From CLEF to TrebleCLEF: the Evolution of the Cross-Language Evaluation Forum

Nicola Ferro[1] and Carol Peters[2]

[1]Department of Information Engineering, University of Padua, Italy
ferro@dei.unipd.it

[2]ISTI-CNR, Area di Ricerca, Pisa, Italy
carol.peters@isti.cnr.it

Abstract

The Cross-Language Evaluation Forum (CLEF) has been running for nearly ten years now; the aim of this paper is to provide a critical assessment of the results achieved so far. In the first part of the paper, we provide a brief overview of the entire activity and summarise the main achievements; in the second part, we focus our attention on the Ad Hoc track with the aim of showing how the results of evaluation can be exploited to increase understanding of the many issues involved in multilingual retrieval system development. In the final part, we outline our main ideas for the future of CLEF.

Keywords: Multilingual Information Access, Cross-Language Information Retrieval, Evaluation, Scientific Data, Data Curation, Test Collections

1 Introduction

The *Cross-Language Evaluation Forum (CLEF)* has been running for almost a decade now; the tenth birthday will be celebrated at the CLEF 2009 workshop. When we launched this activity as a European initiative in early 2000, our declared objectives were the following: “to develop and maintain an infrastructure for the testing and evaluation of information retrieval systems operating on European languages, in both monolingual and cross-language contexts, and to create test-suites of reusable data that can be employed by system developers for benchmarking purposes” [9]. Although it is true to say that this basic idea remains at the core of our activity, over the years our range of interest and our interpretation of these initial objectives have both widened and deepened.

When we began cross language information retrieval had only just started to be recognised as an separate sub-discipline[1], there were very few research prototypes in existence and work was almost entirely concentrated on text retrieval systems running on at most two languages. Our aim via the organisation of annual evaluation campaigns has been to (i) extend the original scope of the activity to encompass truly multilingual multimodal systems covering many languages and diverse media, (ii) build a strong multidisciplinary research community around this area, (iii) create a sustainable technical framework which would not simply support but would also empower both R&D and evaluation activities.

The intention of this paper is to provide a panorama of the CLEF results so far and to show how we have created the foundations which now allow us to shift the focus of our activity, and enable us to concentrate not only on widening our coverage of the various building blocks involved in multilingual system development (tools, components, resources, lexicons) but also on the acquisition of a deeper understanding of the underlying issues. This change in direction has been helped by the launching in 2008 by the European Commission of the TrebleCLEF Coordination Action[2]

The paper is thus structured as follows. The next section will describe the expansion of CLEF from

[1] The first workshop on “Cross-Lingual Information Retrieval” was held at the Nineteenth Annual ACM-SIGIR Conference on Research and Development in Information retrieval in 1996; at this meeting there was considerable discussion aimed at establishing the scope of this area of research and defining the core terminology. Since then, stimulated by the rapid growth of the Internet and the increasing globalization of information, dedicated workshops have been held every year and aspects of the problem are now routinely discussed at conferences on digital libraries, information retrieval, machine translation, and computational linguistics.

[2] TrebleCLEF is a 7FP Coordination Action under the ICT programme; it began activity in January 2008. The Consortium is composed of five academic partners and two important centres: ISTI-CNR, Italy; University of Padua, Italy, University of Sheffield, UK; Zurich University of Applied Sciences, Switzerland; UNED, Spain; CELCT, Italy, ELDA, France. See http://www.trebleclef.eu/

2000 - 2008: the creation of the research community, the building of the technical infrastructure, and the growth of the tracks and the test collections. In the following section, we describe how the TrebleCLEF initiative has been set up in order to build on and expand the results achieved within CLEF. Section 4 takes the Ad Hoc track as a case study to show how we now intend to exploit the valuable resources and experimental collections made available by CLEF over the years in order to gain more insights about the effectiveness of various retrieval techniques with respect to different languages. In the final section, we outline our intentions for the future.

