New Directions in Multilingual Information Access

Fredric C. Gey
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 attempted to present the state-of-the-art in multilingual information access (MLIA) research and development, including cross-language information retrieval and question-answering and multilingual, multi-document summarization. Our goal was to delineate current research areas as well as suggest new areas for future research and development. The workshop also focused on practical issues of scalability and practical application of MLIA in digital libraries and web portals. In addition to an invited keynote, 17 research and position papers were selected for the proceedings which may be found at http://ucdata.berkeley.edu/sigir2006-mlia.htm.

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 was thus consistent with the broadening of research areas in Multilingual Information Access to include cross-language question answering (CLQA) and multilingual, multi-document summarization.

2 ISSUES

In addition to new research directions, another issue is how to transition the research results into practice. This has become important because initiatives by Google and Yahoo have inspired the European Commission to launch an effort aimed at building The European Library. Enabling multilingual access to the contents of Europe's national libraries will play a major role in creating this Library. The Quaero project for the development of a European search engine was announced last summer by the French president Jacques Chirac. We wished 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 seemed time for the research and application communities to get together and examine these questions in depth.

3 PAPERS AND PRESENTATIONS

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 initially scheduled a session on "Lesser Studied Languages, however visa problems prevented the two authors from India from attending.

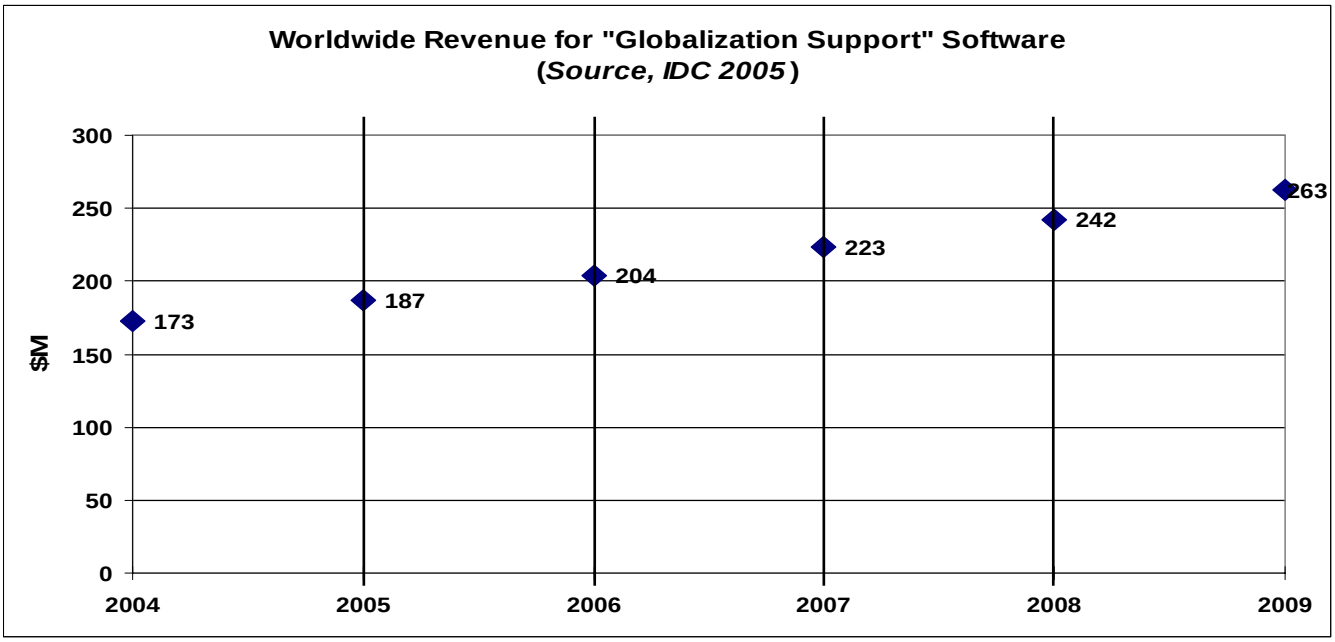
4 CONTENT OF THE WORKSHOP

In addition to an invited keynote, the workshop was organized into six sessions: two sessions on new research directions, one session on user studies and interactivity, a session on evaluation, and a session on the topic of "From Research to Practice." The final session included a panel of four researchers and practitioners leading a discussion among workshop participants on the road ahead for both research and practice in Multilingual Information Access.

