Cross-Language Information Retrieval: the Way Ahead

Fredric C. Gey[1], Noriko Kando[2], Carol Peters[3]

[1] University of California, Berkeley, USA
[2] National Institute of Informatics, Tokyo, Japan,
[3] Italian National Research Council, Pisa, Italy,

Abstract

This introductory paper covers not only the research content of the articles in this special issue of IP&M but attempts to characterize the state-of-the-art in the Cross-Language Information Retrieval (CLIR) domain. We present our view of some major directions for CLIR research in the future. In particular, we find that insufficient attention has been given to the Web as a resource for multilingual research, and to languages which are spoken by hundreds of millions of people in the world but have been mainly neglected by the CLIR research community. In addition, we find that most CLIR evaluation has focussed narrowly on the news genre to the exclusion of other important genres such as scientific and technical literature. The paper concludes by describing an ambitious five-year research plan proposed by James Mayfield and Paul McNamee.

Keywords: cross-language information retrieval, system evaluation campaigns, research roadmap

1. Introduction

Cross-Language Information Retrieval (CLIR) has been recognized as an independent research sub-field for more than a decade now. The field has sparked three major evaluation efforts: a Cross-Language Track at TREC (Text Retrieval Conference) from 1997 to 2002, the Cross-Language Evaluation Forum (CLEF) – a spin-off from TREC - covering many European languages, and the NTCIR Asian Language Evaluation (covering Chinese, Japanese and Korean)[1]. These efforts have had significant impact, providing CLIR researchers and system developers with infrastructures for system testing and tuning, and with the opportunity to discuss and compare ideas and approaches. The result has been considerable progress in system design and development and the building up of a consolidated community of researchers around this topic.

A little more than two years ago, we decided that it was time to review and assess the progress that has been made so far and discuss what research and development remains to be done to make CLIR a practical rather than a research-oriented enterprise. We thus organized a workshop at SIGIR 2002 with the goal of developing a roadmap of research still to be undertaken. 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), with position papers suggesting the directions that research should take in the next half decade (Gey et al. 2002).

One of the results of the workshop was the decision by the editorial board of Information Processing and Management to produce a reference issue on Cross-Language Information Retrieval. The aim as stated in the Call for Papers was to present a landmark set of research papers which present the most significant research in the field on different aspects of multilingual information

[1] The authors of this paper have each been heavily involved in the coordination of these evaluation campaigns. See Gey and Oard (2001), Oard and Gey (2002), Kando (2003), Braschler and Peters (2004).

access and cross-language information retrieval. We received twenty-one submissions in response to this Call covering a wide variety of those areas and arguments that impact on the multilingual information access domain. Seven of these papers, highly representative of the current state-of-the-art, were selected for publication in this special issue.

The rest of this paper is organised as follows. Section 2 present the main areas of discussion and questions addressed at the Workshop, whereas Section 3 focuses specifically on the role that has been played by evaluation campaigns in influencing the direction that CLIR research has taken. In Section 4, we provide an overview of the papers in this volume while Section 5 make proposals for future research directions aimed at the development of fully functional, user friendly multilingual and multimodal retrieval systems. The paper concludes with the ambitious five-year research plan proposed at the Workshop by James Mayfield and Paul McNamee.

2. CLIR Workshop at SIGIR 2002

The workshop was organized in six thematic sessions: *Approaches to CLIR* described various techniques which 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 discussed CLIR for image and speech retrieval across languages. *User Studies/Interactive* presented papers on the role of user interaction in CLIR. A session on *Evaluation* described the Cross-Language Information Retrieval evaluations underway in Europe and Japan (CLEF and NTCIR, respectively) 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.

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, while the many local languages and dialects of Africa have been almost totally ignored.
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 60 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 offering CLIR? –** 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 when applied to the pages found.

Although there was a considerable exchange of ideas on these points, no exhaustive answers were found. The main discussion revolved around the third challenge: why is there still little take-up of the results of the R&D activity by the relevant application communities? A number of questions were asked: 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? In the final section of this paper we will address some of these issues.

3. CLIR Evaluation Campaigns

At the workshop, there was general agreement on the importance of the role that can be played by evaluation campaigns in promoting research in system development and in influencing the directions that future research can take. There is a duality between research and evaluation. Good research is validated by evaluation and good evaluation environments stimulate further research. Modern information retrieval evaluation began with the first TREC conference (Harman, 1993) and has continued within its subsequent eleven other conferences (Harman, 2003). TREC introduced a number of innovative evaluation ideas and approaches, including results pooling, known-item searching, reciprocal rank evaluation (for evaluation of factoid question answering), evaluation of interactive retrieval, and scene boundary detection for video. The success of TREC as an objective forum for evaluation led to the formation of the NTCIR series in Japan (Kando, 2003) and the CLEF campaigns in Europe (Braschler and Peters, 2004)[2]. An evaluation environment consists of a set of topics which describe information needs and a collection of documents which are to be searched to identify those documents which satisfy the information needs. The ground truth of evaluation is a set of "relevant" documents for each information need which has been identified by a human judge. Evaluation is then done for each ranking of documents with respect to the topic by the usual computation of recall, precision and other measures. For cross-language information retrieval evaluation, another factor is introduced – how well cross-language IR performs with respect to monolingual information retrieval on the same document collection.

3.1 The Cross-Language Track at TREC

Originally designed for (and still mainly focused upon) the English language, TREC expanded into other languages with the implementation of the first Spanish foreign language track in TREC-3. In TREC-3, retrieval of 25 topics against a Mexican newspaper corpus was tested by four groups. Spanish language retrieval was evaluated in TREC-3, TREC-4 (another 25 topics for the same Mexican corpus), and TREC-5 (where a European Spanish corpus was used). The TREC-3/4 Spanish collections were used by Ballesteros and Croft in their widely cited paper "Statistical Methods for Cross-Language IR" (Ballesteros and Croft, 1998). In TREC-5, a Chinese language track was introduced using both newspaper (People's Daily) and newswire (Xinhua) sources from People's Republic of China, and 25 Chinese topics with an English translation supplied. The TREC-5 corpus was represented with the GB character set of simplified Chinese. The Chinese monolingual experiments on this collection in TREC-5 and TREC-6 sparked research into the application of Chinese text segmentation to information retrieval using dictionary-based methods and statistical techniques, and simpler overlapping character-bigram segmentation methods were also found to be effective. TREC-6, TREC-7 and TREC-8 introduced the first cross-language

[2] For information on the current activities of these two initiatives see: http://research.nii.ac.jp/ntcir/ and http://www.clef-campaign.org

tracks, which focused upon European languages - first English, French and German, and later Italian (Harman et al., 2001). Following the TREC-8 conference, the venue for European-language retrieval evaluation moved to Europe with the creation of the Cross-Language Evaluation Forum, and the first CLEF workshop was held in Lisbon in September 2000. For TREC-9, the CLIR task used Chinese documents from Hong Kong. In distinction from the earlier TREC-5/6 Chinese corpus, these sources were written in the traditional Chinese character set and encoded in BIG5. Following TREC-9 the evaluation of English-Chinese retrieval moved to the NTCIR Evaluation which is coordinated by the National Institute of Informatics in Japan.