2 The CLEF Evaluation Campaigns

As has been described elsewhere, CLEF actually began life in 1997 as a track for *Cross Language Information Retrieval (CLIR)* within the *Text REtrieval Conference (TREC)* organized in the US by NIST and DARPA[3]. The aim was to provide researchers with an infrastructure for evaluation that would enable them to test their systems and compare the results achieved using different cross-language strategies [6]. However, after three years within TREC, it was decided that Europe was better suited for the coordination of an activity that focused on multilingual aspects of information retrieval. A major motivation for this decision was that it was far easier in Europe to find the people and groups with the necessary linguistic competence to handle the language-dependent issues involved in creating test collections in different languages.

Interestingly, the decision to launch CLEF in Europe in 2000 came just one year after the first *NII-NACSIS Test Collection for IR Systems (NTCIR)* workshop on Text Retrieval System Evaluation was held in Asia[4]. NTCIR-1 also included a track for cross-lingual information retrieval with a task in which Japanese queries were used to search an English document collection. Since 1999 NTCIR has added test collections and tasks with Chinese and Korean as well as Japanese and English.

While the first efforts within TREC concentrated on assessing the performance of cross-language systems in which queries in one language were matched against target collections in another, CLEF and NTCIR have taken the concept of "cross-language system evaluation" much further by also including monolingual retrieval in multiple languages and truly multilingual retrieval, i.e. retrieval against target collections containing documents in several languages, in their evaluation exercises.

Both initiatives recognise the importance of good procedures for monolingual retrieval in all or most of the languages concerned when building a multilingual system[5]. The provision of test collections in multiple languages has helped to stimulate research into language-specific aspects of IR.

CLEF first introduced multilingual retrieval in 2002 with a task where the objective was to retrieve and rank relevant documents from a collection in five languages (English, French, German, Italian and Dutch); this was repeated in 2003 with collections in four and eight languages. The NTCIR 2001-2002 campaign also offered a retrieval task on a multilingual collection containing English, Chinese and Japanese documents. These tasks and the test collections produced have offered researchers the opportunity to experiment with the problem of results merging - not just over different collections but also over collections in different languages.

Since 2000, TREC, CLEF and NTCIR have done their best to coordinate their efforts with the aim of promoting complimentary activities. In 2007, in response to requests from colleagues in India, CLEF organised a mono- and cross-language document retrieval task dedicated to Indian languages. This preliminary action helped to lead to the birth of a new evaluation initiative in India: the *Forum for Information Retrieval and Evaluation (FIRE)*[6]. The objective of FIRE is to stimulate the development of IR systems capable of handling the specific needs of the languages of the Indian sub-continent. Contacts have already been established between FIRE, TREC, NTCIR and CLEF.

2.1 Growth of CLEF

When we launched CLEF in 2000, our focus was on text and document retrieval. However, over the years our scope has gradually expanded to include different kinds of text retrieval across languages (ie not just document retrieval but question answering and geographic IR as well) and different kinds of media (ie not just plain text but collections also containing images and speech). The goal has been not only to meet but also to anticipate the emerging needs of the R&D community and to encourage the development of next generation multilingual IR systems. This has meant that the number of tracks offered by CLEF has increased over the years, from just two in 2000 to nine separate tracks in 2008. Each track is run by a coordinating group with specific expertise in the area covered by the track[7]. Most tracks offer several different tasks

[3] See `http://trec.nist.gov/`

[4] See `http://research.nii.ac.jp/ntcir/`

[5] This was noted, for example, by [3] in the blueprint proposed for a successful cross-language system based on an analysis of the results of the TREC CLIR experience and the first four years of CLEF (CLEF2000 - CLEF2003).

[6] `http://www.isical.ac.in/~clia/`

[7] It is impossible to acknowledge all the research organisations that are involved in the coordination of CLEF. A complete list can be found on the homepage of the CLEF website at `http://www.clef-campaign.org/`.

and these tasks normally vary each year, according to the interests of the track coordinators and participants. Figure 1 shows when tracks have been introduced and when they have been terminated.

The growth in tracks has resulted in a rise in participants; with one exception, the number of participating groups has increased every year. This can be seen in Figure 2 which shows the growth in participation by continent, while Figure 3 shows the participation track by track. Note that many groups participate in more than one track.

Full details of the activities and results of each track, year by year, can be found on the CLEF website[8] in the working notes which are produced at the end of each campaign and which contain detailed reports of the experiments of all participating groups plus track overviews in which the results are analysed. In the next section, we comment on the tracks offered in CLEF 2008.