4.1 Keynote Address

The workshop opened with a keynote address **From R&D to Practice -- Challenges to Multilingual Information Access in the Real World** by 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's position is summarized by this quote from its abstract "Despite the remarkable success of cross-language information retrieval (CLIR) and translingual information retrieval (TLIR) systems to perform on a par with monolingual IR systems in research and evaluation contexts, there has been relatively little commercial development (or success) of TLIR systems and applications. This is due, in part, to lack of demand in the marketplace, but also, in perhaps greater measure, to the special requirements that may be associated with TLIR applications – requirements that are not typically addressed (or assessed) in our research evaluations."

In his presentation, David Evans presented the following graph of current and projected worldwide demand for products for globalization, including multilingual search and software for machine translation. With a 5-year projected growth from 173 million US dollars (2004) to 263 million (2009), Dr Evans wryly observed "These are not numbers which excite venture capitalists."



Evans then went on to describe the failed attempts by his company and its Japanese partners to introduce multilingual aspects into existing monolingual information management products.

His conclusions from these experiences were sobering:

- "In 2006, the market for multilingual globalization support is 'not there yet'
- Quality and scope of machine translation is a major gating factor
- The demand for CLIR, *per se*, is low
- To be successful today, CLIR Systems (already very complex) must be fashioned around "solutions" – integrated into systems that may need CLIR functionality only as a means to other ends
- We must be (very) patient; or perhaps we should rethink our goals and refocus our applications"

In the discussion following the keynote, other participants from commercial firms offering similar products said that their organizations had reached the same conclusions.

4.2 Session I: New Research Directions 1

Session 1 had four papers on new genres and formats as well as improved methods for CLIR and entity extraction:

The first paper **Combining Evidence from Homologous Datasets** by Ao Feng and James Allan defined "homologous datasets" as datasets where the content was identical but the form in which the data was presented was different. Current archiving capabilities allow for information content to be available in multiple forms.: For example, there may be audio, video and text versions of a single broadcast, the content may be available in multiple languages, and may have been converted using automatic speech recognition and machine translation. The paper presented a model for combining evidence from these desperate formats which has produced improved search results.

The second paper: **Translation Disambiguation in Web-based Translation Extraction for English-Chinese CLIR** by Chengye Lu, Yue Xu and Shlomo Geva presented a method for enhanced CLIR between English and Chinese using web resources for disambiguation to choose the best translation

from among multiple translation possibilities. Previous work had focused primarily on finding any translation to resolve the out-of-vocabulary problem in query translation.

The third paper, **Named Entity Processing for Cross-lingual and Multilingual IR Applications** by César de Pablo-Sánchez, José Luis, Paloma argues that multilingual named entity extraction is a serious research challenge for the future of multilingual information access systems. Named Entities are a cheap first step that can help with multilingual processing. The goal is to look at scalable methods for web data, and approaches that are applicable to multiple languages. The approach is to use simple rule based methods, bootstrap using weakly supervised methods, and use multilingual resources like Wikipedia, and that are reusable.

