CROSS LANGUAGE INFORMATION RETRIEVAL: A RESEARCH ROADMAP

Summary of a Workshop at
SIGIR-2002: 22nd Internationl Conference On Research And Development in Information Retrieval
August 15, 2002, Tampere Finland

Web site: http://ucdata.berkeley.edu/sigir-2002

Fredric Gey[1], Noriko Kando[2] and Carol Peters[3], Organizers
(summary produced October 15, 2002)

Cross-Language Information Retrieval (CLIR) has been a research sub-field for more than a decade now. The field has sparked three major evaluation efforts: the TREC Cross Language Track which currently focuses on the Arabic language, the Cross-Language Evaluation Forum (CLEF) – a spinoff from TREC - covering many European languages, and the NTCIR Asian Language Evaluation (covering Chinese, Japanese and Korean). During this one-day workshop we reviewed and assessed the progress that has been made so far and discussed what research and development remains to be done to make CLIR a practical enterprise. Presentations focused on the major techniques and accomplishments of the field (e.g. utilization of corpus, dictionary, and machine translation techniques for crossing language barriers, strategies for sense disambiguation and query expansion) and position papers suggested the directions that research should take in the next half decade. The goal of our workshop was to develop a step-by-step, year-by-year roadmap of research to be undertaken, with each year addressing progressively more difficult goals and expected accomplishments. While the workshop produced suggestions for such a roadmap, unfortunately, time was too short to actually develop a practical plan

Our call for position papers brought 15 submissions, four of which were chosen by the organizers and program committee as main talks and papers:

- **"When You Come to a Fork in the Road, Take It: Multiple Futures for CLIR Research",** by Douglas Oard (University of Maryland, USA)
- **"Towards a Unified Approach to CLIR and Multilingual IR",** by Jian-Yun Nie (University of Montreal, Canada)
- **"Cross-Language Information Retrieval: Consolidating and Moving Forwards",** by Gareth Jones (University of Exeter, United Kingdom)
- **"Three Principles to Guide CLIR Research",** by James Mayfield and Paul McNamee (Johns Hopkins University APL, USA)

The other papers were allotted a short summary time in the program. The workshop was organized into six thematic sessions: **Approaches to CLIR** described various techniques which

[1] University of California, Berkeley, USA, gey@ucdata.berkeley.edu

[2] National Institute of Informatics, Japan, kando@nii.ac.jp

[3] Italian National Research Council, Italy, carol@iei.pi.cnr.it

have been applied to CLIR in the past, including query translation, pivot languages, and thesauri, with speculation for the future. **Strategies for Languages with Little Resources** described techniques for languages for which there are few linguistic resources available, with examples from Indonesian, Tamil and Zulu, and also included a proposal for the standardization of lexical resources for CLIR. The **Multimedia** session had two papers which discussed CLIR for image and speech retrieval across languages. **User Studies/Interactive** presented two papers on the role of user interaction in CLIR. A session on **Evaluation** described the Cross-Language Information Retrieval evaluations (CLEF and NTCIR) underway in Europe and Japan and discussed their contribution to CLIR research and development. The final session **Building a Roadmap** began with a main talk in which a detailed five-year plan for research was outlined; this was followed by participant discussion and a review of the entire day of presentations.

Challenges:

At the beginning of the workshop the organizers presented three challenges:

1. **Where to get resources for resource-poor languages –**outside of the most spoken languages of Europe (English, French, German, Italian, Spanish) and Asia (Chinese and Japanese) or the additional official languages of the United Nations (Arabic and Russian), resources in terms of parallel corpora or commercial machine translation are very difficult to obtain. In particular, the languages of the Indian subcontinent have received very little attention.
2. **Why do we not have a sizeable Web corpus in multiple languages?** -- aside from the issues of cost of construction and maintaining realistic links (which have taken several years to be addressed by the TREC Web track for the English languages), we have the complication of English language dominance (approximately 70-75 percent of web pages currently) and low percentage representation beyond the top ten languages, as well as lack of standards for character and font representation for many other languages. Chinese has at least two major representations (GB and BIG5) and Japanese three, while for Indian subcontinent languages standards are only beginning to be developed (i.e. each site has its own font and internal character representation). This means that if English is included a ranked list of pages will be dominated by pages in English and many languages will not even make in the top 100 pages found. Work is clearly needed here in order to define suitable criteria for the construction of a valid multilingual Web corpus for R&D.
3. **Why aren't search engines using our research?** – several search engines now offer monolingual search in a number of languages coupled with machine translation software to translate pages into English (AltaVista and GOOGLE are prominent examples). Cross-language search would seem to be a natural extension of these offerings. Part of the answer is found in the question of utility – if users are presented with a ranked list of documents that they cannot read, what is the utility? An exacerbating factor is in the weakness of current machine translation software to be applied to the pages found.

