The NTCIR Workshop : the First Evaluation Workshop on Japanese Text Retrieval and Cross-Lingual Information Retrieval

Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, Soichiro Hidaka, Jun Adachi
Research and Development Department
National Center for Science Information Systems (NACSIS) , Tokyo, Japan
URL: http://www.rd.nacsis.ac.jp/~{kando,ntcadm}/

Abstract

This paper introduces the outline of the first NTCIR Workshop, which is the first evaluation workshop designed to enhance research in Japanese text retrieval and cross-lingual information retrieval. The test collection used in the Workshop consists of more than 330,000 documents with more than half are English-Japanese paired. Twenty-three groups from four countries have conducted IR tasks and submitted the search results. Various approaches were tested and reported at the Workshop. Finally some thoughts on the future directions of the NTCIR Workshop and evaluation of cross-lingual information retrieval with Asian languages are suggested.

1. Background and Aims

The First NTCIR Workshop was held on August 30-Septebmer 1, 1999, in Tokyo[1]. The participation to the Workshop was limited to the active participants, *i.e.* the members of the research groups that submitted the results of the tasks, advisors and members of the organizing group. Many interesting papers with various approaches were presented at the Workshop and it ended in great enthusiasm. The third day of the Workshop was organized as the NTCIR/IREX Joint Workshop. IREX Workshop, the another evaluation workshop of IR and information extraction (named entity) using Japanese newspaper articles were held consecutively.

The NTCIR Workshop was planed as part of the NTCIR project[1][2], which is intended to provide sound infrastructure to evaluate the search effectiveness of information retrieval systems with Japanese language and facilitate the IR research with Japanese language and cross-lingual retrieval including Japanese.

The project is motivated by the recognition of the following situations:

(1) Needs for a standard Japanese test collections

(2) Need for cross-lingual retrieval

(3) Need for the variety in text types

(4) Need for the fundamental data for research into the intersection of IR and NLP

The importance of the large-scale standard test collection in IR research are widely recognised. Stopping, stemming and query analysis are language depended procedures. Especially indexing texts written in Japanese or other East Asian languages like Chinese or Korean are quite different from those with English, French or other European languages since there is no explicit boundary (i.e. no space) between words in a sentence. Regarding other East Asian languages, there are large-scale test collections of Chinese and Korean. For Japanese, there is a standard test collection called BMIR-J2, consisting of 5,080 Japanese newspaper articles and ca.60 queries [3]. Although its contribution to Japanese IR research is tremendous, enhancement of the collection in both variety of text types and scale was needed.

Cross-lingual retrieval is critical in the Internet environment. Moreover in the scientific texts, foreign language terms, sentences, or abstracts are often appeared in a Japanese text in their original spelling. Therefore cross-linguistic strategies are also critical for retrieval of Japanese scientific documents [4]. In order to respond the needs stated above, we aim to construct a large-scale test collection which is usable for cross-lingual retrieval and application of NLP to IR, and organize an evaluation workshop using it.

The NTCIR Workshop has the following goals;

(1) to encourage research in information retrieval, cross-lingual information retrieval and related areas by providing a large-scale Japanese test collection 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 effective methods for constructing large-scale test collections and IR laboratory-type testing.

The test collection used in the Workshop consists of more than 330,000 documents and more than half are English-Japanese paired.

In the next section, we describe the tasks performed in the Workshop. Section 3 shows the test collection (NTCIR-1) used in the Workshop and section 4 introduces the evaluation results. The final section discusses some thoughts on future direction.

2. The Tasks

A participant conducted one or more of the tasks below:

The Ad Hoc Information Retrieval task : to investigate the retrieval performance of systems that search a static set of documents using new search topics

The Cross-Lingual Information Retrieval task : an ad hoc task in which documents are in English and topics are in Japanese.

The Automatic Term Recognition and Roll Analysis task : (1) to extract terms from titles and abstracts, and (2) to identify the terms representing the "object", "method" and "main operation" of the main topic.