For the TREC-2001 and TREC-2002 conferences, the cross-language task was a bilingual retrieval of English topics against Arabic document collections (Gey and Oard 2002, Oard and Gey 2003). The paper by Xu and Weischedel in this issue demonstrates their experimental research utilizing the evaluation resources in Chinese and Arabic developed by the TREC tracks between 2000 and 2002.

3.2 Cross-Language System Evaluation at NTCIR

NTCIR is a series of evaluation workshops, organised at 18 month intervals and designed to enhance research in information access technologies, including information retrieval, text summarization, question answering and text mining. The first NTCIR was held in 1997 and the fourth series is now under way. In Asian countries, CLIR between English and native-languages can be critical for international information transfer, at least in an initial stage, and CLIR between languages with completely different structures and origins, such as English and Chinese, or English and Japanese, is a challenging task. There is thus great demand for efficient CLIR technology. For this reason, CLIR has been one of the central interests of NTCIR from the beginning, and has attracted many international participants. In addition, over this period, personal contacts between the populations of East Asia have dramatically increased and the countries in this region are gradually becoming part of a “multilingual multi-cultural society”, in which many people can understand more than one language to some extent in ordinary life, in business and entertainment, but are still not sufficiently fluent to formulate fully expressive queries. NTCIR thus started with Japanese and English bilingual CLIR and has gradually increased the number of document languages.

In NTCIR-1 (Kando and Nozue, 1999) and NTCIR-2 (Kando, 2001), tests were conducted on English-Japanese scientific abstracts. More than half of the documents were English-Japanese paired, but in NTCIR-2 the correspondence between paired documents was not revealed to participants. Interesting characteristics of the document collections were the number of technical terms and the existence of a partially paired corpus with an associated list of bilingual keywords, easily found in non-English speaking countries. Transliteration of technical terms (Fujii and Ishikawa, 1999) was proposed and corpus-derived lexicons were used by several groups. Corpus-based CLIR proved effective in NTCIR-1 but not in NTCIR-2. However, MRD or MT-based approaches supplemented by corpus-derived lexicons improved search effectiveness. Pre- and post translation query expansion and disambiguation, and PRF were used with success. Several enhancements of Okapi were proposed. Some groups tested partial document translation approaches as well. NTCIR-2 also tested on Mandarin Chinese news articles from Taiwan, with traditional Chinese characters using BIG5 encoding (Chen and Chen 2001)[3].

NTCIR-3 and NTCIR-4 used multilingual CtJKE (traditional Chinese, Japanese, Korean and English) news articles published in East Asia with CtJKE topics, and Japanese patents with English-

[3] The language used differs from the Chinese collections used in the TREC 5-6 Chinese Tracks (Simplified Chinese) and in TREC-9 CLIR (Cantonese dialects). Traditional and simplified Chinese (Ct and Cs) differ not only in character encodings but also in vocabulary.

Japanese translation-equivalent paired abstracts with CtCsJKE topics. All language collections are derived from multiple source newspapers and, in NTCIR-4, the size and publication years were well balanced. This change in document genre was decided in response to social needs, i.e., the recent increased interest in the social and cultural behaviour of neighbouring countries in East Asia reported above, and the growing importance of technological information transfer in the business and industrial sectors. Consequently, the handling of named entities[4], and cultural or local/domestic topics were typical targets for intensive investigation on the news document collections, whereas, the use of technical terms, very long documents, large-scale exactly translated paired corpus, etc., were issues for experiments on the patent collection.

At NTCIR-3, many groups tested monolingual retrieval on every language in the multilingual CJKE collections, and compared various language-dependent techniques including segmentation. Only seven groups submitted multilingual CLIR runs and many participants focused on using English topics to search Asian language documents for bilingual CLIR. This was partly because of the availability of translation resources that included English as one of the languages. Different groups experimented with query translation using MRDs or MT with PRF, translation disambiguation, corpus-based translation, or merging of ranked lists for retrieval from a collection in multiple languages. He and Gao (2003) proposed the "decaying co-occurrence model" for translation disambiguation. Chen and Gey (2003) adopted a corpus-based translation approach using web search engines. They also found that a hybrid of query translation and cognate matching between Chinese and Japanese through encoding conversion worked well. Lin and Chen (2003) explored a unique MultiLingual IR (MLIR) strategy for merging ranked lists based on translation difficulties. Sakai et al. (2003) intensively investigated variations of PRF, and Murata et al. (2003) proposed a weighting strategy considering keyword location. Moulinier et al. (2003) and Tomlinson (2003) compared various indexing techniques on CJK, and Luk et al. (2003) examined the comparative effectiveness of several indexing methods and retrieval models on Chinese. Such comparative studies have contributed greatly to clearer insights into segmentation and search mechanisms for CJK. For patent retrieval, follow-up studies were conducted in order to compare various conditions on the same implementation (Iwayama et al 2003).

Experiments using the NTCIR test collections are described by Fujita and by Seo et al. in this issue.

3.3 The Cross-Language Evaluation Forum

The Cross-Language Evaluation Forum (CLEF) has just completed its fourth year of independent activity. The move to Europe has made it possible to build on and extend the initial results achieved within TREC. The multilingual environment provided by Europe has made it easier to add new languages and has stimulated participation. It has also facilitated the organisation of CLEF which is organised on a distributed basis with native-language groups responsible for the creation of test data in each language. Since its beginning, there has been a concerted coordination of efforts and exchange of ideas between CLEF, NTCIR and the cross-language track at TREC; the aim has been to offer complementary CLIR evaluation activities to the R&D community. The first campaigns of CLEF have had as main goals:

- to accommodate as many European languages as possible;
- to provide facilities for monolingual system testing and tuning in European languages other than English, which was already well covered by TREC
- to stimulate systems to move from monolingual searching to the implementation of a full multilingual retrieval service

[4] A cognate-matching approach for named entities often does not work in the CJKE environment owing to the difference in characters used for each languages, and to the fact that the transliteration of names into English is not stable across newspapers or even in the same newspaper.