Keynote Address:

In his talk **"When You Come to a Fork in the Road, Take It: Multiple Futures for CLIR Research,"** Doug Oard opened with the rhetorical statement "in 2002 CLIR is a solved problem!" We have achieved nearly 100% of monolingual effectiveness in many tests (European languages, Chinese, Arabic). A small bag of tricks works across many languages (stemming, stop-word translation, term selection, segmentation for character languages, weight mapping from one language to another, etc). We seem to have adequate resources for many language pairs (bilingual term lists, monolingual corpora, parallel Web text and software to mine the web). The research success rests on several factors: 1) there was an undeniable need; 2) most of the component technology (IR, NLP) already existed; 3) there was a low cost to entry (which led to broad participation); 4) there was a widely-accepted evaluation methodology (so good ideas could be easily recognized). If we accept this scenario, how are we to convince the funding community to pour another $50M in funding into the research area (Oard's estimate of the cost to support research in CLIR between 1996 and today). The question to be asked (and solved) is if the R&D has been a success then why hasn't the technology been transferred to general use? There seem to be three reasons: 1) Utility – what do users do with a ranked list they cannot read? 2) Efficiency and integration – little work has been done on efficiency of multilingual indexing, and research has usually separated the translation component from the retrieval component. 3) Genre – research to date has been mostly about news corpora in textual form – little has been done with web pages or scientific and technical corpora (i.e. patents, case law). Other relevant issues are that research has concentrated almost exclusively on the text modality, with little attention to cross-language retrieval of spoken documents or OCR (there are still a lot of paper documents out there in many languages). A successful future for this research will rest upon collaboration across research communities (web search, speech recognition, OCR, text summarization, data mining, etc.) and user communities (medical, legal, humanities, etc.), further lowering the barrier to entry (supply re-usable software components and resources), and crafting a compelling message of need.

Session 1: Approaches to CLIR:

In his introduction to this session, "**Towards a Unified Approach to CLIR and Multilingual IR",** Jian-Yun Nie argued that current CLIR approaches are deficient for several reasons:

- Translation and retrieval are decoupled, thus
- Translation is often decoupled from the corpora being retrieved
- Languages are retrieved independently, then merged

Nie argued for tighter coupling of translation and retrieval into a unified probabilistic model: steps in this direction have already been made by Kraaij and others at TNO/TPD and University of Twente and Xu and Weischedel at BBN. It is noticeable that monolingual retrieval has become language dependent, relying upon specialized stemmers and stop-words. This means that attention is directed independently to each language without consideration of the other languages being searched. A consequence is that the merging process is deficient because it is carried out without information from the translation and retrieval processes. In the future Nie claims that a unified approach is required for CLIR; one in which language characteristics are considered as additional parameters which specify a document collection, rather than as constituting a barrier to collection cohesion Nie also argued for the development of a multilingual Web collection upon which unified models could be experimented.

The paper **"UTACLIR - An Extendable Query Translation System" by** Turid Hedlund, Heikki Keskustalo, Eija Airio, and Ari Pirkola, described an extendable architecture in which bilingual dictionaries could be rapidly incorporated for languages where machine translation capabilities have not yet been developed. The paper **"Cross-Language Information Retrieval Based on Multilingual Thesauri**" by Natalia Loukachevitch and Boris Dobrov described a system at Moscow State University which has developed a special "information retrieval thesaurus' and argued that IR and CLIR systems based upon traditional thesauri were inadequate and that development of a useful thesaurus needs to be corpus-based. The paper **" Translation via a Pivot Language Challenges Direct Translation in CLIR**" by Raija Lehtokangas and Eija Airio described experiment in the use of pivot languages where direct translation pairs are not available and concluded that CLIR using transitive translation via pivot languages showed exceptional promise.