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

2.1 The Procedures

In November, 1998, the document data, 30 ad hoc topics, 21 cross-lingual topics and their relevance assessments were delivered for each IR tasks participant to train their systems. The 53 new test topics were distributed on February 8, 1999 and the search results for them were submitted by March 4 as official test runs. The test topics are common for both IR tasks.

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 had to submit at least one set of results of the searches using only <Description> fields of the topics as the mandatory runs. For optional automatic runs and manual runs, any fields of the topics could be used. Also each participant had to fill and submit a system description form describing the detailed feature 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 on June 12 to active participants who submitted search results. Based on them, inter-polated 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.

2.2 The Participants

Thirty-one groups including participants from six countries have enrolled to participate the first NTCIR Workshop. Among them, 28 groups have enrolled in IR tasks (23 in the Ad Hoc Task and 16 in the Cross-Lingual Task), and nine in the Term Recognition task.

The below is the list of active participating groups that submitted results of the tasks.

Communications Research Laboratory (MPT)	
Fuji Xerox	RMIT & CSIRO
Fujitsu Laboratories	Tokyo Univ. of Technology
Hitachi	Toshiba
JUSTSYSTEM	Toyohashi Univ. of Technology
Kanagawa University (2 groups)	Univ. of California Berkeley
KAIST/KORTERM	Univ. of Lib. and Inf. Science
Manchester Metropolitan Univ.	Univ. of Maryland
Matsushita Electric Industrial	Univ. of Tokushima
NACSIS	Univ. of Tokyo
National Taiwan Univ.	Univ. of Tsukuba
NEC (2 groups)	Yokohama National Univ.
NTT	Waseda Univ.

Regarding IR tasks, 23 groups submitted search results of 117 runs. There were 48 runs for the Ad Hoc Task from 17 groups and 69 runs for the Cross-Lingual Task from 10 groups. Nine groups are from Japanese companies (six in the Ad Hoc Task and four in the Cross-Lingual, one did both), 11 are from Japanese universities or national research institutes, and four are non-Japanese groups. Two groups are from the United States, one group is from Australia, one group is from Taiwan and 19 groups are from Japan; some of this latter group have non-Japanese members or have collaborated with research groups outside Japan. Two groups worked without any Japanese language expertise.

3. The Test Collection

The test collection used in the Workshop consists of; documents, topics, and relevance judgments for each search topic.

3.1 Documents

The documents are author abstracts of the papers presented at conferences hosted by 65 Japanese academic societies [5]. Since one of the purposes of the original database is to provide an alerting information service about 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.

Documents are SGML-like tagged plain text. A record may contain document ID, title, a list of author(s), name and date of the conference. abstract, keyword(s), and name of the hosted society. (See Fig. 1)

<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 on electronic publishing, SGML (Standard Generalized Markup Language), are assumed to be important, and the results from an experiment at NACSIS to publish "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 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. 1. A Sample of the Document Record

The Collection contains three document collections, *i.e.* JE, J, and E. The JE Collection contains 339,483 documents, more than half are English-Japanese paired. The J and E Collections are constructed through extracting Japanese or English parts of the

documents, respectively, from the JE Collection.

In the Workshop the JE Collection is used in the Ad Hoc task since Japanese operational IR environment, especially, retrieval of scientific documents and Web documents, retrieving both Japanese and English documents at a time is quite natural. The E Collection is used in the Cross-lingual Task. The J Collection is used in the monolingual retrieval, which will be the baseline for comparing the search effectiveness with the results in the cross-lingual runs.

3.2 Topics

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

Its format is similar to the one once used in the TREC-1 and 2 and contains SGML-like tags. A topic consists of a title of the topic, a description, a detailed narrative, a list of concepts and field(s). The title is a very short description of the topic and can be used as a very short query which resembles the one often submitted by an end-user of internet search engines. Each narrative may contain detailed explanation of the topic, term definition, background knowledge, purpose of the search, criteria of relevance judgment, and so on.

