Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop

Noriko Kando, Kazuko Kuriyama, Masaharu Yoshioka
National Institute of Informatics (NII), Japan
{kando, kuriyama, yoshioka}@nii.ac.jp

Abstract

This paper is a report of the Japanese and English Information Retrieval Tasks (JEIR) at the second NTCIR Workshop. There are six subtasks in the JEIR: Japanese monolingual retrieval (J-J), English monolingual retrieval (E-E), and four types of cross-lingual information retrieval among Japanese and English (J-E, E-J, J-JE, E-JE, where the left character shows the topic language and the right characters show the document languages). The paper discusses the scope, schedule, test collections used (NTCIR-1 and NTCIR-2), search results, evaluation and initial analyses of search results of the JEIR.

1. Introduction

This paper serves as an introduction to the Japanese and English Information Retrieval Tasks (JEIR) at the second NTCIR Workshop[1] [1] and research using the NACSIS-NII Test Collections 1 and 2 (NTCIR-1 and NTCIR-2), which are described in detail in the rest of the volume. Information on the participating groups and their systems can be found in the individual group reports in the volume, also available from the NTCIR web site.

There are two sub-categories, Monolingual Information Retrieval (MonoLIR) and Cross-Lingual Information Retrieval (CLIR). Each participating group conducted one or more of the tasks listed below:

MonoLIR including;

Retrieval of Japanese documents by Japanese search topics (J-J)

Retrieval of English documents by English topics (E-E)

CLIR including;

Retrieval of Japanese documents by English topics (E-J)

Retrieval of English documents by Japanese topics (J-E)

Retrieval of a collection of a mixture of Japanese documents and English documents by either of Japanese topics (J-JE) or English topics (E-JE).

The relationships of tasks, topics, and documents are shown in Fig.1.

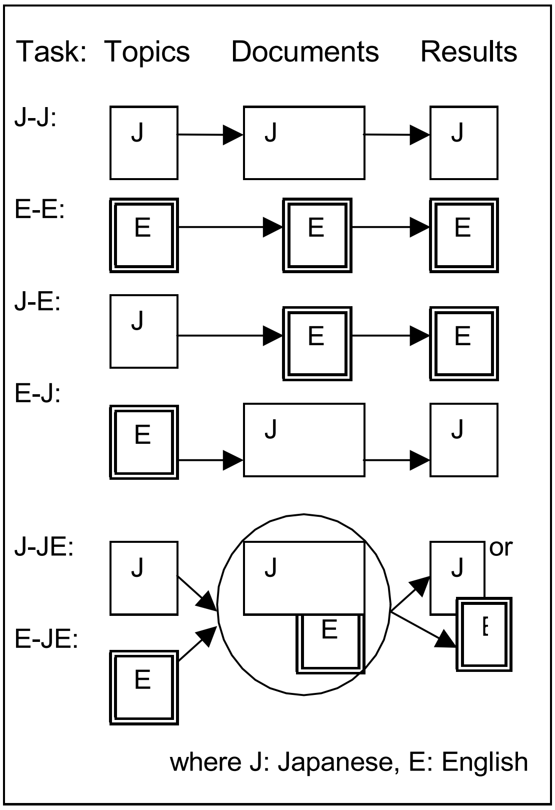

Fig. 1 Collections and Subtasks

The ad hoc IR Task is a retrieval of a mixture of Japanese, English and Japanese-English paired documents by Japanese search topics, since retrieving such a mixture is natural in the operational environment in Japan, and the CLIR Task is ad hoc retrieval of English Documents by Japanese topics. In the first NTCIR Workshop [2], many participants did the ad hoc Task as a Japanese monolingual task by discarding English documents and the English parts of documents. Therefore, the task definition changed to specify the language explicitly and divided paired documents into couples of single language documents in the second workshop.

The J and E Collections in "Test Collection 1

[1] This project is supported by "Research for the Future" Program JSPS-RFTF96P00602 of the Japan Society for the Promotion of Science

(NTCIR-1)" are used for training. The number of document records and size are 332,918 (312MB) and 187,080 (218MB), respectively. They also contain 83 topics, and their relevance judgments. More than half of the documents are English-Japanese paired (document alignments). The NTCIR-1 JE Collection containing 339,483 documents (577MB) was not used in the second workshop.

A new collection, the NTCIR-2 (preliminary version) CD-ROM was distributed to the participants in August 2000. It contains 403,248 Japanese documents (600MB), 134,978 English documents (200MB), and 49 topics. The test is to investigate the search effectiveness for the retrieval of the documents of NTCIR-1 and -2 by 49 new topics in the NTCIR-2. The results are submitted as the ranked top 1000 documents retrieved for each topic.

Thirty-one groups from six countries enrolled to participate in the JEIR tasks in the second NTCIR Workshop. Among these, 22 groups enrolled in Monolingual IR tasks and 17 in the CLIR tasks. Table 1 shows the distribution of groups that enrolled and submitted search results. As shown in Table 2, the search results of 206 runs from 25 groups were submitted.

Table 1. Distribution of Participants

	Enrolled		Submitted	
	MLIR	CLIR	MLIR	CLIR
Canada	1	0	1	0
Japan	17	14	8	8
Korea	0	0	1	1
Taiwan	1	1	2	1
UK	1	0	2	1
USA	2	2	4	3
Total	22	17	17	14

Table 2. Participants for JEIR Tasks

	MLIR			CLIR					
	J-J	E-E	sum	J-E	E-J	J-JE	E-JE	sum	total
#Groups Enrld	22	11	22	16	14	11	11	17	31
#Groups Submitted	17	7	17	12	10	6	4	14	25
#Submitted Runs	93	18	111	40	30	14	11	95	206

Table 3 shows the combination of tasks among the groups submitting search results. Eleven groups did Monolingual IR subtasks only, eight groups did CLIR subtasks only, and six groups did both. Among these, four groups conducted tasks without any Japanese language expertise.

In the next section, we describe the tasks performed for the JEIR Tasks. Section 3 shows the test collections (NTCIR-1 and NTCIR-2) used for the tasks. Section 4 describes the methods of evaluation and Section 5 analyzes the search results submitted by participants. The final section lists issues to be discussed.

