topic fields and relevance degrees are not problematic since the concepts expressed by them are equivalent and the difference did not affect the evaluation results. However the difference in query types and pooling methods may affect the evaluation results. Some of the differences were found after Workshop Meeting was over, and then, to avoid the confusion, these differences should be indicated below. Influence may caused by these


difference should be analyzed and investigated further.

Table 4 shows the query types, or combination of topic fields used in a search. In NTCIR Workshop 1 (Ad Hoc and CLIR Tasks) and JEIR of NTCIR Workshop 2 have placed emphasis to enhance the comparison between systems. Therefore they set "mandatory run", a search only using <DESCRIPTION> of the topic for automatic query construction. They are also keen to the difference between search using <CONCEPT> or without it since it has been known that searches using <CONCEPT> generally obtain better results than those without them from experiences of earlier TREC and NTCIR Workshop 1. Whereas CHTR placed emphasis on the "length" of the query. CHTR and NTCIR Workshop 1 used similar terminology of query types but the contents, the actual topic fields used in the search may differ between the two. To avoid the confusion, JEIR in NTCIR Workshop 2 used combination of topic fields to express the query types rather than using terms like "Very short", "Short", or "Long".

Table 4. "Query Types" in CHTR and JEIR

workshops	NTCIR WS1	NTCIR WS2	
topic field(s) used*	Ad Hoc & CLIR (JE)	Chinese Text Retrieval	Japanese & English IR
T	Very Short	TI (Title)	T
TD			T+D
D	Short without Concept	SO (Short)	D only
C or DC or TDC	Short with Concept		C without N
TC		VS (Very Short)	
N or DN or TN or TDN	Long without Concept		N without C
NC or DNC or TNC or	Long with Concept	LO (Long)	N+C

* where T=TITLE, D=DESCRIPTION
(QUESTION in CHTR), N=NARRATIVE,
C=CONCEPT

 : mandatory for automatic query construct

Regarding the pooling methods, both JEIR and CHTR started from the same scripts for pooling and instructions that were used in IR tasks of the First NTCIR Workshop. JEIR reduced the number of top ranked documents collected from each submitted run to reduce the size of document pool with keeping the number of documents collected from each run for a specific topic are the same across the systems. Whereas according to the final task report [31], CHTR reduced the pool size by

discarding the documents searched by only one system.

4. Discussion and Future Directions

4.1 Round Tables at the Workshop Meeting

In the round table sessions at the Second NTCIR Workshop Meeting, the following tasks were proposed for the Third NTCIR Workshop.

- (1) Multilingual CLIR Task
- (2) Web Challenge
- (3) Patent IR Challenge
- (4) Text Summarization Challenge (TSC)
- (5) Q & A Challenge (QAC)

Based on the round table discussion, each task organizers and discussion groups started to prepare the tasks. The content of the discussion regarding each task can be partly available in task overviews in this volume and later will be circulated through the Mailing List of the discussion group. The preliminary call for participation will be circulated in July 2001. For details, please subscribe ntcir@nii.ac.jp mailing list and consult the NTCIR web site at <http://research.nii.ac.jp/ntcir/>.

4.2 Future Directions

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

- (1) Evaluation of CLIR systems
- (2) Evaluation of retrieval of new document genres and more realistic evaluation
- (3) 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.

With respect to new genres, we are especially interested in Web documents and multimedia documents. For these document types, the user group, usage, and purpose of search, and the criteria for successful retrieval are quite different to

those for traditional text retrieval, and the investigation of these aspects is challenging.

Regarding (3), we should look at the interaction between systems and users not only the text processing technology, such as extraction, summarization, or Q&A.

ACKNOWLEDGMENTS

We thank all the participants for their contributions, the assessors and the program committee.

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] Sung, H.M. "HANTEC Collection". Presented at the panel on IR Evaluation in the 4th IRAL, Hong Kong, 30 Sept.-3 Oct. 2000.
- [5] Kando, N.: Cross-Linguistic Scholarly Information Transfer and Database Services in Japan. Annual Meeting of the ASIS, Washington DC. Nov. 1, 1997
- [6] TREC URL: <http://trec.nist.gov/>
- [7] Smeaton, A., Harman, D. K. "The TREC experiments and their impact on Europe", *Journal of Information Science*, Vol.23, No.2, pp.169-174, 1997.
- [8] Sparck Jones, K., Rijsbergen, C. J." Information retrieval test collections", *Journal of Documentation*, Vol.32, No.1, pp.59-72, 1975.
- [9] Panel of IR Evaluation of the World. RIAO 2000, Paris, France, April 2000.
- [10] TDT URL: [http://www.nist.gov/speech/Selecting "benchmark tests" and then "TDT"](http://www.nist.gov/speech/Selecting%20benchmark%20tests%20and%20then%20TDT).
- [11] CLEF URL: <http://www.iei.pi.cnr.it/DELOS/CLEF/>
- [12] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/
- [13] 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
- [14] Dunlop, M.D. Reflections on Mira, *Journal of the American Society for Information Sciences*, Vol.51, No.14, pp.1269-1274, 2000
- [15] Spink, A., Greisdorf, H. Regions and levels: Measuring and mapping users' relevance judgments. *Journal of the American Society for Information Sciences*, Vol.52, No.2, pp.161-173, 2001
- [16] Campbell, I., Interactive evaluation of the ostensive model using a new test collection of images with multiple relevance assessments, *Information Retrieval*, Vol.2, No.1, pp.87-114, 2000
- [17] Reid, J. A task-oriented non-interactive evaluation methodology for information retrieval systems, *Information Retrieval*, Vol.2, No.1, pp 115-129, 2000
- [18] Chiang, Yu-ting: A Study on Design and Implementation for Chinese Information Retrieval Benchmark. Master Thesis, National Taiwan University, 1999, 184 p.
- [19] Shaw, W.M., Jr, et al.: The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research*, Vol.13, pp.347-366, 1991.
- [20] Hersh, W., Buckley, C., Leone, T.J., Kichman, D.H.: OHSUMED: an Interactive Retrieval Evaluation and New Large Test Collection for Research. In *Proceedings of 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. p.192-201, Dublin, Ireland, 1994.
- [21] Losee, R.M.: Text retrieval and filtering: analytic models of performance. Kluwer, 1998
- [22] Borlund, P., Ingwersen, P.: Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In *Proceedings of 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. p.324-331, Melbourne, Australia, August. 1998.
- [23] Jarvelin, K., Kekalainen, J.: IR evaluation methods for retrieving highly relevant documents. In *Proceedings of 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. p. 41-48, Philadelphia, PA, USA, July 2000.
- [24] Yoshioka, M., Kuriyama, K., Kando, N.: Analysis on the Usage of Japanese Segmented Texts in the NTCIR Workshop 2. In *NTCIR Workshop 2 : Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, June 2000- March 2001 (ISBN : 4-924600-96-2) (to appear)
- [25] Kando, N., Kuriyama, K., Yoshioka, M. Evaluation based on multi-grade relevance judgements. *IPSJ SIG Notes*, July 2001 (to appear)
- [26] 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].
- [27] 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].
- [28] 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].
- [29] 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.
- [30] 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) (to appear)
- [31] 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