<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>
<CONCEPTS>
特徴選択、主成分分析、情報の粒度、幾何クラスタリング
</CONCEPTS>
<FIELD>
1.電子・情報・制御
</FIELD>
</TOPIC>

Fig. 2 A sample Topic

3.2.1 Topic Preparation

Topics were collected from users who gave permission to use them as part of a test collection. Some were collected from researchers in several fields, some were from reference counters of research libraries, and others were created by the analysts based on their research interest or needs. Analysts were mainly graduate students with backgrounds in computer sciences, pharmacology, biochemistry, social sciences such as education, linguistics, and so on.

The Collection contains 30 training topics and 53 test topics. Among them, 21 training topics and 39 test topics are usable for cross-lingual retrieval. All the topics are written in Japanese. English and Korean versions will be available.

Each topic was examined for its clarity and difficulty by the analysts and project members in NACSIS. 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: Simple word matching of query terms cannot retrieve every relevant document and a document containing query terms can be non-relevant.

(4) Five or more relevant documents in the top 100 documents retrieved by the retrieval system that we used in NACSIS.

We put the criteria (3) since in the real world documents, a concept can be represented by different terms and a term can represent different concepts and this ambiguity is on of the essential characteristics of the text retrieval.

3.3 Relevance Judgments (Right Answers)

The relevance judgments were done in three grades, i.e., relevant, partially relevant, non-relevant. Two analysts assessed the relevance of a topic separately, then the primary analyst of the topic who created the topic decided the final judgment.

Relevance judgment files contain not only the relevance of each document in the pool but also contain extracted phrases or passages showing the reason why the analyst assessed the document as "relevant". Since a narrative of topics may contain some description related to the user's situation or the purpose of the search, situational-oriented relevance judgments were conducted 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.

3.4. Linguistic Analysis

A part of the J collection contains detailed part-of-speech tags [6]. Because of absence of explicit boundary between words in Japanese sentences, we set the three levels of lexical boundaries (i.e., word boundary, strong and week morpheme boundary), and assigned detailed POS tags based on the boundaries and types of origin. This part was used in the Term Recognition Tasks.

3.5 Robustness of the System Evaluation using the Test Collection 1

The Test Collection 1 itself has been tested from the following aspects so that it is usable as a reliable tool for IR system testing:

(A) exhaustivity of the document pool

(B) inter-analysts consistency and its effect for system evaluation

(C) topic-by-topic evaluation.

The results of these studies have been reported and published on various occasions [7-11]. As results, in terms of exhaustiveness,

pooling the top 100 documents from each run worked well for topics with less than 50 relevant documents. For the topics with more than 100 relevant documents, although the top 100 pooling covered only 51.9% of the total relevant documents, the coverage reached higher than 90% if combined with additional interactive searches. Therefore we decided to use the top 100 pooling and conducted additional interactive searches for the topics with more than 50 relevant documents.

We found strong correlation between the system rankings produced using different relevant judgments and different pooling methods regardless of the inconsistency of the relevance assessments among analysts and regardless of the different pooling methods [7-9,11]. The similar analysis using has been reported by Voorhees [12]. We concluded that the test collection is reliable as a tool for system evaluation based on these analyses.

4. Evaluation Results

4.1 Ad Hoc IR Task

The R/P graphs of the top Ad Hoc all runs and top Ad Hoc short queries without <Concepts> runs are shown in Fig.3 and 4. The term "long query" represents any query using <Narrative> in the topic. A "short query" is any query that did not use <Narrative>. A "long query" includes <Concept> otherwise specified and "short query" does not include <Concept> otherwise specified.

4.1.1 All Runs

Results are summarized as follows;

(1) The runs used long queries obtained better results than ones used short queires, but some runs were opposite.

(2) Interactive runs were often better than automatic runs but the effectiveness of the interactive runs and the levels of intervention of human searchers varied.

(3) The runs used <Concept> fields of the topics obtained better results than runs without <Concept>.

(4) Both n-gram and word or word and phrase-based indexing were used. As an extension of n-gram, an adaptive segmentation was proposed by UTS group.

