EVIA 2007: The First International Workshop on Evaluating Information Access

Mark Sanderson
University of Sheffield
m.sanderson@shef.ac.uk

Tetsuya Sakai
NewsWatch, Inc.
sakai@newswatch.co.jp

Noriko Kando
NII
Noriko.kando@nii.ac.jp

Abstract

The first workshop on Evaluating Information Access was held at the National Institute of Informatics, Tokyo, Japan on May 15th, 2007. It was composed of a five invited speakers and two sessions of refereed papers and posters.

1 Introduction

Evaluation of Information Retrieval, Question Answering and Text Summarization systems has been central to Information Access research for decades. As retrieval has become more pervasive and diverse, the need for effective and efficient evaluation has never been more important. Following the success of Open Submission Sessions held in the previous two NTCIR meetings, it was decided to hold a Pre-Meeting Workshop (PMW) on evaluation in Information Access, which ran the day before the main sessions of NTCIR 2007. An online proceedings for the workshop can be found at

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/EVIA/

The workshop was split into invited and refereed presentations. The two Sections are now described.

2 Invited Talks

The theme of the invited presentations was to illustrate the broad diversity of evaluation research ongoing beyond the well known three centers of evaluations research: TREC, CLEF and NTCIR. Speakers from China, Vietnam, India and Thailand each spoke about their work on building test collections for the languages and documents of their country.

Le Sun (Chinese Academy of Sciences, China) spoke about the HTRDP IR evaluation exercise (also known as the *863 evaluation*), which has been running since 2004. The IR component (based on the TREC model) was part of a wider range of language technology evaluations conducted by HTRDP. Le Sun described the running of the exercise highlighted the challenges of encouraging participation in China and outlined future plans centered on use of query logs.

Ho Bao-Quoc (Vietnam National University, Vietnam) described the issues in adapting an IR system to work with the Vietnamese language: word boundaries, morphology and part of speech all need to be addressed if retrieval from this language is to succeed. A Vietnamese test collection was also presented in the talk along with the results of a set of experiments examining the success the different language processing tools for Vietnamese.

Prasenjit Majumder (Indian Statistical Institute, Kolkata, India) detailed a new initiative for Indian Language Information Retrieval (ILIR). As with HTRDP, ILIR is starting to run an evaluation exercise based on the TREC model. Prasenjit described the design of the exercise:
it will be based on six languages spoken in India - Bengali, Hindi, Marathi, Punjabi, Tamil and Telugu.
There will be a collection of around 50,000 documents for each language
There will be 50 topics for each language
Cross language searching will be an important research theme of the exercise.

Virach Sornlertlamvanich (TCL, NICT Asia Research Center, Thailand) described work on evaluating search engines using Thai queries. As with Vietnamese, word boundaries in the Thai language are not explicitly marked, neither are sentence boundaries. Virach detailed his work using a query log analysis from a Thai search engine and then described a comparison of seven search engines that search Thai web pages.

Ian Soboroff (NIST, USA) rounded out the invited speaker session with a description of work on going at NIST to try to understand how re-usable test collections are. Almost all modern test collections are built using the pooling strategy first suggested by Spärck Jones and Van Rijsbergen (1975). As test collections are built from larger sets of documents, there is an increasing concern that the pooling strategy isn't locating a diverse enough selection of documents to ensure that future systems measured on test collections will be measured accurately. Ian's work shows potential problems but as yet, the "smoking gun" of a badly measured IR system has yet to appear.

3 Refereed Papers

There were four presentations in this Section followed by a small number of posters. The presentations were from

Tetsuya Sakai (NewsWatch, Japan) who described his work on assessing the quality and flexibility of evaluation measures that deal with graded relevance judgments. Showing that the Q measure was better than the more popular DCG suite of measures.

Yaoyong Li (University of Sheffield, UK) who presented a new measure for assessing the submissions to NTCIR patent classification task.

Maarten de Rijke (ISLA, University of Amsterdam, The Netherlands) described the novel WiQA track of CLEF, where participants were challenged to work with Wikipedia identifying snippets from articles that could be added to a source article. The additions could come from articles written in languages different from the text of the source.

Giorgio Maria Di Nunzio and **Nicola Ferro** (University of Padua, Italy) described the DIRECT system developed at Padua, which is being used extensively by CLEF track organizers to create test collections and by participants to measure and to understand the effectiveness of their retrieval systems. All other

evaluations exercises such as TREC and NTCIR have such systems; however, CLEF's system appears to be the most complete.

4 Conclusions

The workshop was well attended and thanks to the diverse contributions from the invited speakers and presenters of refereed work, the feedback the organizers received from participants was that a workshop devoted to evaluation is needed at this time. Consequently, a 2nd EVIA workshop to coincide with the next NTCIR (taking place in December, 2008) is already being planned.

5 References

K. Spärck Jones and C. J. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. *British Library Research and Development Report 5266*, Computer Library, University of Cambridge, 1975.