Information Access in a Multilingual World:

Transitioning from Research to Real-World Applications

Fredric Gey
University of California, Berkeley USA
Jussi Karlgren
Swedish Institute of Computer Science, Stockholm, SWEDEN
Noriko Kando
National Institute of Informatics, Tokyo, JAPAN
gey@berkeley.edu, jussi@sics.se, kando@nii.ac.jp

Abstract

Multilingual Information Access (MLIA) is at a turning point wherein substantial real-world applications are being introduced after fifteen years of research into cross-language information retrieval, question answering, statistical machine translation and named entity recognition. Previous workshops on this topic have focused on research and small-scale applications. The focus of this workshop was on technology transfer from research to applications and on what future research needs to be done which facilitates MLIA in an increasingly connected multilingual world.

1 Introduction and Overview

The workshop*: Information Access in a Multilingual World: Transitioning from Research to Real-World Applications* was held at SIGIR 2009 in Boston, July 23, 2009. The workshop was held in cooperation with the InfoPlosion Project of Japan[1] The workshop was the third workshop on the topic of multilingual information access held at SIGIR conferences this decade. The first, at SIGIR 2002 in Tampere, was on the topic of "Cross Language Information Retrieval: A Research Roadmap". The second was at SIGIR 2006 on the topic of "New Directions in Multilingual Information Access". Over the past decade the field has matured and significant real world applications have appeared. Our goal in this 2009 workshop was to collate experiences and plans for the real-world application of multilingual technology to information access. Our aim was to identify the remaining barriers to practical multilingual information access, both technological and from the point of view of user interaction. We were fortunate to obtain as invited keynote speaker Dr Ralf Steinberger of the Joint Research Centre (JRC) of the European Commission, presenting the Joint Research Centre's multilingual media monitoring and analysis applications, including NewsExplorer. Dr. Steinberger provided an overview paper about their family of applications, which was the first paper in the workshop proceedings.

[1] http://www.infoplosion.nii.ac.jp/info-plosion/ctr.php/m/IndexEng/a/Index/

In our call for papers we specified two types of papers, research papers and position papers. Of the 15 papers initially submitted, two were withdrawn and two were rejected. We accepted 3 research papers and 8 position papers, covering topics from evaluation (of image indexing and of cross-language information retrieval in general), Wikipedia and trust, news site characterization, multilinguality in digital libraries, multilingual user interface design, access to less commonly taught languages (e.g. Indian subcontinent languages), implementation and application to health care. We feel these papers represent a cross-section of the work remaining to be done in moving toward full information access in a multilingual world.

2 Keynote Address

The opening session was the keynote address on "Europe Media Monitoring Family of Applications." Dr. Ralf Steinberger presented a detailed overview of a major initiative of the European Commission's Joint Research Center at Ispra, Italy to provide just-in-time access to large scale worldwide news feeds in approximately 50 languages. At the heart of the system is the Europe Media Monitor news data acquisition from about 2,200 web news sources to gather between 80,000 and 100,000 news articles daily (on average). The 'monitor' visits news web sites up to every five minutes for latest news articles. The news gathering engine feeds its articles into four public news analysis systems:

- **NewsBrief** – which provides real-time (every ten minutes) news clustering and classification, breaking news detection, and an email subscription facility
- **MedISys** – a real-time system which filters out only news reports of a public health nature, including threats of chemical, biological, radiological and nuclear nature
- **NewsExplorer** – which displays a daily clustered view of the major news items for each of the 19 languages covered, performs a long-term trend analysis, and offers entity pages showing information gathered in the course of years for each entity, including person titles, multilingual name variants, reported speech quotations, and relations. Languages cover 14 European Union languages plus Arabic, Farsi, Norwegian, Russian, and Turkish.
- **EMM-Labs** – which includes a suite of tools for media-focused text mining and visualization, including various map representation of the news, multilingual event extraction, and social network browsers.