- to study the emerging needs of both system developers and system users in order to promote the introduction of new tasks.

The results have been encouraging in terms of numbers and of impact. Separate tracks to test monolingual, bilingual and multilingual systems were provided with the aim of allowing groups to work their way up gradually from mono- to multilingual retrieval. The test collections have continued to grow and the main corpus now consists of comparable news documents from the same time period in ten languages: Dutch, English, French, Finnish, German, Italian, Portuguese, Russian, Spanish, and Swedish. Participation of both academic and industrial groups, and especially of European groups, has increased rapidly. Additional tracks have been added to supplement the core tracks.

By creating a forum for the comparison of results using different approaches and technologies, CLEF appears to have had a real effect on CLIR research and system development. Over the years, we have seen considerable take-up of ideas and methodologies and sharing of resources among participating groups. By promoting the multilingual track as the main task, groups have been strongly encouraged to extend their systems in order to be able to handle a large number of languages and the various problems involved. All kinds of indexing methods have been tried: the merits of simple stemming have been compared with more complex morphological analysers, and different types of compound splitting have been tested on agglutinate languages. For multilingual retrieval, several alternatives for the handling of all the languages exist. They can be handled simultaneously, or they can be handled one at a time, through a succession of bilingual retrieval steps, and then subsequently merged into one, multilingual result. Most groups at CLEF have adopted the second approach. Many experiments aimed at identifying the best merging technique have been made, but so far no clear answer has emerged: the merging of various bilingual results to produce an optimum ranking for multilingual retrieval still remains an unsolved problem. In order to cross the language barrier between query and target collection, groups have experimented with both query and document translation, and with combinations of the two. Different kinds of translation resources have been employed: machine translation systems, electronic dictionaries, corpus-based techniques, and the use of pivot languages when no translation resources are available for direct translation between two languages. Groups have even experimented with methods that use no translation resource but match on character n-grams over languages. CLEF has actively encouraged groups to try out innovative ideas and some very interesting experiments have been presented (Peters, 2001; Peters et al, 2002; 2003; Peters, 2003).

Over the years the attention of the CLEF campaigns has gradually shifted from a focus on text retrieval systems and the measurement of document rankings towards the provision of a wider range of tasks. Increasing attention has been given to issues that interest the end users and their interaction with the system. For example, the ways in which a system can help the user when formulating a query or the ways in which the results of a search are presented are of great importance in CLIR where it is common to have users retrieving documents in languages with which they are not familiar. The paper by López-Ostenero et al. in this issue describes a tool to assist the user in query formulation and refinement and in foreign-language document selection that has been evaluated in the Interactive track at CLEF.

The 2004 campaign offers eight different evaluation tracks, and will include tasks to assess systems for multilingual question answering, for cross-language image retrieval, and for cross-language spoken document retrieval. The goal is to offer a comprehensive set of tasks covering all major aspects of multilingual, multimedia system performance with particular attention to the needs of the end-user.

3.4 The TIDES Surprise Language 2003

The major U.S. program which has funded cross-language information search (as well as other language technologies such as information extraction and summarization and machine translation) is the TIDES (Translingual Information Detection Extraction and Summarization) program of DARPA. The goal of TIDES is to dramatically improve the state of language technology to support the rapid response to new world crises. In the early days of the program the phrase "machine translation for a new language in a week" was coined. In 2003 the program developed a test scenario called the "TIDES Surprise Language Exercise" to evaluate the rapid response capability of the funded technology. The central idea is the announcement of the 'surprise' language on the first day of an evaluation and by the 30th day all technology is adapted to the new language and evaluated according to standard evaluation metrics. In Spring 2003 this was tested in the following ways: a dry-run exercise for 15 days in March to prepare and identify shortcomings in readiness for the actual exercise during the month of June. The dry-run surprise language was chosen as "Cebuano" spoken by about 15 million persons in the Philippine nation, the lingua franca of the southern Philippines. The June evaluation language was Hindi, spoken by 200 million persons in India. Each language presented special challenges: Cebuano because of the scarcity of electronic resources and Hindi because of the multiplicity of encodings of its scripted written language found abundantly on the web. In neither language was a body of aligned parallel text available outside of translations of the Bible. Printed bilingual dictionaries in both languages were scanned and made available and, for Hindi, an innovative web-based translation utility was set up to allow for online translation of Hindi news stories. By the end of the exercise a great deal had been learned and both translation resources developed, and evaluations had been performed on the fundamental language technologies involved in the TIDES program. More information may be obtained by reading the Special Issues on the Surprise Language Exercise of ACM Transactions on Asian Language Information Processing (Oard 2003).

4. CLIR Research in This Issue

The papers in this special issue of IP&M provide a cross-section of many of the important issues currently being investigated by the CLIR research community, both within and externally to the evaluation programs described in the previous section.

The first paper by Kazuaki Kishida provides a survey of the principal technical issues in cross-language information retrieval, including cognate matching, translation types (query, document, interlingual), dictionary-based mapping, disambiguation of multiple translations, machine translation, phrasal translation, parallel and comparable corpus-based methods for probabilistic translation, mining web resources for translation, merging issues for retrieval against multilingual corpora, use of pivot languages for indirect translation, and language-specific issues such as tokenization and segmentation for Asian languages, stopword lists, stemming, decompounding, part-of-speech tagging, etc.

When discussing query translation versus document translation, the paper notes the principal disadvantage of query translation to be short queries with a few disconnected words without sufficient context for disambiguation, while the principal disadvantage of document translation (where context is not a problem) is the overwhelming cost in terms of computing resources needed to translate large document collections. In the area of dictionary-based mapping from the query language terms to the document language terms, methods of disambiguation using query word pair context, part-of-speech mapping, and reverse translation are discussed. For parallel/comparable corpora, the paper describes algorithms for probabilistic matching between term pairs using maximum likelihood and other estimation techniques as well as statistical machine translation models. Interlingual techniques such as latent semantic indexing also depend upon parallel corpora alignment. The paper's examination of merging recognizes that the multilingual merging problem is, in principle, equivalent to the problem of distributed collection retrieval and that many of the techniques developed there apply directly to CLIR.

