Lessons from NTCIR-4: Focusing on Evaluation of CLIR on East Asian Languages, Patent and QA

Outline of Invited talk at CLEF 2004 Workshop

Noriko Kando, National Institute of Informatics, Tokyo, Japan

This talk will present the fourth NTCIR Workshop, which is the latest in a series of evaluation workshops designed to enhance research in information access (IA) technologies including information retrieval (IR), cross-lingual information retrieval (CLIR), automatic text summarization, question answering, text mining and so on by providing large-scale test collections and a forum for researchers. Seventy-four groups from 10 different countries and areas participated in the workshop and submitted results.

The talk will briefly describe the background, the tasks, the participants, and the test collections of the workshop. NTCIR-4 selected five areas of research as "tasks":

1. Cross-Lingual Information Retrieval Task (CLIR)
2. Patent Retrieval Task (PATENT),
3. Question Answering Challenge (QAC)
4. Text Summarization Challenge (TSC), and
5. WEB Task (WEB).

These tasks were enhancements from previous editions of NTCIR. Each of them increased the size of the test collection used.

PATENT proposed experiments within the different information seeking tasks regarding “invalidity search” and the challenging topic of “automatic patent map generation” as a feasibility task in a long-term three year research project lasting until NTCIR-5. For Patent, invalidity search is basically "search patents by patents" and it included a combination of content-based IR and the use of the metadata such as publication dates, classification codes, etc. Relevance judgments were done at both document- and passage-levels.

TSC included the automatic evaluation of summaries and the building of a re-usable test collection for summarization.

CLIR and QAC basically continued with minor changes in task design to remedy the major problems found in the third workshop. In CLIR, for every languages, documents were collected from multiple sources.

The talk will focus in particular on the CLIR activities at NTCIR-4. Pivot CLIR and named entities were focus points, and interesting approaches and strategies included bi-directional CLIR, i.e. combination of document translation and query translation, sophisticated relevance feedback such as Flexible Pseudo Relevance Feedback, out-of-vocabulary problems and the use of web, indexing and word segmentation especially focusing on the decompounding of Korean terms and the comparison of word- and n-gram-based indexing.

We will also outline some of the problems that remain to be addressed, these include a few relevant document issues and the evaluation of system robustness. Mention will be made of the feasibility studies were proposed on Automatic Patent Map Creation (Text mining), Geographic Oriented Web Retrieval, and Search Results Classification of WEB search and on the use of new, completely different document types. The talk will conclude with some thoughts on future directions.