2.2 CLEF 2008: Tracks and Tasks

As can be seen from Figure 1, CLEF 2008 offered nine separate tracks. Two, VideoCLEF and INFILE, were offered this year for the first time as pilot tasks. In addition, MorphoChallenge 2008 was organized in collaboration with CLEF as part of the EU Network of Excellence Pascal Challenge Program[9]. Here below we outline the main features of CLEF 2008.

Multilingual Textual Document Retrieval (Ad Hoc)
The Ad Hoc track is considered as our core track. It is the one track that has been offered each year, from 2000 through 2008, and will be offered again in 2009. The aim of this track is to promote the development of monolingual and cross-language textual document retrieval systems. From 2000 - 2007, the track exclusively used collections of European newspaper and news agency documents and worked hard at offering increasingly complex and diverse tasks, adding new languages each year. Table 1 shows that in the first eight years of the Ad Hoc track, monolingual and bilingual tasks were offered for target collections in twelve different European languages, with bilingual tasks often being proposed for unusual pairs of languages, such as Finish to German, or French to Dutch. In addition multilingual tasks were offered with varying numbers of languages in the target collections.

The results have been considerable; it is probably true to say that this track has done much to foster the creation of a strong European research community in the CLIR area. It has provided the resources, the test collections and also the forum for discussion and comparison of ideas and results. Groups submitting experiments over several years have shown flexibility in advancing to more complex tasks, from monolingual to bilingual and multilingual experiments. Much work has been done on fine-tuning for individual languages while other efforts have concentrated on developing language-independent strategies.

There is also substantial proof of significant increase in retrieval effectiveness in multilingual settings by the systems of CLEF participants. [2] provides a comparison between effectiveness scores from the 1997 TREC-6 campaign and the CLEF 2003 campaign in which retrieval tasks were offered for eight European languages. While in 1997 systems were performing at about 50%–60% of monolingual effectiveness for multilingual settings, that figure had risen to 80%–85% by 2003 for languages that had been part of multiple evaluation campaigns. In the recent campaigns, we commonly see a figure of about 85%–90% for most languages. We find that with languages for which testing has gone on for several years there is usually little variation in performance between the top groups with the best results close to monolingual performance, whereas for "new" languages where there has been little CLIR system testing, there is normally room for improvement.

Interestingly, we find that CLEF participants tend to learn from each other and build up a collective knowhow. Thus, as time passes, we see a convergence of techniques and results with very little statistical difference between the top systems. The best systems tend to be a result of careful tuning of every component, and of combining different algorithms and information sources for every subtask.

In 2008 there was a big change in focus in this track: we introduced very different document collections, a non-European target language, and an *Information Retrieval (IR)* task designed to attract participation from groups interested in *Natural Language Processing (NLP)*. The track was thus structured in three distinct streams.

The first task offered monolingual and cross-language search on library catalog records. It was organized in collaboration with The European Library (TEL)[10] and used three collections from the catalogs of the British Library, the Bibliothéque Nationale de France and the Austrian National Library. The underlying aim was to identify the most effective retrieval technologies for searching this type of very sparse multilingual data. In fact, the collections contained records in many languages in addition to English, French or German. The task presumed a user with a working knowledge of these three languages who wants to find documents that can be useful for them in one of the three target catalogs. Records in other languages were counted irrelevant. This was a challenging task but proved popular; participants tried various strategies to handle the multilinguality of the