The remaining papers describe individual experiments, designed to test different aspects of the CLIR paradigm, using European, Asian and Arabic target language collections. Issues investigated include system architecture, indexing techniques, language modelling, translation resource acquisition and employment, pre- and post translation query expansion, target query term disambiguation and user-system interaction. Techniques proposed have been assessed using TREC, NTCIR and CLEF test data.

4.1 Improving Dictionary-Based CLIR

Most of the papers in this special issue describe experiments that include the application of dictionary-based translation techniques when mapping correspondences between languages. The paper by Levow, Oard and Resnik identifies the key issues raised when using the simplest form of dictionary-based CLIR - machine-readable bilingual term lists - and proposes a unified framework for term selection and translation. These authors also study the effects of such techniques on the retrieval effectiveness for languages with different characteristics, using queries in English on document collections in French, mandarin Chinese, German and Arabic under diverse experimental conditions. Points covered include appropriate methods for indexing and term extraction according to the characteristics of the language, query and document term expansion, the merits of the structured query approach (Pirkola, 1998), and a general methodology which can be used to enhance dictionary coverage.

Both the papers by Xu and Weischedel and by Larkey and Connell also utilize and test the effectiveness of bilingual machine-readable dictionaries when applied to the translation part of cross-language information retrieval. Of course, as also described by Levow et al, translation ambiguity is a major problem when using dictionary-based techniques. Bilingual dictionaries can often provide multiple translations for the same source language word or phrase. This can occur because of polysemy (e.g. the English word *bank* as a financial institution or the side of a river) or because the same source word may be expressed in multiple ways in the target language. Since both of these two papers test probabilistic retrieval models, they make the assumption of uniform probability among the translation alternatives. In reality not all translations are equally likely and much work has been done in recent years on improving the choice of best possible translation among alternative choices. This has often been done by looking for translation alternatives among pairs of adjacent (non-trivial) source language query words within a window (sentence, paragraph, fixed size segment) of target language documents being retrieved.

The paper "Improving Query Translation in English-Korean Cross-Language Information Retrieval" by Seo, Kim, Rim, and Myaeng extends this research in a new direction by optimizing over all possible combinations of translations of source query terms. This process can be compute-intensive (a query of 10 words can generate 10^{20} possible combinations of terms) and hence these authors introduce heuristics to reduce the computational overhead of finding the best possible translation alternative. While their experiments are with English to Korean bilingual retrieval, the methodology is general and could be applied to any language pair where translation is being carried out via machine-readable dictionaries.

One of the problems noted by Levow et al when using dictionaries is asymmetry in term selection for translation and matching and consequent variations in performance. This issue is studied in far more depth with comparative evaluation experiments in the paper by Fujita for term translation between Japanese and English and vice versa on comparable collections. He attempts to explain the asymmetry from two aspects: query translation quality, and discrepancies between information needs with respect to the actual query formulated, and with respect to the contents of the target collection. Fujita describes the use of a bilingual Japanese-English dictionary together with a (partially) parallel Japanese-English collection for pre-translation feedback with positive results. He suggests that a challenging direction for future research would also be to test the use of parallel collections in situations where pivot languages and transitive translation methods are involved.

4.2 Translation Resources Beyond Human-constructed Dictionaries

The papers by Xu and Weischedel and by Larkey and Connell primarily concentrate on evaluation of resources used in the translation aspect of cross-language information retrieval. Both papers demonstrate that use of machine translation in CLIR has moved beyond application of commercial off-the-shelf translation systems to the utilization of state-of-the art statistical machine translation software and models. In both papers, bilingual lexicons are induced using the GIZA++ system against parallel corpora (Al-Onaizan et al, 1999), using IBM Model 1 (single word translation) in the case of Xu and Weischedel and IBM Model 4 (more sophisticated bigram dependency translation) by Larkey and Connell.

The paper by Xu and Weischedel utilizes a model of CLIR which incorporates translation into the retrieval process and compares the combination of fixed resources (i.e. a human produced bilingual dictionary) with flexible resources (lexicons derived from bilingual corpora). Their experiments test retrieval effectiveness as a function of dictionary size, showing that beyond a certain threshold, dictionary-based retrieval plateaus, and dictionary-based retrieval improves when supplemented by statistical lexicons. Bilingual retrieval experiments are carried out for English queries against Chinese, Arabic and Spanish documents, using collections and relevance judgments from the various TREC evaluations of these languages.

The paper by Larkey and Connell compares a traditional IR approach (as exemplified by the U. Massachusetts INQUERY system) which allows for structured queries to incorporate alternative dictionary translations as synonyms to advanced language model approaches (incorporating relevance models) applied to cross-language information retrieval. Their experiments are carried out again for English queries against the Arabic or Spanish document collections of TREC. While Xu and Weischedel carry out their experiments without pseudo-relevance (blind) feedback (wishing to have a clean comparison of resource size), Larkey and Connell test their experiments with pseudo-relevance feedback incorporated as an integral part of the process (indeed incorporated into the language model part of the experimentation). The Larkey/Connell experiments exhibit a result where cross-language retrieval performance (at least where English is the source language) exceeds monolingual performance.[5] They describe how the process of blind feedback can produce this result by raising the performance of a single query from abysmal to excellent, thus impacting overall performance.

4.3 Interaction with the User

One of the criticisms frequently made of research in the CLIR domain is that too much attention has been given to questions of retrieval functionality and effectiveness with little regard to the real needs of the end user. Many users of CLIR systems are looking for and retrieving information in languages in which they have little or no competence. This means that they may well need guidance from the system both in formulating or refining their queries and in interpreting the results. Other CLIR system users may be perfectly capable of understanding results in a number of languages but

[5] CLIR performance exceeding monolingual, while sometimes observed for English→X bilingual retrieval, is certainly not the norm for Asian language retrieval at NTCIR, nor for non-English topic languages at CLEF. It is misleading to conclude from these few narrow tests that the CLIR problem has been solved.

want to be able to query a number of information sources simultaneously, using a single query. Different user groups demonstrate different behaviours and system design must take this into consideration.