Table 3. Active Participants and Their Subtasks

MLIR		CLIR				
J-J	E-E	J-E	E-J	J-JE	E-JE	# Groups
*						9
*	*					2
		*				2
			*			1
		*	*			1
		*	*	*	*	3
				*		1
*	*	*	*			4
*	*	*	*	*	*	1
*		*		*		1
17	7	12	10	6	4	25

2. Task Descriptions for JEIR Tasks

2.1 The Procedures

Schedule

***1 June 2000**:* Call for Participation in Tasks and distribution of the JEIR training data

***10 August 2000**:* distribution of the JEIR test data (new documents and 49 J/E topics)

***18 September 2000**:* submission of JEIR results

10 January 2001: distribution of CHTR and JEIR evaluation results

***7-9 March 2001**:* Workshop meeting

From 1 June 2000, delivery to each JEIR tasks-participant of the NTCIR-1 (document data, 83 search topics (0001–0083) and their relevance judgments began), so the participants could train their systems. Among them, 60 topics were used as cross-lingual topics. The topics were written in Japanese. New documents and 49 new test topics (0101–0149) in NTCIR-2 were distributed to participants on 10 August 2000, and the search results for these new topics were submitted from each participant by 18

September as official test runs. The test topics are common across all the subtasks in the JEIR.

For various reasons, some of the participating groups enrolled for the JEIR tasks could not submit search results by 18 September. We set a second due date and accepted search results from three who wished to continue as active participants. These additional submissions were included in the evaluation results but were not included in the document pool for relevance assessment by human analysts.

A participant could submit the results of more than one run. Both automatic and manual query constructions were allowed. In the case of automatic construction, the participants were required to submit at least one set of results for searches using the <DESCRIPTION> fields of the topics only as the mandatory runs. For optional automatic runs and manual runs, any fields of the topics could be used. Each participant could set the objective of the task and test various approaches. In addition, each participant was required to complete and submit a form describing the detailed features of their system.

Human analysts assessed the relevance of retrieved documents for each topic. Based on the relevance assessments, interpolated recall and precision at 11 points, average precision (non-interpolated) over all relevant documents, and precision at 5, 10, 15, 20, 30, 100 documents were calculated using TREC's evaluation program, which is available from the ftp site of Cornell University.

3. The Test Collection

The test collections used in the workshop, "NTCIR-1" and "NTCIR-2", consist of documents, topics, and relevance assessments for each search topic. The relationship between the training and test sets is shown in Fig. 2.

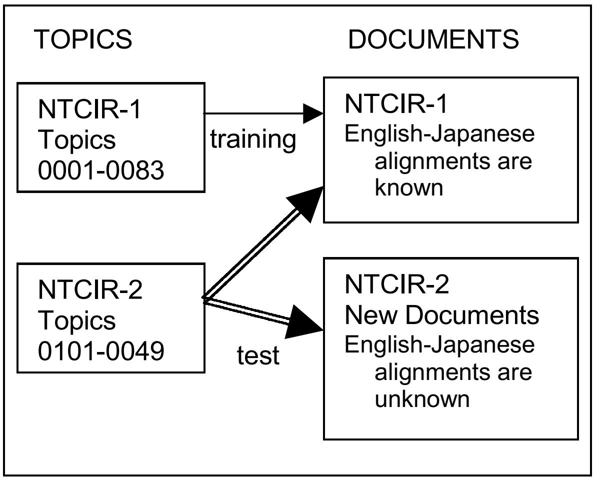
TOPICS
DOCUMENTS
NTCIR-1 Topics 0001-0083
training
NTCIR-1 English-Japanese alignments are known
NTCIR-2 Topics 0101-0049
test
NTCIR-2 New Documents English-Japanese alignments are unknown

Fig. 2 Training and Test Sets

3.1 Documents

Training Set

"*E Collection* (clir/ntc1-e1.tgz)" and "*J Collection* (mlir/ntc1-j1.tgz)" in the *NACSIS Test Collection 1 (NTCIR-1)* CD-ROM were used as training data sets. The document data set "JE Collection (adhoc/ntc1-je1.tgz)" was not used in the NTCIR Workshop 2. The training topics in the NTCIR-1 (topics.tgz) were in Japanese only.

Documents in the NTCIR-1 were part of the documents extracted from "*NACSIS Academic Conference Paper Database*". These are author abstracts of the papers presented at conferences hosted by 65 academic societies in Japan. More than half of the documents are produced as English-Japanese paired (Document level alignments). "*J Collection*" contains Japanese documents and was constructed by extracting the Japanese parts of the documents from the original database, and "*E Collection*" contains the English documents. Most of the documents in the "*E Collection*" have equivalent documents in the "*J Collection*" but some do not. In several cases the relevance judgments are different between paired English and Japanese documents since the detailed expressions are different between them and the differences affected the judgments.

The equivalent pairs of documents in "*E Collection*" and "*J Collection*" share the same ACCNs (document IDs). These alignments are usable for the training, and for constructing bilingual lexicons, or knowledge bases, used for cross-lingual IR.

Test Set

Documents in the *J Collections* and *E Collections* of NTCIR-1 and NTCIR-2 are used for test purposes. The new document set, NTCIR-2 (preliminary version) also contains *J Collection* and *E Collection*. The test topics are both in Japanese and English. Documents in the NTCIR-2 contains two sub-files:

(1) extended summaries of Grant-in-Aid research reports (287,071 Japanese documents and 57,545 English documents), and

(2) author abstracts of conference papers (116,177 Japanese documents and 77,433 English documents).

About 25% of (1) and more than half of (2) are Japanese-English paired (document alignments). The alignments were not announced during the second workshop, but are included in the "NTCIR-2 Research Purpose Version" CD-ROM. The average length of the documents in (1) is about three times that of the documents in NTCIR-1. The statistics are available in Table 4.

Segmented Japanese Texts

Segmented Japanese texts were available as well as non-segmented ordinary Japanese texts for both Japanese documents and Japanese topics in NTCIR-1 and NTCIR-2. In the segmented texts, each sentence is segmented into terms and term components (similar to phrases and words); use of this data set was optional. The purposes are: (1) to enhance the cross-system comparison by providing a baseline of Japanese text segmentation, and (2) to encourage the participation of non-Japanese research groups.