(5) Query expantion was used by several groups and in most cases, it seemed to work well and provided higher search effectiveness for both automatic and interactive runs.

JSCB – *Justsystem* group uses strong NLP-oriented techniques for both indexing and query processing, and also uses normalizing index terms. They utilize phrases and employ relevance feedback for both automatic and interactive runs on a vector space model with weighting scheme based on tf/idf. JSCB3 is an interactive long query, JSCB2 is an automatic long query, and JSCB1 is an automatic short query run.

BK – ***Berkeley*** uses rather simple bi-grams, just discarding HIRAGANA (phonetics, mostly used for functional words)(BKJJBIFU). Berkeley's word-based approach uses the longest match with the dictionary and seems also not so deeply dependent on NLP(BKJJDCFU). Probabilistic model. Weighting scheme uses tf/idf, document length, query length, and collection

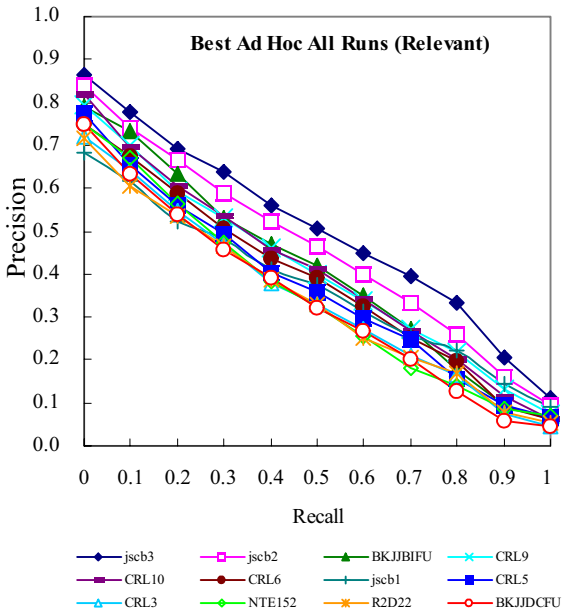

Fig. 3 Top Ad Hoc All Runs (Relevant Level)

length. BKJJBIFU and BKJJDCFU are automatic long query runs.

CRL - Probablistic model with tf/idf and query idf. Indexing using stemming and EDR dictionary. Queries utilize words and phrases. CRL10,9,6,5 and 3 are all automatic long query runs.

NTE15 - *Matsushita* group uses vector space model. The weighting uses tf/idf, document length and cooccurence. Index and query are both word based but index terms are segmented using overlapping longest match using EDR and internally prepared dictionaries. NTE152 is an automatic short query run with <Title> and <Concept>.

R2D2 - *Tokyo University* group uses vector space model with extension called "super impose" of vectors. Index terms are segmented by morphological analyzer then selected by POS, stop words and normalization. R2D22 is an automatic short query run.

One of the most interesting things found in this evaluation is that these two systems of JSCB and BK, which took completely different approaches, both obtained very high scores. JSCB used NLP techniques very well and BKJJBIFU focused on the statistical approach of weighting algorithms based on prolonged experience of expanding probabilistic model. Traditionally, the Japanese IR community has tended to pay too much attention to the methods of segmenting texts into tokens rather than studying retrieval models or algorithms themselves. Some groups used weighting schemes which have been reported worked well against English documents without testing on Japanese documents. It is probably because of time shortness in the schedule of the Workshop and extension of the experiments on the weighting schemes are strongly expected.