The fourth paper: **Real-World Understanding for Multilingual Statistical Tables** by Fredric Gey reviewed existing research in automatic extraction of the numeric content of multidimensional statistical tables found within the text of documents or within web pages and concluded that much more needs to be done than just decoding row and column hierarchies. When transitioning to multilingual tables, issues arise which reflect different lexical forms (uses of the comma, blank and period as separators), different values of currency (dollar, euro, yuan, etc). Because major statistical agencies from the World Bank, OECD and EuroStat have devised an international interchange standard, the Statistical Data and Metadata Exchange (SDMX) which encompasses most of these semantic features of tables, the features can be used as the template for semantic extraction from existing statistical tables presented within text or web pages.

4.3 Session II: New Research Directions 2

The second research directions session presented three papers on the diverse topics of summarization, automatic language identification and multimedia.

The first paper was **The Future of Multilingual Summarization: Beyond Sentence Extraction** by David Kirk Evans (of the National Institute of Informatics of Japan). One distinct feature of multilingual text collection is that it might contain not only documents of similar statement but also of disagreement. Though this might be mainly due to sources instead of languages, traditional summarization methods that focus on finding repeated facts or statements would not work well in this scenario. Evans argued that one promising future direction for multilingual summarization would be identification of differences between documents. To support this type of summarization, one would have to move beyond simple sentence extraction since identification of differences would involve semantic or even pragmatic analysis of documents. To aid users explore these different perspectives, Evans also suggested interactive summarization systems would be one major research branch for multilingual summarization in the future.

The second paper **Identification of Document Language in Hard Contexts** by Joaquim Ferreira da Silva and Gabriel Pereira Lopes remarks that document language identification when the entire document is in one language is more or less a solved problem. It is more difficult when languages are mixed inside a single document, and when language segments are very short. So they look at characterizing documents using character n-grams. They reduce over 300,000 character sequences down to about 18 by looking at component analysis of similar documents to find out what the important character sequences are. They got about 100% precision at the document level, but down to 96% when looking at 23 character minimum size. The discriminative sequences can change depending on the context.

The third paper: **Integrated Content Presentation for Multilingual and Multimedia Information Access** by Gareth J. F. Jones and Vincent Wade suggests re-casting information retrieval as an adaptive hypermedia model. Adaptive hypermedia usually has a strong user model, and uses rich metadata to

present the information to the user. Adaptive hypermedia also uses a diverse range of content by mixing text, graphics, video, etc. One area of application of this model is clearly cross-language image search, where the relevance judgment may be made independent of language. The paper also suggests looking at foreign language documents, and for understanding link chunks of the document to similar in-language document chunks.

4.4 Session III: User Studies / Interactive Access

The session on interactive access and user studies had two papers presented which had quite different content and viewpoints.

The first paper, **Studying the Use of Interactive Multilingual Information Retrieval** by Daqing He, Douglas W. Oard and Lynne Plettenberg focused on interactive multilingual information retrieval. The main thesis was that the systems used to retrieve information are combinations of collections and machines, together with the processes that people use to search the collection(s) using the machine(s). While the effectiveness of the machines is important, it is certainly not sufficient in itself. Of real interest are the processes employed by the users of a cross-language information retrieval (CLIR) system in order to ensure that their requirements are satisfied. These processes must be investigated. Since publicly available CLIR systems are still few and far between, such user studies must of necessity be conducted in the laboratory. Fully integrated multilingual systems should be designed and then extensive user studies performed to construct an understanding of what works, what doesn't and where there is need for improvement. These studies must allow for the fact that that users are unique: there are no typical users and in every case, the users' experience, skills and background affect both their behaviour and their performance. Systems need to assist users to develop their own strategies or tactics of use depending on their own particular needs.

The second paper was **The Remarkable Search Topic-Finding Task to Share Success Stories of Cross-Language Information Retrieval** by Masashi Inoue A persistent criticism of technology transfer from MLIA research community to industry lies in the inability of current MLIA systems to meet real users' needs. Instead, Inoue attributes the difficulty to the users. Because current MLIA technologies are not generally available to potential users, users have little motivation to try them. He proposes the exploitation of compelling example usages of multi-language functionalities as a simple first step in introducing these tools to potential users. A demonstration of impressive benefits obtainable from existing MLIA systems may lead to their adoption by users. The question is: how can we find those example usages? Focusing on the cross-language information retrieval, Inoue suggests to conducting the "persuasive usage finding," or search topic finding in the case of retrieval, as a collaborative effort among the researcher community. Further, he proposes the establishment of an evaluation track as part of existing system evaluation campaigns, wherein researchers compare their algorithms in generating attractive usage scenarios automatically.