Segmentation was done using a commercially available Japanese morphological analyzer called *HAPPINESS/BASE3.5* (Heiwa Information Center Co. Ltd.), which has been used by several operational Japanese IR service systems, and can be thought of as a readily available technical level of Japanese text segmentation for IR purposes. (See Fig. 3a).

```
<REC>
<ACCN>kaken-j-0924516300</ACCN>
<YEAR>1992</YEAR>
<SBJ1 TYPE="kanji">802 : 情報学</SBJ1>
<PJNM TYPE="kanji">文献 の 論理__構造 に 基づく
全文__データベース__検索__システム の 開発__研究
</PJNM>
<ABST TYPE="kanji"><ABST.P>本__研究 は 、 学術__
文献 など の 文書 の 全文 を 収容 する 全文__
データベース に ついて 、 それら の 文書 の 論
理__構造 に 即した 検索 を 可能 と する システ
ム を 研究 ・ 開発 しよう と する もの で あ
る 。 3__年次 に わたって 下記 の 項目 に つ
いて 研究 および 開発 を 行なった 。
</ABST.P><ABST.P>1 . 全文__データベース に 対する
検索__要求 の 詳細__分析 を 行ない 、 SGML の
文書型__定義 に 基づいて 検索 ・ 表示__要求 を
効率的 に 記述 する ため の 表記__形式__DQL
( Document Query Language ) の 詳細__設計 を 行な
った 。 SQL を 拡張 し 、 文書__構造 を 扱う
ため の 記述 を 可能 に した 。
</ABST.P><ABST.P>2 . 文献 の 文書__構造 を 図形
的 に 表示 し 、 要素 を ポインティングデバイス
で 指定 して 検索__条件 ・ 表示__指示 を 行なう
ユーザ系 の ソフトウェア を 設計 し 、 ワークス
テーション__上 で グラフィカルユーザインタフェース
( GUI ) を 用いて 開発 した 。
</ABST.P><ABST.P>3 . 文書__構造 を 各__構成__要素
__間 の 二項__関係 で 関係__データベース__管理__シ
ステム__上 に 表現 し 、 DQL で 記述 された
検索__要求 を 処理 する サーバ系 の ソフトウェア
を 汎用__大型__計算機__上 に 開発 した 。
</ABST.P><ABST.P>4 . サーバ系 と ユーザ系 の 接
続__方式 を 開発 し 、 LAN および ISDN を
介して 連動 させて 動作 を 確認 した 。
</ABST.P><ABST.P>5 . 全体的 な 処理__性能 、 使
い__心地 、 検索__精度 など に ついて 評価 を
行ない 、 実用 に 向けて の 課題 と 解決__方式
を 検討 した 。</ABST.P><ABST.P>現在 の システ
ム には サーバ系 の 性能 に 改善 の 必要性 が
認められ 、 検討 の 結果 、 二項__関係 に 参照
先 の レコード__ID を 含める こと が 有効 で
ある こと が わかった ので 、 今後 これ を 実
現 する こと が 課題 と なる 。 また 、 ユー
ザ系 に おいて は 指定 した 検索__要求 を より
わかりやすく 表示 する 必要 が ある こと が 明
か と なり 、 考案 した いくつか の 方式 に
ついて 実験 に より 検討 する こと が 課題 と
なる 。</ABST.P><ABST.P>以上 の 結果 、 本__シ
ステム の 設計__概念 、 および 実現__方式 の 妥
当性 が 確認 でき 、 課題 へ の 対処 の 方針
も 示す こと が できた ことに より 、 実用化
の 可能性 が 示された 。</ABST.P></ABST>
<KYWD TYPE="kanji">全文__データベース / 情報__検索 / 文
書__構造 / SGML / GUI / 分散__処理</KYWD>
</REC>
```

Fig. 3a. A Sample Document Record (Segmented Data, NTCIR-2 Grant-in-Aid Report Sub-file)

```
<REC>
<ACCN>kaken-j-0924516300</ACCN>
<YEAR>1992</YEAR>
<SBJ1 TYPE="kanji">802: 情報学</SBJ1>
<PJNM TYPE="kanji">文献の論理構造に基づく全文データベ
ース検索システムの開発研究</PJNM>
<ABST TYPE="kanji"><ABST.P>本研究は、学術文献などの
文書の全文を収容する全文データベースについて、それらの
文書の論理構造に即した検索を可能とするシステムを研究・
開発しようとするものである。3 年次にわたって下記の項目
について研究および開発を行なった。</ABST.P><ABST.P>1.
全文データベースに対する検索要求の詳細分析を行ない、
SGML の文書型定義に基づいて検索・表示要求を効率的に記
述するための表記形式 DQL(Document Query Language)の詳
細設計を行なった。SQL を拡張し、文書構造を扱うための
記述を可能にした。</ABST.P><ABST.P>2.文献の文書構造を
図形的に表示し、要素をポインティングデバイスで指定して
検索条件・表示指示を行なうユーザ系のソフトウェアを設計
し、ワークステーション上でグラフィカルユーザインタフェ
ース(GUI)を用いて開発した。</ABST.P><ABST.P>3.文書構
造を各構成要素間の二項関係で関係データベース管理システ
ム上に表現し、DQL で記述された検索要求を処理するサー
バ系のソフトウェアを汎用大型計算機上に開発した。
</ABST.P><ABST.P>4.サーバ系とユーザ系の接続方式を開発
し、LAN および ISDN を介して連動させて動作を確認した。
</ABST.P><ABST.P>5.全体的な処理性能、使い心地、検索精
度などについて評価を行ない、実用に向けての課題と解決方
式を検討した。</ABST.P><ABST.P>現在のシステムにはサ
ーバ系の性能に改善の必要性が認められ、検討の結果、二項
関係に参照先のレコード ID を含めることが有効であること
がわかったので、今後これを実現することが課題となる。ま
た、ユーザ系においては指定した検索要求をよりわかりやす
く表示する必要があることが明かとなり、考案したいくつか
の方式について実験により検討することが課題となる。
</ABST.P><ABST.P>以上の結果、本システムの設計概念、
および実現方式の妥当性が確認でき、課題への対処の方針も
示すことができたことにより、実用化の可能性が示された。
</ABST.P></ABST>
<KYWD TYPE="kanji">全文データベース / 情報検索 / 文書構
造 / SGML / GUI / 分散処理</KYWD>
</REC>
```