[8] http://www.clef-campaign.org/

[9] http://www.cis.hut.fi/morphochallenge2008/

[10] See http://www.theeuropeanlibrary.org/

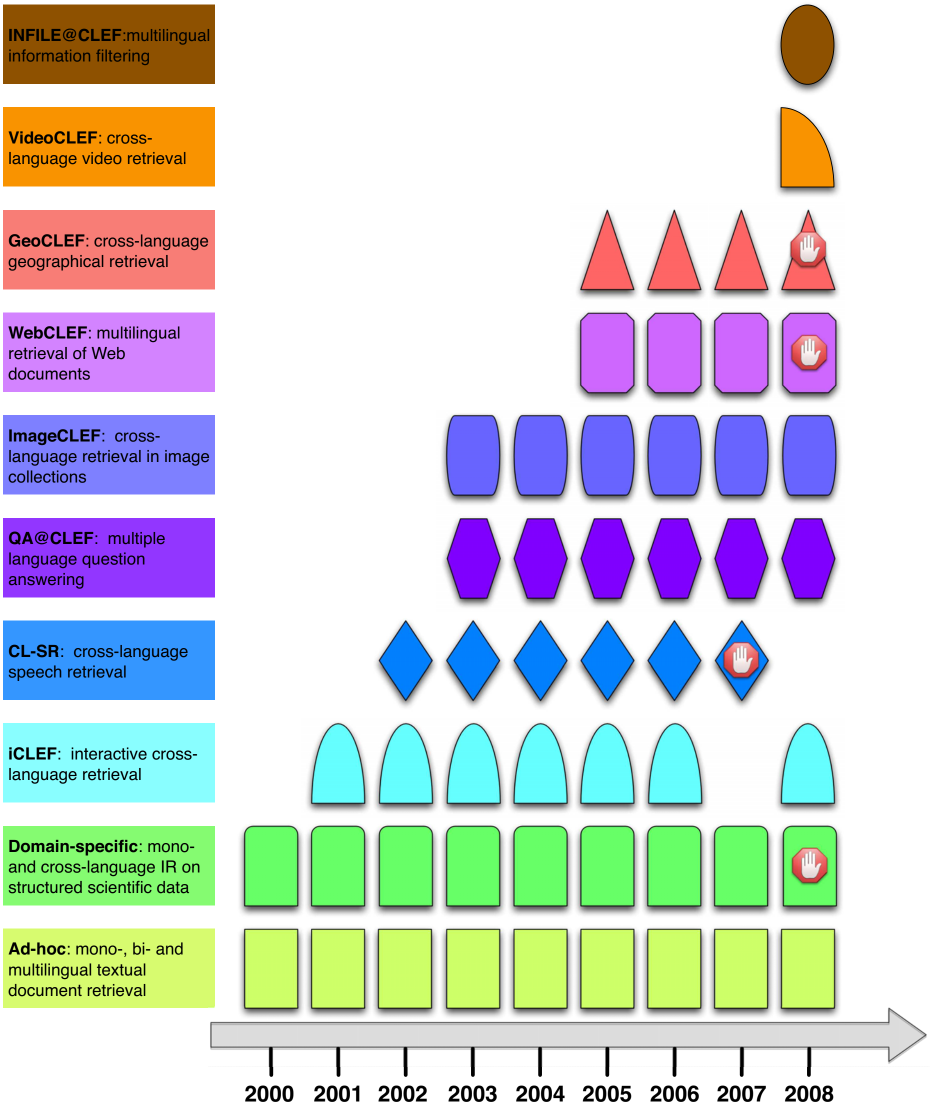

INFILE@CLEF:multilingual information filtering
VideoCLEF: cross-language video retrieval
GeoCLEF: cross-language geographical retrieval
WebCLEF: multilingual retrieval of Web documents
ImageCLEF: cross-language retrieval in image collections
QA@CLEF: multiple language question answering
CL-SR: cross-language speech retrieval
iCLEF: interactive cross-language retrieval
Domain-specific: mono- and cross-language IR on structured scientific data
Ad-hoc: mono-, bi- and multilingual textual document retrieval
2000
2001
2002
2003
2004
2005
2006
2007
2008

Figure 1. CLEF 2000 – 2008 Tracks.

catalogs. The fact that the best results were not always obtained by experienced CLEF participants shows that the traditional approaches used for newspaper document retrieval are not necessarily the most effective for this type of data. The task will certainly be offered again in CLEF 2009.

The Persian@CLEF activity was coordinated in collaboration with the Database Research Group (DBRG) of Tehran University. It was the first time that CLEF offered a non-European language target collection. We chose Persian for several reasons: its challenging script (a modified version of the Arabic alphabet with elision of short vowels) written from right to left; its complex morphology (extensive use of suffixes and compounding); its political and cultural importance. The task used the Hamshahri corpus of 1996-2002 newspapers as the target collection. Monolingual and cross-language (English to Persian) tasks were offered. As was to be expected, many of the eight participants focused their attention on problems of stemming. Only three submitted cross-language runs. The results of the best groups were in line with previous CLEF ad hoc experiments.