4.5 Session IV: Importance of Evaluation

The session on evaluation scheduled three presentations by the co-organizers of the workshop (no papers in the proceedings):

- **New Directions at CLEF** by Carol Peters
- **The Way Ahead of Multilingual Information Access Evaluation** by Noriko Kando
- **Multilingual Summarization at DUC and MSE** by Chin-Yew Lin

In the session on evaluation, the importance of the role of evaluation campaigns in stimulating research into many different aspects of MLIA was a focus of both presentations by Noriko Kando and Carol Peters. Kando discussed new directions for multilingual information access evaluation now being

introduced at NTCIR, stressing in particular the need to study users' information seeking behaviour and to provide user-oriented as well as system-oriented evaluation tasks. New challenges at NTCIR require more involvement of NLP techniques; the aim is to promote the development of systems that make information in documents more usable for the user according to their needs and their situation. Similarly, a main theme of the presentation by Peters was the need for evaluation campaigns to focus more on user satisfaction issues, such as query formulation and results presentation, of particular relevance in the multilingual context. She also urged that it was necessary to study ways to reduce the gap between the research and the application communities by designing tracks that addressed real-world situations and needs more clearly. Chin-Yew Lin gave an update of two recent two evaluations (MSE) focusing on multilingual summarization held in 2005 and 2006. ROUGE (Lin 2004), the de facto automatic evaluation metric, was used in MSE 2005 and 2006. Two manual evaluation phases were carried out in MSE 2006. In MSE 2005, machine translated (MT) documents were found useful to help identify main topics in a mixed MT (Arabic to English) and original English document collection though the quality of MT was still left much to be desired. In MSE 2006, reading or not reading original documents to be summarized by human assessors was shown to be an important factor in rating of summary quality. Goldstein et al. (2006) found that more automatic summaries were rated higher in quality after assessors read the entire document collection than just read the topic description of the given document collection.

The final paper of this session **A Data Curation Approach to Support In-depth Multilingual Evaluation Studies** by Maristella Agosti, Giorgio Maria Di Nunzio and Nicola Ferro proposed a new approach to the curation of the scientific data produced as a result of evaluation campaigns. The aim is to preserve the data and maintain it together with appropriate processing tools so that it is available for future in-depth studies, statistical and failure analyses of various types. In his presentation, Ferro proposed that evaluation campaigns should provide participants with more support, in particular by providing tools for post-campaign work allowing a wide range of comparative experiments and presenting the experimental datasets as objects that can be directly referenced and cited. He claimed that researchers would benefit enormously from an infrastructure of this type as data would be available for new research, existing data could be enhanced and the results made accessible, and published results could be easily validated. He concluded by illustrating DIRECT (Distributed Information Retrieval Evaluation Campaign Tool), a data curation system designed and developed at the University of Padua and now being used in CLEF for this purpose.

4.6 Session V: From Research to Practice

This fifth session had three papers on real-world application of multilingual information access in web accessible digital libraries and commercial applications.

The first paper: **Designing Multilingual Information Access to Tate Online** by Paul Clough, Jennifer Marlow and Mark Sanderson describes the challenges in providing multilingual access to online content from the Tate Gallery, Britain's premier national art gallery, which includes content from internationally-renowned artists such as Constable and Turner. The paper describes user study of what users want in a CLIR site. People were more skilled in reading foreign languages than writing them. Many of the people would like access in other languages, and are very interested in browsing as opposed to search. For cultural heritage people, designers need to think about speed of execution, availability of resources (e.g. wrappers around babelfish) copyright issues, cost and quality.