4.1.2Automatic Short Queries Runs without <Concept>

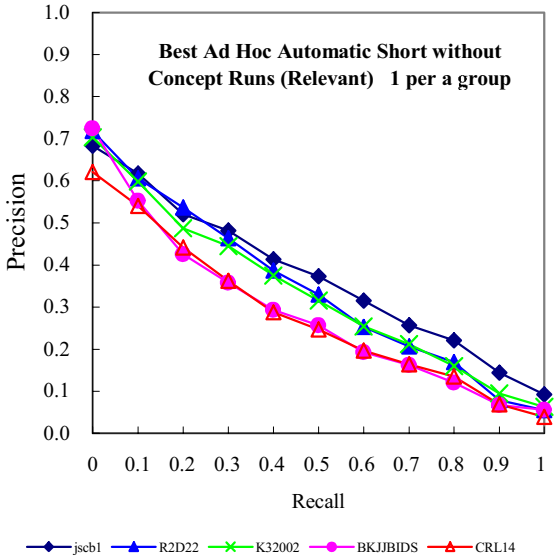

Fig. 4 Best Ad Hoc Automatic Short Queries without <Concept> Runs (Relevant Level) (Mandatory Runs) (chose 1 run per a group)

K32002 - This group used extended n-gram on B-Tree. The system is based on vector space model with the weighting scheme of tf/idf and document length.

BKJJBIDS - It is a short query run of Berkeley group and used simple bi-gram as BKJJBIFU on probabilistic model system stated above.

CRL4 - It is the best short query run of CRL group's system stated above. It also incorporated query expansion.

4.2 Cross-Lingual IR Task

In the Cross-Lingual Task, the effectiveness of the cross-lingual searches of English documents (E Collection) by Japanese topics was compared with that of the monolingual searches of Japanese documents (J Collection) by Japanese topics. In the following, the best monolingual searches in the query type and the query length were shown in the figures in dotted linea as baselines.

4.2.1 All Runs

The results are summerized as follows;

(1) A search using a longer query tended to obtain better results, however, some runs gave opposite results and could not utilize the <Narrative> of the topics to improve search effectiveness.

(2) A search using <Concept> in the topic obtained better results than a search without <Concept>, but the search using <Concept> only worked poorly.

(3) Every run took "query translation approach".

(4) The size of internally prepared dictionaries for query translation varied; from 20 K to 560 K entries.

(5) Technical terms were one of the most difficult problems in NTCIR-1. Using phonetics (transliteration) was proposed by ULIS group and it worked well.

(6) Query expansion and word disambiguation were conducted by several groups; post-translation QE, pre-translation QE, automatic local feedback, more naïve QE of translating into more than one target language terms, and so on.

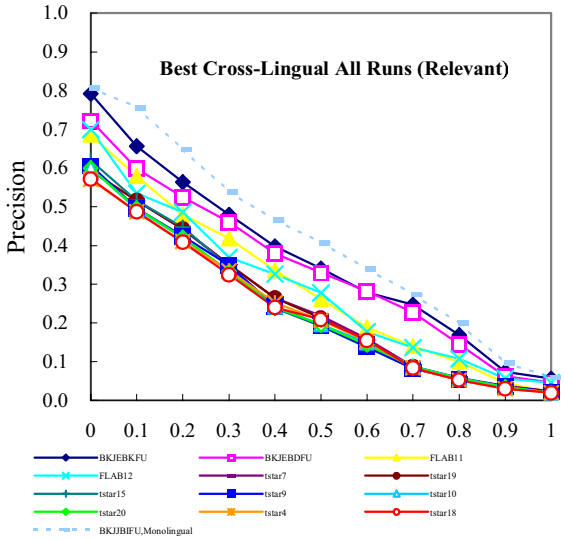

Fig. 5 Best Cross-Lingual All Runs (Relevant Level)

BK - Berkeley group tested both dictionary-based and machine translation-based cross-lingual runs on the system based on probabilistic model. Unfortunately MT system did not incorporated with technical terms and worked not so well. The dictionary used was created internally using keywords of the JE Collection documents and other lexical resources and contained more than 373 K entries and the size was 23MB. BKJEBKFU and BKJEBDFU are both automatic dictionary-based translation runs without query expansion using long queries whereas the former used every terms in <Concept> and the latter used <Concept> excepts English terms.

FLAB - These were interactive runs using long queries using <Concept> except English terms in it.