3 Research Papers

The research paper by Nettleton, Marcos, and Mesa-Lao of Barcelona, Spain, "The Assignment of Tags to Images in Internet: Language Skill Evaluation" was presented by Ricardo Baeza-Yates. The authors had performed a study on differences between native and non-native users when labeling images with verbal tags. One of the results presented was that the diversity was lower for non-native users, reasonably explained through their relatively smaller vocabulary. The authors studied tags related to concrete image characteristics separately from tags related to emotions evoked by the image: they found, again reasonable in view of likely relative exposure of users to concrete and abstract terminology, that the difference was greater for evocative terms than for concrete visual terms. This study elegantly demonstrated the limits of linguistic competence between native and non-native, simultaneously giving rise to discussion of which usage is the more desirable in a tagging application: do we really wish to afford users the full freedom to choose any term, when many users are likely to be content with a more constrained variation in terminology?

Elena Filatova of Fordham University USA, presented her paper on "Multilingual Wikipedia, Summarization, and Information Trustworthiness**."** Her experiment showed how a multilingual

resource such as Wikipedia can be leveraged to serve as a summarization tool: sentences were matched across languages using an established algorithm to find similarities across languages. Sentences that were represented in many languages were judged as more useful for the purposes of the summary than others. This judgment was verified by having readers assess the quality of summaries. The research corpus was a subset of Wikipedia on biographies utilized in the DUC (Document Understanding Conference) 2004 evaluation.

The paper "A Virtual Evaluation Track for Cross Language Link Discovery" by Huang, Trotman and Geva was presented by Shlomo Geva of Queensland University of Technology, Australia. The authors propose a new evaluation shared task for INEX, NTCIR and CLEF, where participating projects will contribute towards an interlinked universe of shared information across languages, based on internet materials. The objective is to create a low-footprint evaluation campaign, which can be performed off-line, asynchronously, and in a distributed fashion.

4 Position Papers

Masaharu Yoshioka of Hokkaido University, Japan presented a paper on "NSContrast: An Exploratory News Article Analysis System that Characterizes the Differences between News Sites" Yoshioka's idea was that news sites from different countries in different languages might provide unique viewpoints of reporting the same news stories. The NSContrast system uses "contrast set mining (which) aims to extract the characteristic information about each news site by performing term co-occurrence analysis." To test the ideas, a news article database was assembled from China, Japan, Korea and the USA (representing the 4 languages of these countries). In order to compensate for poor or missing translation, Wikipedia in these languages was mined for named entity translation equivalents.

John Tait of the Information Retrieval Facility in Vienna, Austria, presented a provocative view of "What's wrong with Cross-Lingual IR?" Tait argued that laboratory-based evaluations as found in TREC and other evaluation campaigns have limited generalizability to large scale real-world application venues. In particular, patent searches within the patent intellectual property domain involve a complex and iterative process. Searches have a heavy recall emphasis to validate (or invalidate) patent applications. Moreover, in order to validate the novelty of a patent application, patents in any language must be searched, but the current dominance is with English, Japanese, and possibly Korean. In the future, Chinese will become a major patent language for search focus.

Jiangpen Chen presented her paper co-authored with Miguel Ruiz "Towards an Integrative Approach to Cross-Language Information Access for Digital Libraries." The paper described a range of services which are and might be provided by digital libraries, including multilingual information access. The authors described an integrative cross-lingual information access framework in which cross-language search was supplemented by translational knowledge which integrates different resources to develop a lexical knowledge base by enlisting, among other, the users of the systems to participate in the development of the system capability. Chen's presentation provided a number of example systems which provided some level of bilingual capability upon which future systems might be modeled.

Michael Yoshitaka Erlewine of Mozilla Labs (now at MIT in Linguistics) presented a paper "Ubiquity: Designing a Multilingual Natural Language Interface" about the development of a multilingual textual interface for the Firefox browser which aims at an internationalizable natural language interface which aligns with each "user's natural intuitions about their own language's syntax." The shared vision is that we can put theoretical linguistic insights into practice in creating a

user interface (and underlying search and browse capability) that provides a universal language parser with minimal settings for a particular language.