The robust task ran for the third time at CLEF 2008. This year it used English test data from previous campaigns but, in addition to the original documents and topics, the organizers provided word sense disambiguated (WSD) documents and topics. Both monolingual and bilingual experiments (topics in Spanish) were activated. The results for the eight participating groups were mixed: while some top scoring groups did manage to improve the results using WSD information in both monolingual and bilingual settings, and the best monolingual robustness (GMAP) score was for a WSD run, the best scores for the rest came from systems which did not use WSD information. Given

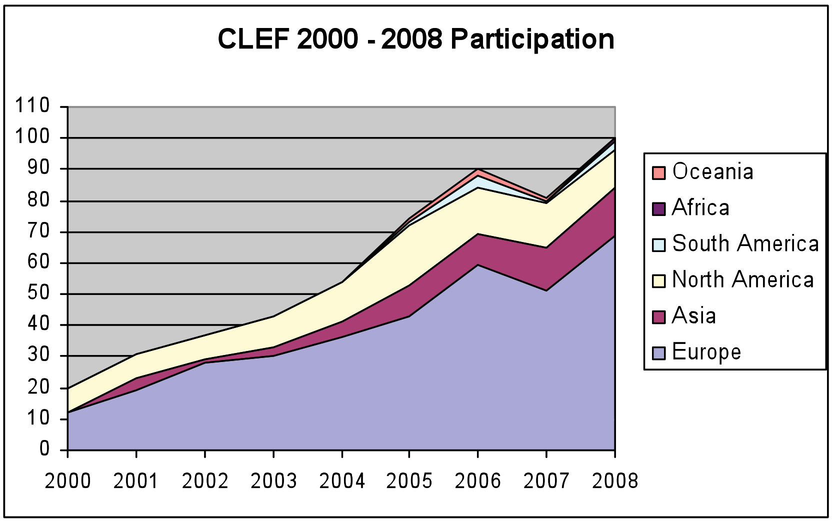


Figure 2. CLEF 2000 – 2008 participation.

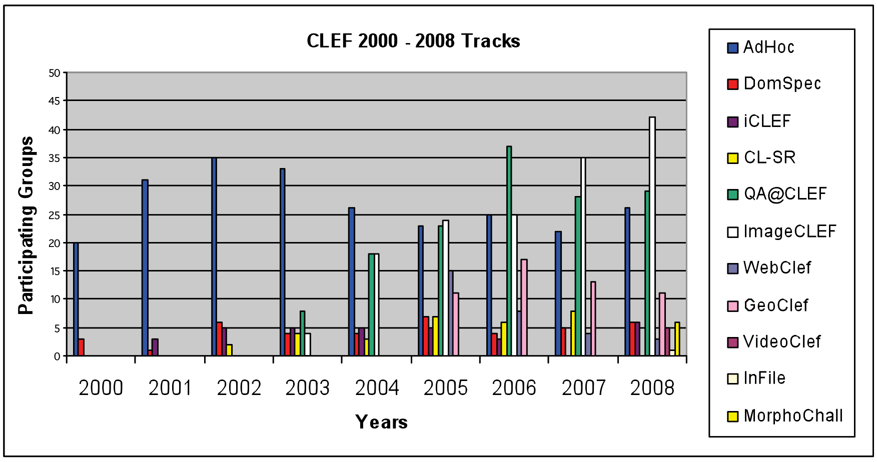


Figure 3. CLEF 2000 – 2008 participation by track.

the relatively short time that the participants had to try effective ways of using the word sense information we think that these results can be considered positive; a subsequent evaluation exercise would be needed for participants to further develop their systems.

Cross-Language Scientific Data Retrieval (Domain-Specific) The focus of this track has been research into how the structure of data in collections (i.e. metadata, controlled vocabularies) can be exploited to improve search. Mono- and cross-language domain-specific retrieval has been studied in the domain of social sciences using structured data (e.g. bibliographic data, keywords, and abstracts) from scientific reference databases. The track has used German, English and Russian target collections in the social science domain. A multilingual controlled vocabulary (German, English, Russian) was also provided. It was decided to terminate this task in 2008 as we felt that it had fulfilled its purpose in providing the community with the opportunity to compare differences between free-text search over languages with structured document retrieval.