The recognition of these needs has given rise recently to a new line of investigation aimed at studying the most effective means for user-system interaction. Interesting work in this area has been done by the Clarity project[6]. Clarity has studied CLIR for the so-called low-density languages, those with few translation resources. The project has investigated a number of techniques aimed at enabling users to better interact with the CLIR system by presenting and organising cross-language retrieval results in an efficient way (Petrelli et al, forthcoming). The paper by López-Ostenero et al in this volume continues this line of investigation, presenting an interactive system which is based on the belief that noun phrases are basic conceptual units and as such can be usefully exploited for both query formulation and refinement and for document selection. The paper claims that a cross-language summarization algorithm based on translations of the noun phrases in a document is much faster and less resource consuming than full machine translation, and provides the user with all the information needed to select relevant documents or choose appropriate translations for the query terms. The problem, as with all experimental work in this field, is that the user studies reported are limited and the results are assessed on the basis of a very small data sample. It is hard and time consuming work to set up an extensive user study and this is one of the reasons that has constrained research in this area so far. More studies of user behaviour and user needs in the CLIR context are needed in order to design and build systems that are not only efficient but respond to the users' expectations.

5. The Future of CLIR Research

As evidenced in the papers in this volume, CLIR research so far has mainly concentrated on text collections and on a limited set of languages. However, in the discussions at the SIGIR workshop, it was observed that CLIR was simply a means to an end – access to information regardless of the language or media in which it is presented - and the user is primarily interested in the end result. It was thus felt that there should be more attention given to end-user issues such as results presentation, multilingual question answering, cross-language filtering and summarization and also to the implications of multilingual searching in collections in multimedia. Furthermore, there was general agreement that future research in CLIR should focus more specifically on several additional areas:

- new languages, particularly lesser studied languages
- different genres and media
- the multilingual web
- fundamental models and unsolved problems

5.1 CLIR for all Languages

Thus far CLIR research and resource development has involved just a small fraction of the nearly 2,000 widely spoken languages in the world today. Indeed, only about ten of the top 25 most commonly used languages have been subjected to any kind of consistent CLIR experiments. Not surprisingly, these are also the languages that have the most economic and commercial influence, i.e. a number of European languages, certain East Asian languages and Arabic.

[6] http:// www.dcs.shef.ac.uk/research/groups/nlp/clarity/

Rank	Language	Principal Country	Population speaking
1	*CHINESE MANDARIN [CHN]	China	885,000,000
2	*SPANISH [SPN]	Spain	332,000,000
3	*ENGLISH [ENG]	United Kingdom	322,000,000
4	BENGALI [BNG]	Bangladesh	189,000,000
5	*HINDI [HND]	India	182,000,000
6	*PORTUGUESE [POR]	Portugal	170,000,000
7	*RUSSIAN [RUS]	Russia	170,000,000
8	*JAPANESE [JPN]	Japan	125,000,000
9	*GERMAN STANDARD [GER]	Germany	98,000,000
10	CHINESE WU [WUU]	China	77,175,000
11	JAVANESE [JAN]	Indonesia Java Bali	75,500,800
12	*KOREAN [KKN]	Korea South	75,000,000
13	*FRENCH [FRN]	France	72,000,000
14	VIETNAMESE [VIE]	Viet Nam	67,662,000
15	TELUGU [TCW]	India	66,350,000
16	CHINESE YUE [YUH]	China	66,000,000
17	MARATHI [MRT]	India	64,783,000
18	TAMIL [TCV]	India	63,075,000
19	TURKISH [TRK]	Turkey	59,000,000
20	URDU [URD]	Pakistan	58,000,000
21	CHINESE MIN NAN [CFR]	China	49,000,000
22	CHINESE JINYU [CJY]	China	45,000,000
23	GUJARATI [GJR]	India	44,000,000
24	POLISH [PQL]	Poland	44,000,000
25	ARABIC EGYPTIAN SPOKEN [ARZ]	Egypt	42,500,000

Table 1: World's Top 25 most widely spoken languages. *Languages for which consistent CLIR experiments have been performed are indicated by *.* [Source: Ethnologue language list, http://www.ethnologue.com]

In total, no more than fifteen of the world's languages have been subjected to extensive formal evaluation exercises and test corpus development, and most of these are European languages. Indeed, nine of the eleven official languages of the European Union are included in the CLEF test suites[7]. It could be felt that such concentration of attention gives these languages an unfair advantage. This issue also has far reaching social implications when we are talking about global access to information. The diversity of the world's languages and cultures gives rise to an enormous wealth of knowledge and ideas. CLIR research should contribute to the ensuring the survival of endangered languages not to giving unfair advantages to a chosen few.

We are still a long way from the creation of instruments and methodologies that will make it possible to overcome all language barriers, although the TIDES surprise language exercises described above has been instrumental for the investigation and understanding of many of the problems and difficulties involved (Oard 2003). Particular obstacles are represented by:

- font and representation anarchy (Strassel et al 2003)
- the lack of machine-readable resources and NLP tools

While it is feasible that the first of these problems will be overcome with a gradual acceptance and adoption of common globally recognized encoding standards for the representation of information

[7] The official languages of the European Union are currently Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish. Following enlargement it is likely that a further eight official languages may be added : Estonian, Latvian, Lithuanian, Polish, Czech, Slovak, Hungarian and Slovene.

in digital form, the second is more difficult to solve. It will probably be necessary to investigate two directions: (i) proposals for quick and inexpensive ways to create resources and tools for a “new” languages, as in the TIDES surprise language exercise described above, (ii) language-independent methodologies which can be adopted for all languages, or more likely for groups of languages with the same characteristics.

5.2 Different Genres and Media

Much of the research work and almost all CLIR evaluations have focused upon news stories as the main genre of study. However, news media have characteristics which may not hold true for other genres: wide use of proper nouns (names and places), of which the names have a brief lifespan in a temporally tagged corpus, association of date stamps, particular style of writing and a rapid evolution of general-purpose vocabulary. Certain features may facilitate access and retrieval, others may hinder it. By contrast, scientific and technical terminology is difficult to locate in standard machine translation resources, but is, when located, relatively stable (except in rapidly developing research areas such as e-commerce). Although there has been some important CLIR research in domain-specific areas, in particular in the medical and legal domains[8], much more evaluation work is needed in order to understand which approaches are the most successful. In fact, real-world running applications in domain-specific sectors, even those that contain collections in multiple languages, are reluctant to adopt any serious cross-language functionality. Most of those that do adopt strategies that use some kind of controlled vocabulary. This is confirmed by a recent study of digital library projects under the Fifth Framework programme of the European Commission which revealed that while 14 contained collections in multiple languages, only six had implemented cross-language search mechanisms. And five of these used a multilingual controlled vocabulary or thesaurus. But why did they chose this solution? And was it really the most appropriate?

