NTCIR Workshop: an Evaluation of Cross-Lingual Information Retrieval

Noriko Kando
Research and Development Department,
National Center for Science Information Systems (NACSIS), Japan
kando@rd.nacsis.ac.jpl

Abstract

This paper introduces the first NTCIR Workshop, Aug.30 - Sept.1, 1999, 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 of English and Japanese. 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 are suggested.

1. Introduction

The NTCIR Workshop [1] [1] has the following goals;

(1) to encourage research in information retrieval (IR), cross-lingual information retrieval (CLIR) 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.

The test collection used in this Workshop is called “NACSIS Test Collection 1” or “NTCIR-1” and consists of more than 330,000 documents, with more than half presented as English-Japanese pairs. Although there is a Japanese test collection called BMIR-J2 consisting of 5,080 newspaper articles [2], enhancement of the Japanese test collection in both variety of text types and scale was needed. We place emphasis on CLIR since it is critical in the Internet environment and for Japanese scientific information retrieval [3].

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.

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, 11 are from Japanese universities or national research institutes, and four are non-Japanese groups. Some of the Japanese groups have non-Japanese members or have collaborated with research groups outside Japan. Two groups worked without any Japanese language expertise

In the next section, we describe the tasks performed in the Workshop. Section 3 shows the test collection used in the Workshop and section 4 introduces the evaluation results. The final section discusses future direction.

2. The Tasks

A participant conducted one or more of the tasks below:

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

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

C) **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.

2.1 The Procedures

In November, 1998, the document data, 30 ad hoc topics, 21 cross-lingual topics and their relevance judgments 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

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

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. 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.

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 [4]. 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.

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 was 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 was used in the Cross-lingual Task. The J Collection was used in the monolingual retrieval, which was 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.

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.

3.2.1 Topic Preparation

Some topics were collected from users who gave permission to use them as part of a test collection. 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 by the analysts and project members in NACSIS according to the criteria below;

(1) Not too easy: Simple word matching of query terms cannot retrieve every relevant document and a document containing query terms can be non-relevant.

(2) Five or more relevant documents in the top 100 documents retrieved by the initial searches that we used in NACSIS.

We put the criteria (1) 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.

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 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 were used in the formal evaluation of this Workshop.

3.4. Linguistic Analysis

A part of the J collection contains detailed part-of-speech tags. 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:

(1) exhaustivity of the document pool
(2) inter-analysts consistency and its effect for system evaluation
(3) topic-by-topic evaluation.

The results of these studies have been reported and published on various occasions [6-10]. 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 [6-8,10]. The similar analysis using has been reported by Voorhees [11]. We concluded that the test collection is reliable as a tool for system evaluation based on these analyses.

4. Evaluation Results

Many interesting investigation with various approaches were reported at the Workshop and it ended in great enthusiasm. The results were summarized as follows;

(1) The searches used longer queries obtained better results than ones used short queries, 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>, but the search using <Concept> only worked poorly.

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

(5) Query expansion was used by several groups both for Ad Hoc and cross-lingual retrieval. In most cases, it seemed to work well and provided higher search effectiveness for both automatic and interactive Ad Hoc runs. For cross-lingual retrieval, post-translation QE, pre-translation QE, automatic local feedback, more naïve QE of translating into more than one target language terms, and so on. Further investigation and update are expected.

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

5. Summary and Future Directions

NACSIS will change its name into National institute of Informatics in the next spring and plans to run a second evaluation after the change. It will include at least Japanese and English with training data available in May 2000, test data available in approximately in September, and the workshop itself scheduled for February or March, 2001. Some part of the schedule may be changed through the effort to avoid the schedule overlap of other evaluation project like TREC, TDT, or TIDES.

Meanwhile, we are planing details of the tasks, subtasks, evaluation scheme, collection, and resources. The needs of training courses and tutorials on evaluation of information retrieval systems including interactive systems for Japanese new comers in Japanese language and ones on Japanese text processing and available resources for non Japanese researchers in English are suggested from the advisory group. Any comments, advises, and leads are welcome.

Acknowledgements

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. Online version is available at [2]
[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