Fig. 3b. A Sample Document Record (J Collection, NTCIR-2 Grant-in-Aid Report Sub-file)

A sample of the document record

Documents are SGML-tagged plain text. A record in NTCIR-1 and the Conference Paper Sub-file in NTCIR-2 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 record in the Grant-in-Aid Report Sub-file in NTCIR-2 may contain document ID, title, classification code, abstract, keyword(s) that were assigned by the author(s) of the document (See Fig. 3b, 3c).

```
<REC>
<ACCN>kaken-e-2469487463</ACCN>
<YEAR>1992</YEAR>
<SBE1 TYPE="alpha">802: *</SBE1>
<PJNE TYPE="alpha">Development of a Full Text Retrieval System based on Logical Structure of Documents</PJNE>
<ABSE TYPE="alpha"><ABSE.P>The investigators have conducted research and development of a system with which users can retrieve documents from a full text database containing full documents such as academic papers. Results obtained were as follows: </ABSE.P><ABSE.P>1. Based on detailed analysis of retrieval requests for full text databases, they made detailed design of a formal notation DQL (Document Query Language) for describing retrieval and display requests efficiently according to the document type definition of SGML. </ABSE.P><ABSE.P>2. They designed a user system software which displays document structures and let users select elements with pointing devices and specify retrieval conditions and display instructions. The system was developed on work-stations using graphical user interface systems. </ABSE.P><ABSE.P>3. They developed a server system software on a main frame computer which stores documents in a relational database management system by expressing the structure as binomial relations among elements and processes retrieval requests described in DQL. </ABSE.P><ABSE.P>4. They developed a communication method between sever and user systems and confirmed the functionality by experiments using LAN and ISDN. </ABSE.P><ABSE.P>5. They evaluated the overall performance, usability, retrieval precision etc and considered the problems and their solutions toward practical use.</ABSE.P><ABSE.P>Through evaluation, the necessity of performance improvement of the server system was revealed. Further investigation has made it clear that including referenced record identifiers within the binomial relation records is effective. A future issue is to implement this method. The necessity for the user system to represent specified retrieval requests more understandably was also revealed. Several methods already proposed should be studied by experimenters.</ABSE.P><ABSE.P>As the result of the research, the feasibility of the design concept and the implementation of this system was confirmed, the approach to existing problems was presented, and the reality of the full-fledged system was shown.</ABSE.P></ABSE>
<KYWE TYPE="alpha">Full Text Database / Information Retrieval / Document Structure / SGML / GUI / Distributed Processing</KYWE>
</REC>
```

Fig. 3c. A Sample Document Record (E Collection, NTCIR-2 Grant-in-Aid Report Sub-file)

Since one of the purposes of the original database is to provide an alerting service for papers presented in Japanese academic conferences as soon as possible, documents are put in the database without any revision or modification by professional abstractors or editors. Some of them are refereed, and others are pre- or non-refereed. As part of the philosophy of leaving the data as close to the original as possible, and because it is impossible to check all the data manually, there are "errors" in the data. These range from errors in the original data or other typographical errors, to errors in the reformatting undertaken by NII and the NTCIR Project Group. The error checking has concentrated on allowing readability of the data rather than on correction.

The comparisons of the length of the free-text part of each Japanese text collection are shown in Table 4 and Fig. 4. Details reported in [3]. For IREX see [4].

Table 4. Length of Document

	min	max	average	stdev
NTCIR-1 Conference Paper Abstracts				
J Collection	0	8,003	521.7	224.0
E Collection	1	4,363	531.9	254.9
NTCIR-2 Grant-in-aid Report Sub-file				
J Collection	5	17,111	1,366.8	291.4
E Collection	15	20,396	1,614.6	492.7
Newspaper articles (Japanese only)				
IREX	20	15,770	895.9	977.8