Session 2: Strategies for Languages with Little Resources:

This session had four papers submitted but only two presenters were able to attend the workshop. The summary will cover all four papers. In the paper and presentation **"Starting from Nothing: Resources of First Resort in CLIR",** Fredric Gey showed examples from the Tamil language spoken by 70 million people in southeast India and argued that even with a language with no resources there might be a methodology to achieve some results. The methodology would be fuzzy or phonetic matching with languages using the Roman alphabet, similar to the work by Buckley in TREC-6 which considered French words to be just English words miss-spelled. For scripted languages like Tamil, the scripts can be first transliterated (or Romanized) into the Roman alphabet and then searched with phonetic matching techniques. In the paper and presentation **"CLIR Access in Indigenous Languages: a case study of the Zulu language",** Erica Cosijn, Ari Pirkola, Antti-Pekka Känsälä , and Theo Bothma argued that increasing attention is being paid to indigenous populations and languages for cultural and medical reasons and that it is important for the CLIR community to devise approaches to deal with those languages for which there will be few resources. CLIR can become the route to accessing Indigenous Knowledge (information about medicines, food preparation, natural resource management). The paper presents a case study of the Zulu language, the most widely spoken of nine indigenous languages of South Africa, with challenges of inflection, lexical ambiguity, paraphrasing and borrowed words. The paper **"Evaluating Indonesian Resources for CLIR"** by Mirna Adriani presents research on bilingual retrieval from English to Indonesian, using freely available but limited dictionaries found on the WWW and with lexicons constructed from parallel corpora of news articles. The paper **"Computational Lexicons for Cross-Language Information Retrieval"** by Nicoletta Calzolari and Allessandro Lenci presents a series of requirements to which computational lexicons must adhere in order to be effective resources for CLIR and a model framework for evaluating these requirements, with a discussion of how EuroWordNet fits into the model.

Session 3: Multimedia

This session had two papers. The main paper **"Cross-Language Information Retrieval: Consolidating and Moving Forwards"** by Gareth Jones presented a summary of the progress to date in CLIR in which he remarked that average precision as a performance measure hides the vast differences in performance for particular queries and that further analysis of why some queries are challenging is called for. In addition he stated that there are logistical barriers to

progress in CLIR – the field needs more resource and expertise sharing (for example non-English search engines) and that publication is widely dispersed, making it hard to have a grasp of what really works best. He suggested a future in which many new tasks are now being investigated in monolingual IR: multimedia, question-answering, summarization and web retrieval; all these tasks could extended into CLIR tasks. The question to be asked is whether this would be worthwhile and whether the monolingual (primarily English) versions of the technologies are sufficiently mature at this time to be extended. In the case of multimedia, technologies are developing for speech, digital images and video with existing retrieval work in the first two areas and some developing work in video. Indeed, some aspects of such retrieval are language independent (images and video can often be understood independent of language) and are amenable to query-by-example: "find me more objects that sound/look like this". However in the area of speech, the challenge of retrieving speech content is compounded by errors in speech recognition technology, e.g. if the actual spoken item "the meeting begins at two thirty" is transcribed automatically as "the meeting begins too the day" the point of the message is lost. The research question to be addressed in speech CLIR is the quantification of information loss due to transcription errors versus translation errors and whether research in this area could lead to the development of better tools which exploit this interaction to achieve better performance. Jones also remarked that there are formidable logistical barriers to research in these areas – the lack of availability of suitable corpora for evaluation (often potentially available material is perceived to have commercial value by its owners and acquisition cost becomes a barrier), and the fact that the technology required for indexing is scarce and expensive. In addition, as with the commercial machine translation technology, use of commercial out-of-the-box speech recognition precludes "under the hood" investigation. In the paper **"EuroVision: An Image-Based CLIR System",** Mark Sanderson and Paul Clough describe their EuroVision project which aims at CLIR for an image collection using textual headings associated with the collection. The advantage is that the content can often be understood without knowledge of the text associated with the image. In addition, image libraries might be able to charge for content, which means that investment in the technology can be capitalized. Among the research challenges are that image captions are often short (and thus may need more language processing than longer textual corpora used for CLIR), and image queries seem to contain more significant verbs than has been the case with traditional CLIR to date. Finally, the fact that traditional image content retrieval (either through vectorization of image shapes or color histograms) is not going to solve the retrieval problem soon means that image CLIR has the potential for extensive use and study of user feedback.