Fredric Gey of the University of California, Berkeley (one of the workshop organizers) presented a paper on "Romanization – An Untapped Resource for Out-of-Vocabulary Machine Translation for CLIR." The paper noted that rule-based transliteration (Romanization) of non-European scripts has been devised for over 55 languages by the USA Library of Congress for cataloging books written in non-latin scripts, including many variations of Cyrillic and the Devanagiri scripts of most Indian sub-continent languages. The paper argued that rule-based Romanization could be combined with approximate string matching to provide cross-lingual named entity recognition for borrowed words (names) which have not yet made it into general bilingual dictionaries or machine-translation software resources. The approach should be especially beneficial for less resourced languages for which parallel corpora are unavailable.

Kashif Riaz of the University of Minnesota presented a paper "Urdu is not Hindi for Information Access." The paper argued for separate research and development for the Urdu language instead of piggy-backing on tools developed for the Hindi language. Urdu, the national language of Pakistan, and Hindi, the major national language of India, share a major common spoken vocabulary such that speakers of each language can be as well-understood by speakers of the other language as if they were dialects of a common language – however written Urdu is represented by the Arabic script while written Hindi is represented by a Devanagari script. The paper differentiates the separate cultural heritage of each language and argues for significant additional and independent natural language processing development for the Urdu language.

The paper "A Patient Support System based on Crosslingual IR and Semi-supervised Learning" by Isozaki and others of NTT Communication Science Laboratories Kyoto, Japan, was presented by Hideki Isozaki. The authors are constructing a system for aiding medical patients in their quest for information concerning their condition, including treatments, medications and trends in treatments. Because considerable medical information is available in English, the system incorporates a cross-language retrieval module from Japanese to English. The content being accessed is both technical articles (PubMed) and patient-run web, government information sites focused on medical conditions and local information about doctors and surgeons. For technical terms which may not be understood or used by patients, the system provides a synonym generator from lay terms to medical terminology. The system's cross-language goal is to analyze multiple English medical documents "with information extraction/data mining technologies" to generate a Japanese survey summarizing the analysis. Currently the system supports medical literature searches (which have high credibility) and is in the process of expanding to patient sites for which credibility judgment criteria and methods will need to be developed.

5 Discussion of the Future of Multilingual Information Access

The final session was a free-ranging discussion of future research needs and the remaining barriers to widespread adoption of well-researched techniques in multilingual information access into real-world applications.

Discussion on what usage needs to be supported by future systems for cross-lingual information access took as its starting point the question of what usage scenarios specifically need technical support. The requirements for professional information analysts with a working knowledge of several languages are different from the needs of lay users with no or little knowledge of any second language beyond their own and with only passing knowledge of the task under consideration. Most of

the projects presented here did not explicitly address use cases, nor did they formulate any specific scenario of use, other than through implicit design. The long time failure of machine translation systems was mentioned as a negative example: engineering efforts were directed towards the goal of fluent, high quality sentence-by-sentence translation which in fact seldom has been a bottleneck for human language users. The alternative view, held by many, is that most users have been satisfied by approximate translations which convey the content of the original document.

The suggestion was put forth that the field of cross-lingual information access might be best served by a somewhat more systematic approach to modelling the client they are building the system for; that would in turn better inform the technology under consideration and allow system building project to share resources and evaluation mechanisms.

Action items suggested were, among others, creation of a permanent web site dedicated to research and development of multilingual information access. The first task of the web site would be to accumulate and identify available multilingual corpora to be widely distributed as a goal of further development of equal access to information regardless of language.

6 Conclusion

This workshop recognized that the time has come for the significant body of research on cross-language retrieval, translation and named entity recognition to be incorporated into working systems which are scalable and serve real customers. Two example systems were presented, news summarization (by the keynote speaker) and by researchers trying to provide information support for medical patients. In addition another speaker provided an architecture for integrating multilingual information access within the digital library environment, and one presentation suggested a distributed, low-footprint shared task for evaluation purposes. The discussion sessions generated directions and suggested next steps toward this agenda of developing real-world application systems.

These next steps necessarily will involve sharing experiences of real-world deployment and usage across systems and projects. To best encourage and accommodate such joint efforts, those experiences must be documented, published, and presented in some common forum. If evaluation is to proceed beyond system benchmarking, finding and leveraging these common real-world experiences are crucial to achieve valid and sustainable progress for future projects.