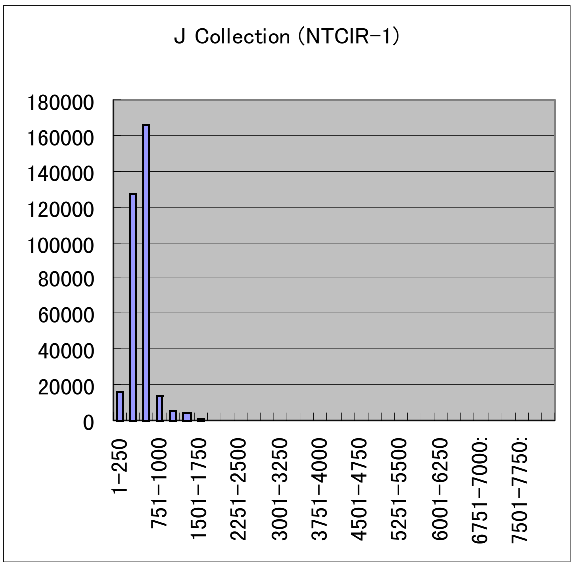

Fig. 4. Length of Free-Text Part of Japanese Document Collections

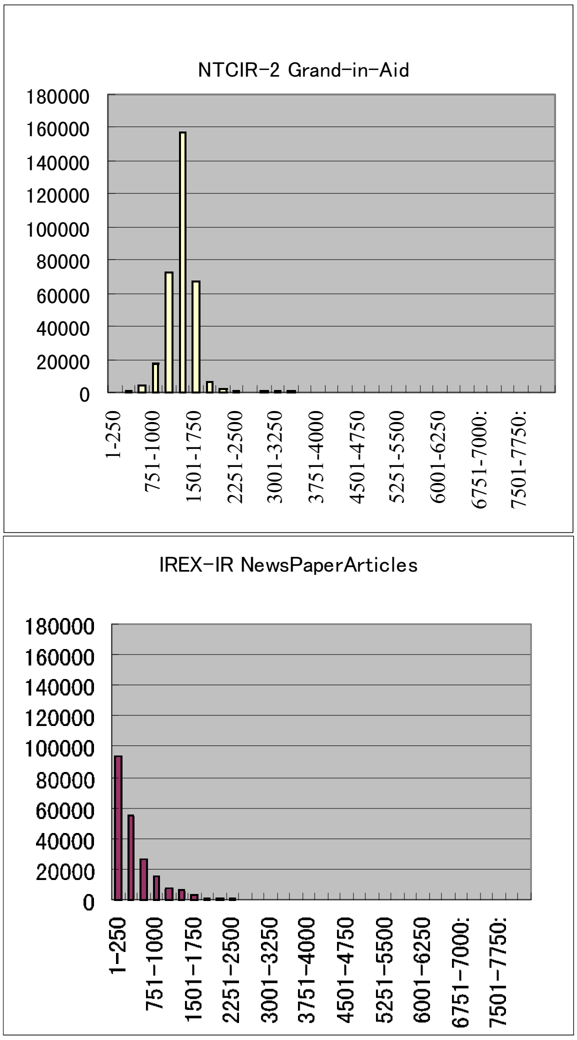

Fig. 4. Length of Free-Text Part of Japanese Document Collections (Cont'd)

3.2 Topics

A topic is a formatted description of a user's information needs. We defined the topics as statements of "user need" rather than "queries", which are the strings actually submitted to the system, since we desire to allow both manual and automatic query construction from the topics.

The query format is similar to that used in TREC-1 and-2 and contains SGML-like tags. A query consists of a title of the topic (T), a description (D), a detailed narrative (N), a list of concepts (C) and field(s) (F). 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. (See Fig. 5)

Topic Preparation

Topics were created by the assessors, who are the users of the document type of the subject domain, i.e., researchers (mainly graduate students) and specialized professional information specialists with sufficient subject knowledge of the subject domain. In the first workshop some of the topics were collected from researchers or library reference users, and then modified by assessors. The relevance assessments of those topics originally not created by the assessors themselves are difficult and time-consuming. Then all the topics in NTCIR-2 were created by assessors based on their information needs. Assessors' backgrounds are computer science, physics, chemistry, microbiology, architecture and civil engineering, social sciences, such as education, linguistics, library and information science, mass communication, etc.

```
<TOPIC q=0001>

<TITLE>
bibliometrics
</TITLE>

<DESCRIPTION>
Are there any documents of bibliometrics that deal with the
proper treatment of the types which are unseen in the given
sample?
</DESCRIPTION>

<NARRATIVE>
There are many studies that deal with the mathematical
structure of given samples in the field of bibliometrics. I
wonder whether there are studies that deal with the unseen
types. Especially, I would like to know the extension of
Lotka's law or Bradford's law to the theoretical population.
</NARRATIVE>

<CONCEPT>
bibliometrics, bibliometrics, population, sample, Lotka's
law, Bradford's law
</CONCEPT>

<FIELD>
1.Electricity, information, and control,
8.Cultural, information, and social science
</FIELD>

</TOPIC>
```

Fig. 5 A sample Topic (English Translation)

The NTCIR-1 and NTCIR-2 contains 83 and 49 topics, respectively. Among 83 NTCIR-1 topics, 60 are usable for cross-lingual retrieval. NTCIR-1 topics are written in Japanese. In NTCIR-2, Japanese topics and English translations are available. Among 83 NTCIR-1 topics, 20 are used in the HANTEC Korean Test Collection [5].

After collecting the topics, each topic was examined for its clarity and difficulty by the analysts and project members in NII. The criteria are as follows.

(1) Statements of "user need" rather than "queries"

(2) <Description> containing every concept needed to describe the topic

(3) Not too easy:

(3-1) Simple word matching of query terms cannot retrieve every relevant document

(3-2) A document containing query terms can be irrelevant

(4) Five or more relevant documents

The function category of each topic was analyzed and assigned based on the function category proposed by BMIR [6]. The category indicates the required level of techniques and knowledge to conduct a search of the topic. Among the above conditions, (3) is not mandatory and we tried to balance the topic length, number of relevant documents, and "difficulty". Some analysis of "topic difficulty" is found in [7].

3.3 Relevance Judgments (Right Answers)

The relevance judgments were undertaken by pooling methods. A certain number of top-ranked documents were collected from each submitted run and these created a pool of possibly relevant documents. Human analysts assessed the relevance of each document in the pool against the topic. The relevance assessment was undertaken using four grades: highly relevant (S), relevant (A), partially relevant (B), irrelevant (C). 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.

Since we had many submissions, compared to the limited work force for the assessments, selected numbers of the runs were contributed to the document pool. We chose the same number of runs from each participating group and the same number of top ranked documents from each run for the topic, in order to retain the "fairness" and "equal opportunities" among each participating group. A detailed description of the pooling procedure and the analysis of "fairness" are reported in Kuriyama et al. [8] in this volume.

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". The relevance judgments are often called "right answers" in the Japanese IR community.

4. Evaluations and Examples of Discussion Points

Based on the relevant assessment, the following measures were calculated and are reported in the appendix of the proceedings of the Second NTCIR Workshop Meeting [9], also available from the NTCIR web site:

Non-interpolated average precision over all relevant documents,

Recall-level Precision: Interpolated recall and precision at 11 points,

Document-level Precision: Precision at 5, 10, 15, 20, 30, 100 documents, and

R-precision: precision after the "R (the total number of the relevant documents for the topic)" documents are retrieved

These were calculated for each submitted result set, for each topic and the mean value over all topics.

We would like to emphasize here that the aims of the workshop were to provide a forum for IR researchers interested in comparing results, and exchanging ideas, experiences, or opinions in an informal atmosphere, and to encourage research in IR and cross-lingual IR by providing test collections. Therefore, we expected that various approaches would be proposed and tested in this workshop, and that this workshop would encourage intensive discussion through the mailing list of participants and at the workshop meeting. At the same time, we also desired to improve the quality of the test collection and explore the possibility of the resources available for the evaluation of various IR systems based on feedback, comments, advice, and leads from participants.