TSTAR - National Taiwan University group tested heavily on the automatic dictionary-based translation runs with post-translation query expansion and word disambiguation using various English corpora on vector space model system. Tstar7, 19, 15, 9 used <Title> and <Concept> in the topic, tstar 10 and 20 used <Title>, <Description> and <Concept>, and tstar 4 and 18 used <Description> and <Concept>. The size of dictionary is 3MB and one of the smallest among the ones used in the Workshop.

4.2.2 Automatic Short Queries without <Concept> Runs

BKJEBDDS - it is a short query run of the BK group and the system used was the same as stated above and used the dictionary-based translation without query expansion.

1KE3 - it is the run by the organizing group and put in the pooling just to show the effectiveness of the system used for the initial pooling done in the NACSIS. It was dictionary-based translation using automatically generated multiligual keyword clusters which were created from less than 10% of the JE

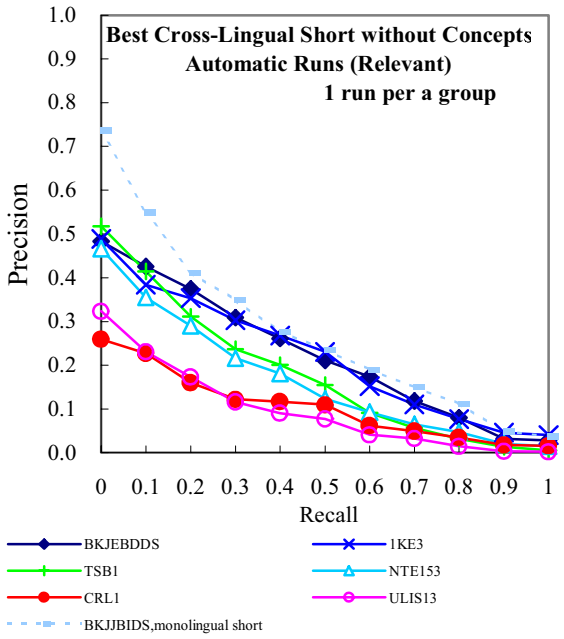

Fig. 6 Best Cross-Lingual Automatic Short Queries Runs without <Concept> (Relevant Level) (chose 1 per a group)

Collection based on graph theory. Each query term was translated into more than one target language terms using the clusters. The detail of the approach was reported in [13] .

NTE15 - Matsushita group is the only group used corpus basaed translation in this Workshop. The system is based on vector space model and without query expansion.

TSB - Toshiba group tested both automatic and interactive, and both vector space and probablistic models, and two different machine translation systems. The interactive runs obtain better results than the runs shown here. The TSB1 run used machine translation based query translation on a probablistic model system with weighting scheme using ti/idf and document length. The group also tested local feedback.

CRL - The group tested dictionary-base translation on a probabilistic model system. CRL1 used EDR to translate the queries with post-translation query expansion and word disambiguation using TREC 4 and 5 collections. Another run used internally prepared dictionary using the JE Collection.

ULIS - The group tested dictionary-based query translation on a vector space model system using both a subject-oriented dictionary and a general term dictionary and transliteration of Katakana terms in the topics into English terms. Both dictionaries are internally prepared from EDR. ULIS13 used both dictionary and transliteration. Query expansion was not used but some runs adopted translation of each query term into up to 3 or 10 terms in the target language but not for ULIS13.

5. Summary and Future Directions

Through the above overview of the Workshop, we can see that various approaches and investigation have been tested using the NTCIR Collection. The results of the research have already been reported at several international conferences. Lists of publications on NTCIR and research using NTCIR-1 are available at http://www.rd.nacsis.ac.jp/~ntcadm/paper1-en.html. For further study, we need to consider the following issues ;

(1) Schedule

IREX and NTCIR will join and organize the NTCIR Workshop 2. The call for participation is planned for April 2000; documents and training topics will be distributed in May, test topics will be distributed in September or October, and the Workshop meeting will be held in March 2001. The effort to avoid the overlaps of the schedule with other evaluation projects like TREC, TDT, and TIDES should be considered.

(2) Evaluation of cross-lingual retrieval

We used the evaluation method that searches English documents by Japanese queries and compares the system effectiveness against monolingual searches of Japanese documents by Japanese queries. Further consideration and discussion are required for the validity of it. The English and Korean topics will be prepared.