Session 4: User Studies and Interactive CLIR

The two papers in this session presented ideas and challenges (to the common assumptions):

- Where are the postulated monolingual searchers? – in Europe many users are polyglot (Petrelli)
- English is used in combination with in other languages – should be incorporated into extended search engines (Petrelli)
- Users adapt to systems capabilities easily, but not to the document language (Gonzalo)
- Interactive Cross-Language question answering is more realistic than monolingual question answering (Gonzalo)

In the paper (and presentation) **"Should the Real Use Affect CLIR Research**", Daniela Petrelli, Micheline Beaulieu and Mark Sanderson say that "It would appear that little effort is being made to identify who the users of CLIR systems are and to fully understand how actual users can make use of such systems". In a study of actual users of a CLIR system in England, the authors found the following surprise results: most users were not solely monolingual but polyglot with a smattering of knowledge of several languages and that English is commonly used as a tool when searching other languages because of its international relevance in technical jargon. The implication is that real CLIR systems should allow multilingual queries as well as retrieve documents in more than one language**.** In addition, they found that users' search is use oriented – and their search skills are very low; users search many languages simultaneously; they swap between the known languages, choosing the most appropriate language for search; they use English as pivot language in posing multi-language queries; they have a need to search phrases and proper names; and finally user-created dictionaries can play an important role in CLIR. In his paper and presentation **"Scenarios for Interactive CLIR Systems",** Julio Gonzalo asserted that users adapt to interactive systems capabilities quite easily, but not to an unknown document language. In order to perform evaluation guided interactive CLIR research, questions must be addressed which are amenable to comparative evaluation. Among such dimensions are:

- Language knowledge (e.g. null (Korean), passive (European), active (user's native or second language)
- Media (text, image captions, speech, etc.)
- Information need (bibliographic search, web surfing, question-answering, etc.)

Gonzalo proposes that the next few years of CLEF evaluation campaigns should include a track for Interactive Cross-Language Question Answering which would include not only answer retrieval and evaluation but also application of the answer within an application task. Such systems would be realistic (more realistic than English QA, at least in Europe), challenging, with feasible comparative evaluation attracting a potentially wide research community. For CLEF-2004 he proposed that CLQA be done on the internet using a 1) common question set, 2) common internet search engine (GOOGLE, ALTAVISTA?), 3) Latin-square design of systems/users/questions. The results might be evaluated by having a questionnaire which answers the question set filled in by the searchers. The evaluation would combine accuracy of answers with amount of time and number of interactions with the system.

Session 5: Evaluation

This session focused on the relationship between formal evaluation campaigns and cross-language system development. In her presentation **"The Contribution of Evaluation: the CLEF Experience",** Carol Peters observed that the activity of the Cross-Language Evaluation Forum has shown that an evaluation initiative can provide far more than just a benchmarking infrastructure. In fact, over the last six years CLEF has provided the CLIR research community with a service that goes far beyond the mere provision of evaluation test-suites, offering common ground for the discussion of approaches and ideas that can lead not only to collaborative work and/or the exchange of tools and resources between groups working on similar problems but also proposals for new lines of research. Peters felt, however, that if CLEF was to continue to satisfy the emerging requirements of the R&D community, it must shift its main focus from cross-language text retrieval and the measuring of system performance in the terms of document rankings to the provision of a comprehensive set of tasks covering all aspects of multilingual, multimedia system performance with particular attention being paid to the needs of the end-user.

In her presentation **"Evaluation – the Way Ahead: the case of the NTCIR",** Noriko Kando utilized the succession of Asian language retrieval evaluation workshops (NTCIR) to generalize about the role and utility of evaluation workshops in stimulating research and transferring technology. An evaluation workshop is characterized by a set of data and unified evaluation procedures, wherein each participant conducts research with their own approach using the data provided. In this way a wide variety of approaches are tested on identical data sets in a forum designed to learn from the experience of others. Kando described the following products of successful evaluation workshops: 1) the creation of large-scale test collections 2) a forum for research idea exchange and technology transfer 3) a "showcase" for new technologies 4) motivation for research 5) discussion of evaluation methods 6) development of experimental design models and 7) facilitation of newcomers to the research area. She presented the following dimensions by which to characterize CLIR systems: 1) languages covered 2) type of media 3) tasks and users 4) relevance judgments or success criteria 4) document genres 5) layers of CLIR technologies and 6) the information access processes. In its series of three workshops NTCIR has evolved from a single language pair (Japanese – English) to four languages (Chinese, English, Japanese, Korean) and is moving in the direction of cross-language information access in question answering, summarization, text mining, and domain specific retrieval (e.g. patents and scientific domains).