For example, it was expected that further insights into Japanese text retrieval would be explored, with regard to the following points. These were examples of the interests proposed at the call for participation for the task, and not limited to:

(i) Appropriate algorithms and parameters for Japanese text retrieval

(ii) Relationship between text segmentation and retrieval algorithms

(iii) Retrieval of mixed-language texts and English terms in Japanese texts

(iv) Application for interactive systems

(i) Appropriate Algorithms and Parameters for Japanese Text Retrieval

In the NTCIR Workshop 1, the algorithms and parameters, which are known to be effective on English text, were applied to Japanese texts. This was partly because of the limited time for the training period. Various sectors would be interested in "good algorithms and parameters for Japanese text retrieval", or "are the algorithms which are good for English texts the best for Japanese texts?". Various challenges were expected.

In the first and second NTCIR Workshops, several new models have been proposed, including the Relevance-based Super Imposition Model by the R2D2 group, Coordination level scoring (CLS) by the Matsushita group, flexible pseudo-relevance feedback by the CAMUK group, LSI with multiple semantic spaces by the Forst group, and the sstut group proposed a new weighting scheme similar to Berkeley's regression model. The SRGDU group has looked at the modification of the probabilistic model with logistic regression for Japanese text retrieval. In addition to those, various techniques have been proposed, including transliteration, various approaches to construct bilingual dictionaries from the paired documents and so on.

Many groups who participated in both the first and second workshops modified and improved their systems in various ways based on their continuous efforts and experiments using the NTCIR collection. Generally, the retrieval performance of these systems has been improved, although of course we cannot directly compare the performances using different test collections. For example, the LISIF group (ULIS group in the first workshop) conducted a rather drastic change of the system from one based on a vector space model to a probabilistic model.

The list of the publications used in the NTCIR-1 and -2 collections are available at the NTCIR Website, http://research.nii.ac.jp/ntcir/paper-en.html. It is based on the reports from the author(s) of each paper and not an exhaustive list, but covers certain parts of the research and the continuous efforts of the NTCIR participants.

(ii) Relationship between Text Segmentation and Retrieval Algorithms

Many researchers have focused on the segmentation of Japanese texts. Someone said that the bi-gram is the best, and others said the word- and phrase-based indexing is the best. Each system uses its segmentation and its algorithm; thus, the cross-system comparison becomes complicated. Therefore, the segmented Japanese texts of documents and topics were prepared for this workshop.

The purpose of the segmented Japanese texts were: (1) to encourage the participation of non-Japanese research groups, (2) to investigate the effects of the segmentation methods on the search effectiveness, and (3) to encourage the comparison of the retrieval algorithms, minimizing the effects of the segmentation. The use of the segmented texts was optional.

The problems of the segmented texts include not being usable to examine the following issues:

(1) the effect of the tuning of the segmentation methods according to the documents to be retrieved,
(2) the effect produced from the beneficial combination of the segmentation and retrieval algorithms,
(3) the index structure, such as pat trie, not depending on the segmentation.

To cancel the effects of these issues, as much as possible, we asked for the submission of the final query term lists with retrieved results for further analysis. This was done on a voluntary basis and only one group has submitted a full set of results using segmented data. Regardless of these disadvantages, we prepared them as one of the attempts to investigate the characteristic aspects of Japanese text retrieval. The preliminary analysis of the effect of segmentation for the IR is reported by Yoshioka et al. [10] in this volume. Further comments and discussion are welcome.

(iii) Retrieval of Mixed-Language Texts and English Terms in Japanese Texts

The document collections, including ordinary life documents like web documents, produced in Japan are naturally a Japanese and English mixture. Some English documents are produced as a paired document with a Japanese one; some are summaries of sets of Japanese documents; others are produced on their own. Moreover, Japanese documents naturally contain English or other foreign language terms with original spellings or as transliterated forms using Japanese phonetic characters, KATAKANA. Such English or other foreign language terms are often newer concepts or specific technical terms, which are important as search keys, but rarely listed in the ordinary lexical resources such as dictionaries and thesauruses. What approaches will be effective for such an environment of Japanese documents? Further insight into the matter was expected to be found through discussion in this workshop.

In the first NTCIR Workshop the ULIS group proposed to use "transliteration". It seemed to work well on their system, and they continued its use in the second workshop (ULIS changed name to LISIF). This is an example of a technique to overcome the problem.

(iv) Application for Interactive Systems

Interactive systems were welcome. Many proposals on and examinations about the applicability of laboratory-type testing to interactive systems were expected. Only four groups submitted the search results of interactive runs: DOVE, FXSD, smlab, and sstut, and they are all J-J monolingual runs.

In addition to this, some groups submitted results of automatic query constructions, but presented the systems with functions to enhance the browsing of the retrieved results or interaction between systems and users. For example, the LISIF system used clustering of the retrieved documents to support the users' browsing on the CLIR system.

We need to see how the collection was used on the interactive system and look at the possibility of organizing a subtask of interactive system evaluation. In the NTCIR-1 and -2, <NARRATIVE> in each topic may contain background of the topic, criteria and relevance judgments, purpose of search, term definition. It was intended to serve as a "pretended situation" of a user in interactive experiments [11-12]. Regardless of the subject specificity of the NTCIR-1 and -2 collections, the detailed <NARRATIVE> seemed to be useful for the test users who participated in the interactive experiments to understand the topic and to do relevance judgments during the interaction with the systems. For example, [12] reported the agreement of the group's test users and the NTCIR's judgments was over 70% and concluded that it was sufficient.

The interactive system on the CLIR in conjunction with the appropriate technology of the text processing was also requested from the various sectors.

5. Retrieval Results

This section reports an overview of the retrieval results from the aspect of system effectiveness, and analyzes some of the similarities and differences of the approaches taken by each participating group.