The second paper: **Implementing MLIA in an Existing DL System** by Martin Braschler, Nicola Ferro and Julie Verleyen discussed the problems involved in implementing MLIA functionality in an existing federated digital library system. At the workshop Ferro illustrated the solutions that are currently being

investigated in order to provide users of The European Library (TEL) with the possibility to access and search any of the collections of the national libraries included in the TEL federated system in their own (or preferred) language. This presentation was exemplary as it showed clearly that MLIA systems cannot be considered as plug-in additions to already running operational systems but rather need to be completely integrated into the existing system architecture if full functionality is to be provided. In fact, the light architecture adopted by TEL, with the aim of providing a low barrier of entry for the national libraries that want to join, imposes severe constraints on what can be offered from the MLIA perspective. Ferro discussed these constraints and presented the initial results of the feasibility study that has been conducted in a collaboration between TEL and the DELOS Network of Excellence for Digital Libraries.

The final paper of the workshop on **What is the Future of Multi-lingual Information Access?** was presented by Isabelle Moulinier (co-authored with Frank Schilder) of Thompson Legal and Regulatory. Thomson is a large scale professional information management firm with 40,000 employees in 45 countries. Primarily legal and regulatory sectors, learning, financial services, and scientific health care. They used to be about 10 years ago about 75% print based for revenue, and now about 42% of revenue comes from online services and search (e.g. Westlaw). The Thompson engines focus primarily on Boolean search. They are not really into CLIR because customers are not asking for it yet. The problem presented by cross-linguality is not just translation, but cultural. The legal systems vary from country to country - some countries are based on common law, some are based on regulations. There are very different approaches to how you search under those different paradigms. How do concepts map across country lines? Since their users are primarily librarians who are used to having lots of control, how do you handle things across language? There is a lot of monolingual demand though, so Thompson focuses on improving that. There is a need to move from words to concepts during translation, to go beyond document retrieval – to better understand how users interact with a cross-lingual system, to look at cultural points of view, to provide controversy detection, summarization, and question answering.

4.7 Session VI: Panel Discussion

The final discussion session was headed by a panel of participant experts with a broad range of experience in all aspects of multilingual information access.

Tetsuya Sakai from Toshiba Research and Development presented on "Possible Future Directions in Multilingual Information Access". His advice: pay more attention to information presentation, interaction, etc. What we are evaluating in some campaigns does not really exist. For example, querying in Japanese, to retrieve Chinese documents back - but in reality the Japanese person can't read those!) A new approach would be to change the nature of querying, and develop a sequence of queries to get to the information that user wants, ending in the result "ok, give me a summary of the information that has been found." On evaluation, he questioned whether off-line evaluation as practiced in TREC, CLEF and NTCIR was realistic, however interactive evaluations are expensive, time-consuming, and error-prone. His proposal: Do off-line evaluation 80 percent of the time and online, interactive evaluation 20 percent of the time and make sure that the results of the separate evaluations are consistent.

Mark Sanderson from University of Sheffield, UK: In CLEF there is a focus on mean average precision, but does that really reflect real-world performance and what users want to do? TREC is planning on moving to several thousand query track. What is the correct measure to use? It only takes one bad query for the user to decide there is a problem. Evaluation measures such as MAP place emphasis on pseudo-relevance feedback, but does that really help in the real world? What about offering up different types of search results depending on the interpretation of the query? Finally, culture and cultural heritage is an interesting application area, particularly for academics. English is becoming the

minority of the worlds spoken languages -- sooner or later English speakers might need more help in accessing non-English information sources.