While it is true that both NTCIR and CLEF have worked to some extent with domain-specific data (scientific abstracts and also patents at NTCIR), the collections have only covered a few languages (Japanese, German, and French) while some of the best scientific and mathematical literature is also being published in Chinese and Russian and important technical literature exists in very many languages. More comparative studies are needed to understand which approaches and techniques are most suitable according to the genre to be handled.

Similarly, it is time that research and evaluation work shifts its focus to collections in diverse media rather than just text, and from assessing system performance only in terms of a list of ranked documents to other equally important issues that affect user satisfaction such as assistance in query formulation and results presentation in appropriate forms, according to the user needs and his/her language competence.

Differently from the IR scenario, a Question Answering (QA) system processes questions formulated into natural language (instead of keyword-based queries) and retrieves answers (instead of documents). QA is a multi-faceted problem requiring contributions from information retrieval, natural language processing and artificial intelligence. The components of a good QA system thus differ from that of a traditional CLIR system and need to be studied independently. A pilot track for multilingual question answering system evaluation was introduced in CLEF for the first time in 2003 in order to encourage investigation of the issues involved (Magnini et al., 2004).

The current expansion in collections of digital documents in various media and languages means that there is a growing need for systems able to automatically access the information contained in

[8] See, for instance, the studies that have exploited versions of the MeSH thesaurus and the Unified Medical Language System in different languages, such as in the MuchMore project (Vinter et al 2003) or work on retrieval of legal documents in multiple languages (Sheridan and Schäuble, 1997)

such documents. However, almost all work in the multimedia area has been on monolingual (generally English) collections. Two years ago CLEF decided to promote research on multilingual multimedia IR systems and two experimental tracks were introduced. So far the cross-language speech track has aimed at evaluating CLIR systems on noisy automatic transcripts of spoken documents comparing performance against a monolingual baseline. First results show the importance of the translation resources used and also that different indexing units can be give better performance depending on whether bilingual or monolingual retrieval is involved (Federico and Jones, forthcoming). Cross-language image retrieval is a very new area for CLIR research. Depending on the collection and on the processing tools available, the retrieval can be purely caption-based (and thus becomes a particular type of text retrieval) or can be based on a combination of language-dependent (text-based) and language-independent (image-based) features (Clough and Sanderson, forthcoming). In our opinion these two sectors represent crucial areas for research in the future and investments are needed in order to be able to create the necessary test collections for serious comparative studies.

5.3 The Multilingual Web

Probably, the first application that comes to mind for most people when they think about the potential of CLIR is searching on the WorldWideWeb. It is however notable that cross-language search and retrieval is not a functionality offered by the most widely used web search engines. We should examine the reasons for this. It was remarked at the workshop that Web search engines do not provide CLIR services because of the poor quality of general-purpose commercial machine translation. But MT is not the only way to go for CLIR. It seems evident that the CLIR R&D community should now be investigating what would be the most successful technology, or combination of technologies, for CLIR in the dynamic context represented by the Web.

In Table 2, we note the following:

- the web is predominated by languages of the developed countries. Further, the top 16 languages account for 90.3% of all web usage and the bottom 23 languages account for 0.93% of all web pages or less than 1 percent;
- while English remains the predominant language, its share has been declining over time (a 1999 study by Excite Corp had English web pages as 72% of the web at that time).

Hence to prepare a web corpus for evaluation purposes (if we presume that CLIR researchers will not have the resources to store the entire WWW) we will need to prepare a stratified sample, one which taps the entire collection for less represented languages and randomly samples some fraction of the more represented languages (say those representing at lease 1 percent of web urls). What the sampling fraction should be for the 'between' languages will require careful thought.

For any representative sample of the web, we will also encounter font and character set representation problems. The panacea of Unicode representation meets reality with the web capability of dynamic font downloading, which enables any web site to choose its own character set and font encoding. Only in Western Europe are the ISO standards widely used and accepted.

Language	Number-of-Pages	Percentage
English	1,690,901,291	57.37%
German	256,997,640	8.72%
French	120,860,523	4.10%
Russian	98,422,529	3.34%
Japanese	95,372,822	3.24%
Chinese(simplified)	93,228,783	3.16%
Spanish	89,429,894	3.03%
Italian	78,104,597	2.65%
Korean	66,390,095	2.25%
Dutch	64,020,603	2.17%
Portuguese	34,193,853	1.16%
Czech	31,429,288	1.07%
Swedish	28,700,267	0.97%
Polish	27,864,012	0.95%
Danish	27,622,742	0.94%
Chinese(traditional)	21,421,052	0.73%
Catalan	19,245,884	0.65%
Norwegian	14,728,161	0.50%
Hungarian	14,504,776	0.49%
Finnish	11,856,905	0.40%
Slovak	9,236,475	0.31%
Turkish	7,587,437	0.25%
Greek	5,261,482	0.18%
Hebrew	4,563,942	0.16%
Romanian	3,993,211	0.14%
Arabic	3,882,050	0.13%
Thai	3,547,806	0.12%
Others †	23,959,095	0.93%
Total†	2,947,327,215	100.00%

Table 2: Language Distribution of 2.3 Billion Web Pages in July 2003 [Source: all-the-web.com for July 2003, prepared by Takagi and Gey] **†** 22 other languages (in order of count): Croatian, Estonian, Bulgarian, Slovenian, Byelorussian, Icelandic, Lithuanian, Indonesian, Ukrainian, Latvian, Galician, Vietnamese, Malay, Afrikaans,. Basque, Latin, Faeroese, Albanian, Frisian, Welsh, Serbian and Swahili) each have less than 0.1% and cumulatively less than one percent of the total web pages.