Table 5. Participants and Topic Types & Methods

	monolingual IR			CLIR					grnd total
	J-J	E-E	total	J-E	E-J	J-JE	E-JE	total	
Donly	16	6	16	11	9	5	4	11	20
T/TD	2	2	3	0	0	0	0	0	3
N noC	1	0	1	2	2	2	2	2	3
N+C	8	5	8	5	4	3	1	6	10
C	3	0	3	2	2	0	0	2	5
auto	16	7	16	11	10	6	4	13	24
inter	4	0	4	0	0	0	0	0	4
total	17	7	17	12	10	6	4	14	25

Table 6. Number of Submitted Runs and Their Topic Types and Methods

	monolingual IR			CLIR					grnd total
	J-J	E-E	total	J-E	E-J	J-JE	E-JE	total	
Donly	48	9	57	25	16	7	6	54	113
T/TD	5	2	7	0	0	0	0	0	7
N noC	8	0	8	4	4	4	4	16	24
N+C	27	7	34	9	8	3	1	21	55
C	5	0	5	0	1	0	0	1	6
auto	81	18	99	39	30	14	11	94	193
inter	12	0	12	0	0	0	0	0	12
total	93	18	111	40	30	14	11	95	206

Since one of the main purposes of the NTCIR Workshops is to enhance research in Japanese text retrieval and cross-lingual retrieval, examination of various approaches using the NTCIR Collection has been encouraged. For further details of each approach, please consult each system paper in this volume.

The search effectiveness is usually affected by the query types (topic fields used) and query methods. Table 6 shows the participating groups' submitted runs of each category of query type and method.
In the following, the graphs show the best runs from each participating group in the category of runs.

5.1 Monolingual Runs

Retrieval results were submitted from the 17 participating groups listed below:

List of Monolingual IR Tasks Active Participants

Communications Research Laboratory (Japan)
Fuji Xerox (Japan)
Central Research Laboratory, Hitachi Co. (Japan)
Johns Hopkins University (US)
JUSTSYSTEM Corp. (Japan)
Matsushita Electric Industrial (Japan)
National Institute of Informatics (Japan)
OASIS, Aizu University (Japan)
Osaka Kyoiku University (Japan)
Ricoh Co. (Japan)
Surugadai University (Japan)
Trans EZ Co. (Taiwan ROC)
Toyohashi University of Technology (2) (Japan)
University of California Berkeley (US)
University of Tokyo (Japan)
Waseda University (Japan)

The R/P graphs of the top-ranked J-J runs of all query categories are shown in Figs. A-1 and 2, and top-ranked J-J runs using the <DESCRIPTION> only are in Figs. A-3 and 4. The R/P graphs of the top-ranked E-E runs of all query categories are shown in Figs. A-5 and 6, and top-ranked E-E runs using the <DESCRIPTION> only are in Figs. A-7 and 8. The CRL, LAPIN, and SRGDU are also tested in runs using the <TITLE> fields.

CRL - this group submitted a large number of runs produced by two different systems. Both are upgraded versions of the system participated and reported in the NTCIR Workshop 1 and IREX-IR in 1999. One used Okapi weighting and tried to reduce the free parameters.

DOVE - this group submitted runs of both the automatic and interactive systems, in which clustering and dendrograms were used.

JSCB – this group uses NLP-oriented techniques for both indexing and query processing, and also uses normalizing index terms. They utilize phrases based on NLP techniques, and employ pseudo-relevance feedback. This system worked very well on NTCIR-1, in which documents are generally short. In NTCIR-2, document length is more diversified than in NTCIR-1. Analysis on that aspect is also included.

LAPIN **-** this group uses a modified Okapi weighting.

Brkly – this is a probabilistic model-based system with logistic regression. The group participated in the NTCIR Workshop 1, but worked without any Japanese language expertise for this year. They focused on the retrieval effectiveness of different segmentation methods and tested uni-grams, bi-grams without overlap, overlapping bi-grams, and NTCIR-2 segmented texts. For the segmented texts, using only "short units", which is equivalent to words rather than phrases.

APL - this group used overlapping character-based n-grams in the indexing and retrieval of text, and compared the effectiveness of bi-grams and tri-grams in Japanese and of 6-grams and word-based segmentation in English. This is the first participation and the group and it worked without any Japanese language expertise.

SRGDU - this group modified the Berkeley-model of logistic regression to be applicable for Japanese text retrieval.

5.2 CLIR Runs

Retrieval results were submitted from 14 participating groups listed below:

List of CLIR Tasks Active Participants

ATT Labs & Duke University (US)
Communications Research Laboratory (Japan)
Johns Hopkins University (US)
JUSTSYSTEM Corp. (Japan)
Kanagawa University (Japan)
Korea Advanced Institute of Science and Technology (KAIST/KORTERM) (Korea)
Matsushita Electric Industrial (Japan)
National. TsinHua University (Taiwan, ROC)
Toyohashi University of Technology (Japan)
University of California Berkeley (US)
University of Cambridge/Toshiba/Microsoft (UK)
University of Library and Information Science (Japan)
University of Tokyo (Japan)
Yokohama National University (Japan)

(1) J-E and E-J Runs

The R/P graphs of the top-ranked J-E runs of all query categories are shown in Figs. A-9 and 10, and top-ranked J-E runs using <DESCRIPTION> only are in Figs. A-11 and 12. The R/P graphs of the top-ranked E-J Runs of all query categories are shown in Figs. A-13 and 14, and top-ranked E-J runs using <DESCRIPTION> only are in Figs. A-15 and 16. Several groups, including CAMUK [14], and panel discussion on CLIR on the first day of the Workshop Meeting, proposed the question of whether the limit of the CLIR effectiveness and the monolingual retrieval can be a baseline of the CLIR or not. In CLIR Track in the TREC-9, the search effectiveness of the English - Chinese CLIR runs also sometimes exceeded those of the Chinese monolingual runs [15]. This is a very fundamental questions and I intended to be open for further discussion at the workshop meeting, but it could not be done fully because of the time limitation.

JSCB - first participation in the CLIR. Used automatically constructed bilingual dictionaries from the NTCIR-1 collection and worked well.

MPINS - this group uses corpus-based term translation and has upgraded from the first workshop. In the second workshop, three groups participated in the CLIR task using corpus-based CLIR and they are generally slightly less effective than dictionary-based translation. This group tested several tactics to improve the effectiveness and seemed to work well.