David A. Evans, Clairvoyance Corporation: A healthy tension has emerged between science and commercial applications. What we need for good science isn't necessarily what we need for good applications: the community should not lose sight of what that is. In the real world, many people don't know what is needed for good science (and don't care.) As a community we have allowed ourselves to play the numbers game, to improve by x% at some significance level, but we should really look at what we need to do to evaluate real performance. We need to have shared resources. Certainly test collections are necessary, but we need resources for replication. Need good lexical and NLP resources for each language. Standard statistical evaluation packages are important with good instructions on how to use them. We need to have good models for explaining why things are the way they are, so good explanatory theories. What do we need for good applications? Need to be able to avail ourselves of reliable resources and techniques. Need clear roles in solving problems. Applications don't exist without a user community and a focus, a user need to solve problems, so we need to define those problems and be realistic about it. Some of the bridge problems which may be useful involve messages, blogs, fragments and excerpts (because these can be useful in the solution space and commercial space.) CLEF and NTCIR do a good job with patents and medical information, but what about advertisements, contracts, insurance documents.

Doug Oard of the University of Maryland presented: "4 good ideas from today". At SIGIR 1996 in the first workshop on cross-language research, there were two things that participants agreed opon:

1. The cross-language document problem was correct problem to work on,
2. It should be called cross-language and not cross-lingual. We've regretted both since then.

Now we think that CLIR is pretty much solved, but maybe it isn't the best problem to focus on. We don't know much about what the rest of the story is. To set the stage, here are four proposals:

1. *Death to average precision*. Ranked lists are a type of multi-document summary, let's try some summarization measures.
2. *Study fully integrated systems*. Users are control freaks. Study the process of retrieval and not just the results. We are in enormous danger of doing bad science with user studies, but we have to at least jump in and try doing it. Should think before taking a step, but we do have to start working on it.
3. *Study cross-cultural retrieval:* CLIR requires matching meaning. Machine translation requires re-expressing meaning. Communication requires negotiating meaning. We should look at cross-cultural communication.
4. *Eat your own dog food:* let's make a searchable archive of CLEF or NTCIR records. Help develop an archival practice. Illustrate the use of our technology in a multinational enterprise.

5 MAIN THEMES EMERGING FROM THE WORKSHOP

From the workshop and the discussion by participants, a number of different themes emerged to both challenge researchers in this area and suggest new avenues of research and development:

1. Identification of the real world (commercial) use case? This was a main topic of the keynote paper and a running theme throughout the workshop.
2. The difficulty of technology transfer into existing applications (see the paper Braschler et al).
3. The need for new evaluation methodologies which evaluate the whole system including aspects relating to usage and not just system performance from the technical perspective

4. Following on from 3, the importance of replication of research results, implying a need for appropriate tools (see the paper on Data Curation)
5. Again related to 3, the need to study user behaviour - and to define reliable techniques for such studies (papers by He/Oard, and Clough et al)
6. The need to study the relationship between cross-language retrieval, machine translation and multilingual summarization (see presentations by Chin-Yew Lin & D.K.Evans))
7. The pressure to move studies from text to mixed media and "new" genres (see Ao Feng, Gey, Jones)
8. The need for digital content to be prepared with access in mind – and the need for markup. The potential of the semantic web for CLIR must be investigated (information processing and retrieval must move from words or features to concepts via markup)

A final conclusion was that CLIR research (and the stimulus provided by evaluation campaigns) has provided valuable insight into the IR problem in general, so the community should be optimistic about the future despite the current obstacles regarding transition to applications and commercialisation.

6 ACKNOWLEDGMENTS

Very helpful notes by David Kirk Evans were provided which considerably eased the task of summarizing some of the papers presented. The organizers wish to thank our transcontinental (Asia, Europe, North America) program review committee who worked hard under tight deadlines. Their names may be found on the workshop web site.

7 REFERENCES

Goldstein, Jade, Lucy Vanderwende, and Liang Zhou (2006). Multilingual Summarization Evaluation 2006 (MSE 2006). http://research.microsoft.com/~lucyv/MSE2006.htm

Lin, Chin-Yew. (2004). ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.