Finally, there are intellectual property problems associated with creating such a corpus, even for research and evaluation. In the United States a published work (such as a web page) is automatically copyrighted under the law, even if the copyright is not registered. However, works of the U.S. federal government are considered to be publicly financed and hence their content in the public domain (not copyrightable). For this reason, the TREC web track uses urls and pages taken from the .gov domain in order to prepare a web corpus for research and evaluation purpose. Whether this unrestricted usage by the public will hold for other countries is a matter of some question. CLEF has now begun discussion on the creation of a multilingual comparable web corpus, by following the TREC example and spidering government sites of a number of European countries for a set of specific sectors such as health, social security, education, etc. but first will need to investigate eventual copyright issues.

5.4 Fundamental CLIR Models and Outstanding Problems

Both Jian-Yun Nie (2002) and Mayfield and McNamee (2002) have suggested that drawbacks of current research in CLIR have stemmed from a haphazard approach to the problem, including the separation of the translation process from the retrieval process, and the difficulty of addressing results merging from monolingual retrieval results for multiple languages. Nie suggested that in a unified approach to CLIR, the translation process and the retrieval process would incorporate a) the word distribution knowledge in both source and target languages and b) the uncertainty of translation (Nie, 2002). The paper by Xu and Weischedel in this issue presents a start at a unified model for CLIR which incorporates the translation uncertainty into the retrieval process, but not the word distribution in the target collection, nor the merging problems (their paper deals with bilingual search). Incorporation of word distribution has been suggested as Pirkola's method, which has been used by a number of researchers (Pirkola, 1998). The problem of merging is a special case of the problem of retrieval from distributed collections; an example of experiments made on different merging methodologies for the CLEF collections can be found in Chen & Gey (2004).

6 A Five Year Plan for CLIR Research

At the workshop, Mayfield and McNamee (2002) proposed a 5 year plan for CLIR research which covers the following areas: tools, standards, resources (both aligned corpora and bilingual dictionaries as well as multilingual ontologies), multiple modalities and media and web corpora:

Period	Resources	Evaluation
Year 1	Development of standards and tools for translation resources, both bilingual and multilingual	Isolation of the resources and retrieval methodology in evaluation conferences
Year 2	Release of large, comparable and aligned corpora including several genres (10GB+ per language pair)	Evaluation of name translation/transliteration and spelling correction in 5-10 languages
Year 3	Evaluation of systems that build bilingual dictionaries in several languages with 100,000 or more entries using publicly available Web sources.	Evaluation of document selection by users with no ability to read a foreign language
Year 4	Tools and methods that simplify building speech models from spoken corpora	Evaluation of multilingual retrieval (ad hoc or classification in 15 (or more) languages, including Asian, European, Indic and Semitic languages
Year 5	A global WordNet available in 15 languages with a kernel of 100,000 synsets in each of the languages	Evaluation workshop in cross-language speech retrieval in 4 or more languages that attracts at least 10 participating groups.

This ambitious plan seems to be a good strawman proposal for the next generation of cross-language research. It will require both commitment and funding to become a reality.

References

Al-Onaizan, Y., Curin, J., Jahr, J., Knight, K., Lafferty, J., Melamed, D., Och, F-J., Purdy, D., Smith, N. A. & Yarowski, D. (1999). *Statistical Machine Translation.* Final Report. Johns Hopkins University Text, Speech, and Dialog Workshop 1999, http://www.clsp.jhu.edu/ws99/final/Stat_Machine_Translation.pdf .

Braschler, M. & Peters, C. (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval*, 7(1-2), 7-31.

Ballesteros, L. & Croft, W.B. (1998). Statistical Methods for Cross-Language Information Retrieval. In Grefenstette, G. (Ed.), *Cross-Language Information Retrieval*, (pp. 23-40). Kluwer Academic Publishers.

Chen, A. & Gey, F. C. (2003). Experiments on cross-language and patent retrieval at NTCIR-3 workshop. In *NTCIR Workshop 3: Proc. of Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo, http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings3/NTCIR3-CLIR-ChenA.pdf

Chen, A. & Gey, F. (2004). Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding, *Information Retrieval*, 7(1-2), 147-180.

Chen, K. H. & Chen, H. H. (2001). The Chinese Text Retrieval Tasks of NTCIR Workshop 2. In *NTCIR Workshop 2: Proc. of Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, Japan, May 2000-March 2001. NII, Tokyo, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/overview-kando.pdf

Clough, P.D. & Sanderson, M. (2004). The CLEF 2003 Cross-Language Image Retrieval Track. In Peters, C., Braschler, M., Gonzalo, J & Kluck, M. (Eds.), *Proceedings of Fourth Workshop of the Cross-Language Evaluation Forum, CLEF 2003. Lecture Notes in Computer Science*, Springer, forthcoming.

Federico, M. & Jones, G. (2004). The CLEF 2003 Cross-Language Spoken Document Retrieval Track. In Peters, C., Braschler, M., Gonzalo, J & Kluck, M. (Eds.), *Proceedings of Fourth Workshop of the Cross-Language Evaluation Forum, CLEF 2003. Lecture Notes in Computer Science*, Springer, forthcoming.

Fujii, A. & Ishikawa, T. (1999). Cross-Language Information Retrieval at ULIS. In *NTCIR Workshop 1: Proc. of First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*. Tokyo, Aug.30–Sept.1,1999, Tokyo, NACSIS, http://research.nii.ac.jp/ntcir/workshop/ OnlineProceedings/ 031-IE-fujii.pdf

Gey, F.C., Kando, N. & Peters, C. (Eds.) (2002). Cross-Language Information Retrieval: *A Research Roadmap. Proceedings of a Workshop at SIGIR-2002*, Tampere Finland August 15, 2002, http://ucdata.berkeley.edu/sigir-2002/sigir2002CLIR.pdf

Gey, F. & Oard, D (2002). The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries, *The Tenth Text Retrieval Conference, TREC 2001*, (pp.16-25), NIST Special Publication 500-250, May 2002.

Harman, D. (1993). Overview of the first Text REtrieval Conference (TREC-1). In *Proceedings of the First Text REtrieval Conference (TREC-1*), (pp.1–20). NIST Special Publication 500-207.

Harman, D. (2003). The Development and Evolution of TREC and DUC. In *NTCIR Workshop 3: Proc. of the Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo, ISBN:4-924600-77-6, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-INV-HarmanD.pdf

Harman, D., Braschler, M., Hess, M., Kluck, M., Peters, C., Schäuble, P. & Sheridan, P. (2001). CLIR Evaluation at TREC. In Peters, C (Ed.) *op cit.* (pp. 7-23). Springer.