In fact, a main finding has been that search on metadata-based documents (just title, abstracts, thesaurus descriptors) can achieve similar results as for full-text archives (ca. 50% in precision as highest result). The results of the monolingual and bilingual domain-specific experiments have been very similar to those achieved in the ad hoc track. Depending on the approach applied, using the controlled vocabulary in the records (thesaurus) can improve retrieval, but may add noise in other cases. Russian has been found to be the hardest language and the results have been much lower than those for English and German. However, this may also depend on the fact that the cor-

Table 1. CLEF 2000–2008 Ad Hoc Tasks. The following ISO 639-1 language codes have been used: am=Amharic; bg=Bulgarian; bn=Bengali; de=German; en=English; es=Spanish; fa=Farsi; fi=Finnish; fr=French; hi=Hindi; hu=Hungarian; id=Indonesian; it=Italian; mr=Marathi; nl=Dutch; or=Oromo; pt=Portuguese; ru=Russian; sv=Swedish; ta=Tamil; te=Telugu.

	Monolingual	Bilingual	Multilingual
CLEF 2000	de;fr;it	x→en	x→de;en;fr;it
CLEF 2001	de;es;fr;it;nl	x→en x→nl	x→de;en;es;fr;it
CLEF 2002	de;es;fi;fr;it;nl;sv	x→de;es;fi;fr;it;nl;sv x→en(newcomers only)	x→de;en;es;fr;it
CLEF 2003	de;es;fi;fr;it;nl;ru;sv	it→es de→it fr→nl fi→de x→ru x→en (newcomers only)	x→de;en;es;fr x→de;en;es;fi;fr;it;nl;sv
CLEF 2004	fi;fr;ru;pt	es;fr;it;ru→fi de;fi;nl;sv→fr x→ru x→en (newcomers only)	x→fi;fr;ru;pt
CLEF 2005	bg;fr;hu;pt	x→bg;fr;hu;pt	Multi8 2yrson (as in CLEF 2003) Multi8 Merge (as in CLEF 2003)
CLEF 2006	bg;fr;hu;pt Robust de;en;es;fr;it;nl	x→bg;fr;hu;pt am;hi;id;te;or→en Robust it→es fr→nl en→de	Robust x→de;en;es;fr;it;nl
CLEF 2007	bg, cz, hu Robust en;fr;pt	x→bg;cz;hu am;id;or;zh→en bn;hi;mr;ta;te→en Robust x→en;fr;pt	
CLEF 2008	fa TEL de;en;fr Robust WSD en	en→fa TEL x→de;en;fr Robust WSD es→en	

pus is less well-formed (fewer abstracts, noisy keywords). The existence of a bilingual thesaurus can help a lot in translating technical language which is usually not contained in general-purpose dictionaries. It was also consistently found that blind feedback mechanisms improved retrieval.

Interactive Cross-Language Retrieval (iCLEF) In iCLEF, cross-language search capabilities are studied from a user-inclusive perspective. A central research question is how best to assist users when searching information written in unknown languages, rather than how best an algorithm can find information written in languages different from the query language. Since 2006, iCLEF has moved from news collections (a standard for text retrieval experiments) in order to explore user behaviour in a collection where the cross-language search necessity arises more naturally for average users. The choice fell on Flickr[11], a large-scale, online image database based on an extensive social network of WWW users, with the potential for offering both challenging and realistic multilingual search tasks for interactive experiments. The search interface provided by the iCLEF organizers was a basic cross-language retrieval system for the Flickr image database presented as an online game: the user is given an image, and must find it again without any a priori knowledge of the language(s) in which the image is annotated. The game was publicized on the CLEF mailing list and prizes were offered for the best results in order to encourage participation. The main novelty of the iCLEF 2008 experiments was the shared analysis of a search log from a single search interface provided by the organizers (i.e. the focus was on log analysis, rather than on system design). Search logs

[11] See http://www.flickr.com/

were harvested from the search interface described above and iCLEF participants could essentially do two things:

- Search log analysis: participants had access to the search logs, and could freely perform data mining studies on them, such as looking for differences in search behaviour according to language skills, or looking for correlations between search success and search strategies, etc.

