

# New Directions in Multilingual Information Access: Introduction to the Workshop at SIGIR 2006

Fredric C. Gey  
UC Data Archive  
University of California  
Berkeley, USA  
gey@berkeley.edu

Noriko Kando  
National Institute of  
Informatics, Tokyo, JAPAN  
kando@nii.ac.jp

Chin-Yew Lin  
Microsoft Research Asia  
Beijing, CHINA  
cyl@microsoft.com

Carol Peters  
Italian National Research  
Council, Pisa, ITALY  
carol.peters@isti.cnr.it

## ABSTRACT

This workshop aims at presenting the state-of-the-art in multilingual information access (MLIA) research and development. Our goal is to not only delineate current research areas but to suggest new areas for future research and development. The workshop will also focus on practical issues of scalability and practical applications of MLIA in digital libraries and web portals.

## Categories and Subject Descriptors

H.3.3 Information Search and Retrieval H.5 Information Interfaces and Presentation

## General Terms

Design, Experimentation

## Keywords

Multilingual Information Access, Cross Language Information Retrieval, Text Summarization

## 1. INTRODUCTION

At SIGIR 2002 in Finland a successful, standing-room only workshop "Cross-Language Information Retrieval: A Research Roadmap" was organized by three of the organizers of this workshop. Since 2002, research has been vigorously pursued not only in cross-language information retrieval through the Cross-Language Evaluation Forum (CLEF) and NTCIR Asian Language Retrieval and Question-answering Workshop, but also in multilingual summarization workshops and cross-language named entity extraction challenges as part of the Association for Computational Linguistics as well as the Geographic Information retrieval (GeoCLEF) tracks of CLEF. The scope for this workshop is thus consistent with the broadening of research areas in Multilingual Information Access to include cross-language question answering (CLQA) and multilingual, multi-document summarization.

*COPYRIGHT IS HELD BY THE AUTHOR/OWNER(S). SIGIR'06 WORKSHOP, AUGUST 10, 2006, SEATTLE, WASHINGTON, USA.*

## 2. ISSUES

In addition to new research directions, another important issue is how to transition the research results into practice. This has become of particular relevance because, following the announcement of enormous digital archives and digital library programs by Yahoo and Google the European Commission recently launched the i2010 Digital Libraries Initiative. Enabling multilingual access to the contents of Europe's national libraries will play a major role. At the same time, the Quero project for the development of a European search engine was announced by the French president Jacques Chirac. We wish to explore whether the research community is ready to meet the challenges posed by these major initiatives. Can current prototype systems scale up or meet the requirements of content and usage that such programs imply? What is needed to move from the lab to the real world, in terms of research, resources and equipment? How much more attention needs to be paid to presentation of multilingual results? It is time for the research and application communities to get together and examine these questions in depth.

## 3. PAPERS

Twenty-one papers were submitted to the workshop and reviewed by our program committee. We were heartened to receive several submissions about MLIA research on languages from the Indian Subcontinent, so we were able to schedule a session on "Lesser Studied Languages." Seventeen were accepted for presentation; one was later withdrawn, so sixteen papers will be presented during the one-day workshop on August 10, 2006.

## 4. CONTENT OF THE WORKSHOP

The workshop will open with a keynote address "From R&D to Practice -- Challenges to Multilingual Information Access in the Real World" by Dr. David A. Evans, CEO of Clairvoyance Corporation (formerly professor of computer science and linguistics at Carnegie Mellon). Clairvoyance has substantial practical experience in multilingual applications for both Asian and European languages.

The keynote will be followed by seven sessions. Session I will be on New Research Directions in MLIA, covering 1) New methods for cross-language access which unify multiple sources of evidence, 2) Web-based disambiguation of English-Chinese retrieval, 3) Named-entity extraction issues in MLIA applications and 4) Research issues in content extraction from multilingual statistical summary tables.

Session II will have two presentations on Lesser Studied Languages: 1) A description of new research initiatives for cross-

language search in India , and 2) A study of web-based cross-language search of Hindi using click-through information.

Session III (New Research Directions-2) will have presentations on 1) The future of multilingual summarization, 2) Document language identification under hard contexts and 3) Presentation research which would integrate multilingual and multimedia content.