He, H. & Gao, J. (2003). NTCIR-3 CLIR experiments at MSRA. In *NTCIR Workshop 3: Proc. of Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo, http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings3/NTCIR3-CLIR-HeH.pdf

Iwayama, M., Fujii, A., Kando, N. & Marukawa, Y. (2003). An empirical study on retrieval models for different document genres: Patents and newspaper articles. In *Proc. of 26th Annual International ACM SIGIR Conference*, pp. 251–258.

Kando, N. (2001). Overview of the second NTCIR Workshop. In *NTCIR Workshop 2: Proc. of Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, Japan, May 2000-March 2001. NII, Tokyo, http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings2/overview-kando.pdf

Kando, N. & Nozue, T. (Eds.) (1999). In NTCIR Workshop 1: Proc. of First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition. Tokyo, Aug. 30 – Sept. 1, 1999. ISBN: 4-924600-77-6, http://research.nii.ac.jp/ntcir/workshop/ OnlineProceedings/index.html

Kando, N. (2003). Introduction to the NTCIR. In *NTCIR Workshop 3: Proc. of Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization,* Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo, ISBN:4-924600-77-6, http://research.nii.ac.jp/ ntcir/workshop/OnlineProceedings3/NTCIR3-INT-KandoN.pdf

Lin, W. C. & Chen, H. H. (2003). Description of NTU at NTCIR3: Multilingual information retrieval. In *NTCIR Workshop 3: Proc. of Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization,* Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo, http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings3/NTCIR3-CLIR-LinW.pdf

Luk, R. W. P., Wong, K. F. & Kwok, K. L. (2003). Different retrieval models and hybrid term indexing. In *NTCIR Workshop 3: Proc. of Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization,* Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo, http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings3/NTCIR3-CLIR-LukR.pdf

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F. & de Rijke, M. (2004). The Multiple Language Question Answering Track at CLEF 2003. In Peters, C., Braschler, M., Gonzalo, J & Kluck, M. (Eds.), *Proceedings of Fourth Workshop of the Cross-Language Evaluation Forum, CLEF 2003. Lecture Notes in Computer Science,* Springer, forthcoming.

Mayfield, J & McNamee, P. (2002). Three Principles to Guide CLIR Research, in *A Research Roadmap. Proceedings of a Workshop at SIGIR-2002,* Tampere Finland August 15, 2002, http://ucdata.berkeley.edu/sigir-2002/sigir2002CLIR-18-mayfield.pdf

Moulinier, I., Molina-Salgado, H. & Jackson, P. (2003). Thomson Legal and Regulatory at NTCIR-3: Japanese, Chinese and English Retrieval Experiments. In *NTCIR Workshop 3: Proc. of Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization,* Tokyo Japan, Oct.2001–Oct.2002. NII, Tokyo, http://research.nii.ac.jp/ntcir/workshop/ OnlineProceedings3/ NTCIR3-CLIR-MoulinierI.pdf

Murata, M, Ma, Q, & Isahara, H. (2003). Applying multiple characteristics and techniques to obtain high levels of performance in information retrieval. In *NTCIR Workshop 3: Proc. of Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization,* Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo, http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings3/NTCIR3-CLIR-MurataM.pdf

Nie, J-Y (2002), Towards a Unified Approach to CLIR and Multilingual IR, in *A Research Roadmap. Proceedings of a Workshop at SIGIR-2002,* Tampere Finland August 15, 2002, http://ucdata.berkeley.edu/sigir-2002/sigir2002CLIR-04-nie.pdf

Oard, D. (2003). The Surprise Language Exercises, *ACM Transactions on Asian Language Information Processing,* 2 (3-4), 79-84.

Oard, D and F Gey (2003). The TREC-2002 Arabic-English CLIR Track, *The Eleventh Text Retrieval Conference, TREC 2002.* (pp 17-26) NIST Special Publication 500-251. May 2003. (available at http://trec.nist.gov) .

Peters, C. (Ed.) (2001). *Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal. Lecture Notes in Computer Science 2069,* Springer, 387p.

Peters, C., Braschler, M. Gonzalo, J & Kluck, M. (Eds.) (2002). *Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Lecture Notes in Computer Science* 2406, Springer, 600p.

Peters, C., Braschler, M. Gonzalo, J, Kluck, M. (Eds.) (2003). *Advances in Cross-Language Information Retrieval. Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Lecture Notes in Computer Science* 2785, Springer, 825p.

Peters, C.,. (Ed.) (2003). Results of the CLEF 2003 Cross-Language System Evaluation Campaign: Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway, Vols I and II., http://www.clef-campaign.org/.

Petrelli, D., Hansen, P., Beaulieu, M., Sanderson, M., Demetriou, G. & Herring, P. (forthcoming). Observing Users - Designing Clarity: A Case study on the user-centred design of a cross-language retrieval system. *Journal of the American Society for Information Science and Technology (JASIST)* in press.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based crosslanguage information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 55-63.

Sakai, T., Koyama, M., Suzuki, M. & Manabe, T.. (2003). Toshiba KIDS at NTCIR-3: Japanese and English-Japanese IR. In *NTCIR Workshop 3: Proc. of Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization,* Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-CLIR-SakaiT.pdf

Sheridan, P. & Schäuble, P. (1997). Cross-Language Information Retrieval in a Multilingual Legal Domain (1997) *Proceedings of ECDL-97, Lecture Notes in Computer Science 1324* (pp. 253-268). Springer.

Strassel, S, Maxwell, Cieri, C, (2003). Linguistic resource creation for research and technology development: A recent experiment, *ACM Transactions on Asian Language Information Processing* (2), 101 – 117.

Tomlinson, S. (2003). Asian language parsing evaluated by Hummingbird SearchServer^{TM} at NTCIR-3. In *NTCIR Workshop 3: Proc. of Third NTCIR Workshop on Research in Information Retrieval, Question Answering and Summarization*, Tokyo Japan, Oct. 2001–Oct. 2002. NII, Tokyo, http://research.nii.ac.jp/ntcir/ workshop/OnlineProceedings3/NTCIR3-CLIR-TomlinsonS.pdf

Vintar, S., Buitelaar, P., Volk, M. (2003). Semantic Relations in Concept-Based Cross-Language Medical Information Retrieval In *Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM*), Cavtat-Dubrovnik, Croatia.