Session 6: Building a Roadmap

In the final session, **Building a Roadmap,** the presentation **"Three Principles to Guide CLIR Research,"** laid the groundwork for open discussion in which the community of researchers began the process of outlining a roadmap of research necessary to move the field to a new level. In their paper and presentation, James Mayfield and Paul McNamee enunciated the following principles: "1) CLIR = CL X IR" or Cross-language information retrieval is equivalent (in the probabilistic sense) to the assumption of independence between the translation step and the retrieval step. Cross-language IR is primarily dependent on the correctness of translation. There is no compelling evidence (i.e. no comparative research has been done with constant resources) either for or against integration of the steps. "2) CLIR > CLDR" or evaluation using the standard of mean average precision is insufficient – it hides detail and the possibility that query subclasses have systematic errors which may be corrected; in addition it treats all errors equally while users may have priorities for error types. The research community should investigate whether other evaluation types may exist which could elucidate these questions. "3) MLIR ≠ BLIR" or multilingual retrieval is different in nature from bilingual retrieval – results merging (which is unnecessary for a single target document collection) is an unsolved problem in MLIR. Probabilistic models no longer hold because you need to compute query prior relevance estimates by collection. The problem is similar to metasearch, but more pronounced for Multi-Lingual Information Retrieval. The presenters also enumerated five dangers facing the research community:

1. Perceived barriers to entry
2. Availability of language resources
3. Waning interest among researchers
4. Waning interest among funders
5. Unclear path to usefulness

An ambitious 5 year roadmap for research and resource goals was outlined, concluding in year 4 with "Evaluation of multilingual retrieval in 15 or more languages, including Asian, European,

Indic and Semitic languages" and in year 5 with "A global WordNet available in 15 languages with a kernel of 100,000 synsets in each language" and "Evaluation of cross-language speech retrieval in four or more languages attracting at least 10 participating groups."

Discussion

During and after each session, discussion points arose and remarks were made which are incompletely summarized below. For languages with few resources, spoken language recognition and transcription into text may become paramount because of the low literacy rates in countries and areas where the languages are spoken. In addition, in CLIR for all languages the problem of proper noun recognition assumes a larger role (as, for example, when Kurt Waldheim is translated to "Kurt forest home"), In evaluation campaigns we need to understand their relationship to technology transfer. Since test collection costs are high, they should leverage a large amount of research to be viable (some of the domain-specific tasks have had difficulty attracting participants). There is a critical need for the data (and resources) to be made available after the workshops.

The final discussion revolved around the questions: Have we solved the CLIR problem? Have we identified the CLIR problem? Do we need a better understanding of the requirements of real users? What are the strategies for moving forward? . It was remarked that Web search engines don't provide CLIR because of the poor quality of general-purpose commercial machine translation, rendering the translated pages inaccurate. Perhaps summarization followed by more accurate translation would improve prospects for further use. It was observed that CLIR is a means to an end -- access to information regardless of language in which it is presented. Thus, concentration on end-use areas such as Cross Language Question Answering, Cross-Language Filtering, image search and Cross Language Summarization might be more fruitful Should the field evolve in this direction from CLIR to Multi-Lingual Information Access?

All workshop papers are available at: http://ucdata.berkeley.edu/sigir-2002/

Special Issue of IP&M on Cross-Language Information Retrieval: During SIGIR-2002 a meeting of the Board of Information Processing and Management approved a Special Issue on Cross-Language Information Retrieval, to be edited by the organizers of this workshop. The target for the issue is to be a reference issue for CLIR; its provisional schedule is: March 2003: Deadline for paper submission; June 2003: Notification of acceptance; September 2003: Publish.

We would like to thank our program committee members (listed below) who reviewed under deadline conditions and provided many helpful suggestions for improving the position papers:

James Allan	**(USA)**	**Jussi Karlgren**	**(Sweden)**
Martin Braschler	**(Switzerland)**	**Wessel Kraaij**	**(Netherlands)**
Hsin-Hsi Chen	**(Taiwan)**	**Chin-Yew Lin**	**(USA)**
Kuang-hua Chen	**(Taiwan)**	**Paul McNamee**	**(USA)**
Bruce Croft	**(USA)**	**Jian-Yun Nie**	**(Canada)**
Julio Gonzalo	**(Spain)**	**Doug Oard**	**(USA)**
Djoerd Hiemstra	**(Netherlands)**	**Ari Pirkola**	**(Finland)**
Gareth Jones	**(United Kingdom)**	**Mark Sanderson**	**(United Kingdom)**