Session IV will have two presentations on Interactivity and User studies: 1) Studying the use of interactive multilingual retrieval and 2) Topic-finding as a means for distinguishing the utility of cross-language search.

Session V on Evaluation will have three brief presentations on existing evaluation campaigns (CLEF for European Languages, NTCIR for Asian Languages and DUC for Multilingual Summarization) as well as one paper on a data curation approach to support evaluation studies.

Session VI will present papers on the topic of "From Research to Practice," 1) Designing multilingual information access for a national art gallery collection of digital images, 2) Implementing multilingual information access within an existing digital library system, and 3) A discussion of the practical future of MLIA within domain-specific areas such as legal information access.

The final session (VII) will be discussion among workshop participants on the road ahead for both research and practice in Multilingual Information Access.

## 5. RESULTS

We might expect at least two possible outcomes of the workshop and post-workshop activities: 1) a special issue on multilingual information access for a major IR journal (the 2002 workshop resulted in a CLIR special issue of Information Processing and Management), and 2) a white paper with guidelines for implementing multilingual information access in large digital library environments.

## 6. ACKNOWLEDGMENTS

We wish to thank our program review committee who worked hard under tight deadlines:

### ASIA:

Hsin-Hsi Chen, National Taiwan University, Taiwan,  
Kuang-hua Chen, National Taiwan University, Taiwan,  
Kazuaki Kishida, Keio University, Japan,  
Gary Geunbee Lee, Pohang University of Science & Technology, Korea  
Robert Luk, Polytechnic University of Hong Kong,  
Tetsuya Sakai Toshiba Corporate R&D Center, Japan,  
Yukata Sasaki, ATR Spoken Language Translation Research Laboratories, Japan

### EUROPE:

Maristella Agosti, University of Padua, Italy  
Martin Braschler, Zurich University of Applied Sciences, Switzerland  
Julio Gonzalo, LSI-UNED, Madrid, Spain,  
Gareth Jones, Dublin City University, Ireland

Bernardo Magnini, ITC-irst, Trento, Italy,  
Thomas Mandl, University of Hildesheim, Germany,  
Daniella Petrelli, University of Sheffield, UK  
Ari Pirkola, Tampere University, Finland  
Mark Sanderson, University of Sheffield, UK,  
Jacques Savoy, University of Neuchatel, Switzerland,  
Maarten de Rijke U. Amsterdam, The Netherlands

### NORTH AMERICA:

Aitao Chen, Yahoo Research, US  
KL Kwok, City University of New York, US  
James Mayfield, Johns Hopkins University, US,  
Isabelle Moulinier, Thomson Legal and Regulatory, US  
Jian-Yun Nie, University of Montreal, Canada  
Doug Oard, University of Maryland, US  
Miguel Ruiz, SUNY at Buffalo, US

## 7. ABOUT THE ORGANIZERS

**Fredric Gey** is co-organizer of GeoCLEF 2005 and 2006, the Geographic Information Retrieval (GIR) track of CLEF. He coordinated the English-Arabic CLIR Track at TREC (2001, 2002). He participated in TREC CLIR tracks since 1995 and in all CLEF European language campaigns and all 5 NTCIR Asian language evaluations. He was the General Chair of the ACM SIGIR'99.

**Noriko Kando** is the coordinator of the NTCIR Asian Language Retrieval evaluations NTCIR-1 through NTCIR-5 (held in Tokyo December 2005). The NTCIR evaluations include Chinese, Japanese and Korean (and English) retrieval evaluations as well as text summarization and question-answering tracks in these languages. She is an Asia/Pacific regional liaison for ACM-SIGIR Executive Committee and Program Co-Chair of SIGIR 2007. She is a Professor at the National Institute of Informatics in Tokyo Japan.

**Carol Peters** is coordinator of the Cross-Language Evaluation Forum, CLEF, sponsored by the European Commission. A main objective of CLEF is to promote the development of fully multilingual multimodal information access systems. The CLEF test suites currently include target collections in thirteen European languages. She has edited volumes of the CLEF-2000 through CLEF-2005 Proceedings, published in the Springer Lecture Notes for Computer Science series. She is a researcher with the Italian National Research Council in Pisa, Italy.