Use of the ntc1-je0 (JE Collection) to train systems or to extract knowledge were allowed in this NTCIR Workshop. However, using the bilingual lexical resource created from it to translate queries is somehow a kind of closed testing for lexical resource preparation. It is highly rewarding to develop appropriate methods to utilize the readily available corpus to create bilingual lexical resources, especially the ones for technical terms, and to investigate the highest ceiling of the search effectiveness of cross-lingual retrieval using this collection. However, to obtain more solid evidence, these results should be tested against a new document set in the future. The workshop organizer hopes to provide appropriate document sets for this purpose in the near future.

(3) International collaboration

International collaboration is needed for enhancement of the research in cross-lingual retrieval and its evaluation. For example, the Korean test collection project, which is headed by Sung H. Myaeng, and the NTCIR plan to exchange the topics. IR and CLIR with Asian languages have been attracted researchers out side Asia. Sharing fundamental resources like dictionaries or morphological analyzers and fundamental knowledge to process Asian languages shall be great help to enhance the research of IR and CLIR using Asian languages.

(4) Enhance the variation of text types. Copyright issues

(5) Further subtasks.

For example, evaluation of interactive system, using real Web documents including hyperlinks, post retrieval processing such as automatic abstracting, pinpointing the answers in the retrieved documents, and so on shall be investigated.

Retrieving documents which may include relevant information is the purpose of the traditional IR systems but it is not the end of the story. Users may wish to have more sophisticated function to support their information works such as decision making, problem solving, writing papers, etc. using retrieved documents. It is obvious that the evaluation methods for those function should be investigated. For the future direction, we have to test IR systems'

effectiveness on the R/P based methods as one of the fundamental functions of IR systems and to see additional aspects as well.

ACKNOWLEDGMENTS

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, Ross Wilkinson, Sung H. Myaeng, and Mun Kew Leong for their substantial advice and continuous support.

REFERENCES

[1] NTCIR Workshop 1 : Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognision, Aug. 30-Sept.1, 1999, Tokyo, ISBN4-924600-77-6. (http://www.rd.nacsis.ac.jp/~workshop/OnlineProceedings/)

[2] NTCIR Project. http://www.rd.nacsis.ac.jp/~ntcadm/

[3] Sakai, T. *et al*. BMIR-J2: Test Collection for Evaluation of Japanese Information Retrieval Systems. SIGIR Forum (to appear).

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

[5] The list of 65 academic societies is available at URL, http://www.rd.nacsis.ac.jp/~ntcadm/acknowledge/thanks1-en.html.

[6] Kageura, K. *et al*. NACSIS Corpus Project for IR and Terminological Research. NLPPR '97, (Dec. 1997), p. 493-49

[7] Kando, N. *et al*. NTCIR-1: Its Policy and Practice. IPSJ SIG Notes, Vol.99, No.20, p. 33-40 (1999) [in Japanese].

[8] Kuriyama, K. *et al*. 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 (May, 1999) [in Japanese].

[9] Kuriyama, K. *et al*. Construction of a Large Scale Test Collection: Analysis of the Training Topics of the NTCIR-1. IPSJ SIG Notes, Vol. 99 (July, 1999) [in Japanese].

[10] Kando, N. *et al*. Construction of a Large Scale Test Collection: Analysis of the Test Topics of the NTCIR-1. In Proceedings of IPSJ Annual Meeting (to appear) [in Japanese].

[11] Kuriyama, K. *et al*. Pooling for a Large Scale Test Collection: Analysis of the Search Results for the first NTCIR Workshop. In Proceedings of IPSJ Annual Meeting (to appear) [in Japanese].

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

[13] Kando, N.; Aizawa, A. Cross-Lingual Information Retrieval using Automatically Generated Multilingual Keyword Clusters In Proceedings of the 3rd International Workshop of IR with Asian Language, p.86-94, Oct.13-14, 1998, Singapore