LISIF **-** this group also participated from the first workshop and have upgraded various aspects of the system. For example, the fundamental model was changed from vector space to an Okapi-like probabilistic model, they enhanced the dictionaries, introduced machine translation systems, and used transliteration as well, as they did in the first workshop.

CAMUK - this group participated with J-E runs, but the main focus of this year's participation is to test the effectiveness of the flexible pseudo-relevance feedback on the collection as a component of the CLIR system that the first author has been investigated. To concentrate that aspect, the Japanese topics were firstly translated using a commercial product machine translation system, and then these translated topics were used as initial topics in the experiments.

(2) J-JE Runs and E-JE Runs

The R/P graphs of the top-ranked J-JE runs of all query categories are shown in Figs. A-17 and 18, and top-ranked J-JE runs using <DESCRIPTION> only are shown in Figs. A-19 and 20. The R/P graphs of the top-ranked E-JE runs of all query categories are shown in Figs. A-21 and 22, and top-ranked E-JE runs using <DESCRIPTION> only are shown in Figs. A-23 and 24.

DLUT - this group aimed at adopting an existing bilingual dictionary by extending it using a large-scale bilingual corpus.

Forst - this group proposed a new LSI model in which the corpus was segmented into several sub-corpora based on the subject. They showed that the

LSI is feasible with reasonable computing cost while keeping the search effectiveness almost the same, or slightly better, when the corpus was segmented appropriately.

6. Summary and Future Directions

Through the above overview of the workshop, we can see that various approaches and investigations have been tested using the NTCIR Collection. Lists of publications on NTCIR, and research using NTCIR-1 and NTCIR-2 are available at http://www.rd.nacsis.ac.jp/~ntcadm/paper1-en.html.

For further study, we need to consider the following issues to increase our understanding.

(1) International Collaboration for Cross-Lingual Evaluation

This year we conducted Chinese Text Retrieval Tasks and Japanese and English Information Retrieval Tasks separately. We are examining the possibility of conducting Chinese-English-Japanese CLIR evaluation and desire to increase the number of languages available. Collaboration with HANTEC and CLEF are important.

(2) Evaluation Metrics Suitable for Multi-grade Relevance Judgments

An initial discussion of the proposed metrics is found [15]

(3) Examining the influence of the reuse of training data in the test and experimental frameworks using paired texts.

The training data was reused as part of test data. It may hide the effectiveness of the systems or their advantages. The influences may be caused by such reuse should be examined. Also more rigid specification of the CLIR experiments using paired documents and discussion on the baseline of the CLIR should be needed.

(4) Encourage Resource Sharing and Exchange

(5) Methods to Estimate the Difficulty of Search Topics

(6) Enhance the Interaction among the Forum

(7) Enhance the Variation of Text Types.

(8) Further Discussion of Evaluation Schemes and Subtasks

In particular, evaluation schemes may include the evaluation of interactive systems, using real Web documents including hyperlinks, post-retrieval processing such as automatic abstracting, pinpointing the answers in the retrieved documents, and so on.

ACKNOWLEDGMENTS

We are grateful for the 65 academic societies for their understanding and permission for the use of documents, and for Heiwa Information Center Co. for their permission to use the segmented data produced by their product. We thank all the participants for their contributions and the analysts who worked so hard and with such a high level of concentration. Special thanks are due to Donna Harman, Ellen Voorhees, Sung H. Myaeng, and Mun Kew Leong for their substantial advice and continuous support.

REFERENCES

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

[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.

[3] Kuriyama, K., Kando, N.: Comparison of NTCIR-1 and NTCIR-2, In Proceedings of the 61st IPSJ Annual Meeting [in Japanese]. p.3U-08, Matsuyama, Japan, Oct. 2000,

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

[5] Sung, H.M. "Hantec Collection". Presented to the Panel on IR Evaluation in the 4th IRAL, Hong Kong, 30 Sept – 3 Oct. 2000.

[6] Sakai, T., Kitani, T., Ogawa, Y., Ishikawa, T., Kimoto, H., Keshi, I., Toyoura, J., Fukushima, T., Matsui, K., Ueda, Y., Tokunaga, T., Tsuruoka, H., Nakawatase, H., Agata, T., Kando, N. "BMIR-J2: A test collection for evaluation of Japanese information retrieval systems". SIGIR-Forum. Vol.33, No.1, p.13-17, December, 1999

[7] K. Eguchi, K. Kuriyama, and N. Kando: "Analysis of the Topic Difficulty for NTCIR (NACSIS Test Collection for Information Retrieval Systems)", In Proceedings of the 3rd International Conference of Asian Digital Library (ICADL 2000), pp.231-238, Seoul, Korea, Dec. 2000.

[8] Kuriyama, K. et al.: 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)

[9] Kando, N., Aihara, K., Eguchi, K., Kato, H(ed). . NTCIR Workshop 2 Meeting: Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization. National Institute of Informatics 874 p., Tokyo, March 2001 (ISBN: 4-924600-89-X)

[10] Yoshioka, M. et al.: 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)

[11] Nozue, T., Kando, N., Kuriyama, K. Reconsideration of the concept of relevance for NACSIS Test Collection. In

Proceeding of the 46th Annual Conference of Japan Society of Library and Information Science. November, 1998, p. 67-70.(in Japanese)

[12] Nozue, T., Kando, N. Primary considerations in the concept of relevance: Relevance judgement of NTCIR (in Japanese). In the Special Interest Group Notes of Information Processing Society of Japan. No. 99-FI-53, March, 1999, p. 49-56.

[13] Iwayama, M., Niwa, Y., Nishioka, s., Takano, A. Hisamitsu, T. Imaichi, O., Sakkurai, H., Fujio, M. The effect of document clustering in interactive relevance feedback. 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)

[14] Sakai, T., Robertson, S.E., Walker, S. Flexible Pseudo-Relevance Feedback for NTCIR-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)

[15] Kando, N., Kuriyama, K., Yoshioka, M. Evaluation based on multi-grade relevance judgements. IPSJ SIG Notes, July 2001 (to appear)