- Interactive experiments: participants could recruit their own users and conduct their own experiments with the interface. For instance, they could recruit a set of users with passive abilities and another with active abilities in certain languages and, besides studying the search logs, could perform observational studies on how they search, conduct interviews, etc.

The 2008 experiments resulted in a truly reusable data set (the first time in iCLEF!), with 5,000 complete search sessions recorded and 5,000 post-search and post-experience questionnaires. 200 users from 40 countries played an active role in these experiments which covered six target languages. Six groups submitted results (4 log analyses, 2 observational studies). It was possible to quantify the differences (in success, behaviour, satisfaction) between different user profiles (active, passive, unknown) and search settings (mono, bi, multilingual). A main observation was that, using the same cross-language search engine (with standard interactive CLIR facilities, such as assisted translation), users with active or passive abilities in the target language find relevant images faster and more accurately than users with no knowledge of the target language. In other words, in addition to better CLIR algorithms, we need more research on interactive features to help users bridge the language gap.

Multilingual Question Answering (QA@CLEF) This track has been offering monolingual and cross-language question answering tasks since 2003. QA@CLEF 2008 proposed both main and pilot tasks. The main scenario was event-targeted QA on a heterogeneous document collection (news articles and Wikipedia). A large number of questions were topic-related, i.e. clusters of related questions possibly containing anaphoric references. Besides the usual news collections, articles from Wikipedia were also considered as sources of answers. Many monolingual and cross-language sub-tasks were offered: Basque, Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish were proposed as both query and target languages; not all were used in the end.

After 6 years, a lot of resources and know-how have been accumulated. However, the tasks offered have proved to be difficult for the systems which have not shown a very good overall performance, even those that have participated year by year. In addition, a result of offering so many language possibilities has meant that there have always been very few systems participating in the same task, with the same languages. This has meant that comparative analysis is extremely problematic. Consequently, the QA organisers have decided to redefine the task for CLEF 2009 to permit the evaluation and comparison of systems even when they are working in different languages. The new setting will also take as reference a real user scenario, in a new document collection in which multilinguality is more natural.

The additional exercises were the following:

- The Answer Validation Exercise (AVE) in its third edition was aimed at evaluating answer validation systems based on recognizing textual entailment. Like last year, all participating systems employed lexical processing. However, this year there were more groups using syntactic processing, mainly chunking or dependency analysis. The application of semantic analysis decreased while the use of WordNet increased (50% of participants used it). 7 out of 9 groups used a Named Entity Recognizer and most groups applied Machine Learning techniques. Support vector machines (SVM) and decision trees were the most used classifiers. There is insufficient evidence to say that one performed better than the other. Overall, it seems that more sophisticated tools do not imply better performance.

- QAST was focused on Question Answering over Speech Transcriptions of seminars. In this 2nd year pilot task, answers to factual and definitional questions in English were to be extracted from spontaneous speech transcriptions related to separate scenarios in English, French and Spanish. Five groups participated across ten tasks that included dealing with different types of speech (spontaneous or prepared), different languages (English, Spanish and French) and different word error rates for automatic transcriptions (from 10.5% to 35.4%). For the tasks where the word error rate was low enough (around 10%) the loss in accuracy compared to manual transcriptions was under 5%, suggesting that QA in such documents is potentially feasible. However, even where automatic speech recognition (ASR) performance is reasonably good, there remain many challenges when dealing with spoken language. The results from the QAST evaluation indicate that if a QA system performs well on manual transcriptions it will also performs reasonably well on high quality automatic transcriptions.

- QA-WSD provided questions and collections with already disambiguated Word Senses in order to study their contribution to QA performance. Unfortunately, just one groups participated in this task so the results were not significant.

Cross-Language Retrieval in Image Collections (ImageCLEF) This track evaluated retrieval of images from multilingual collections; both text and visual retrieval techniques were exploitable. Five challenging tasks were offered in 2008:

- A photo retrieval task: a good image search engine ensures that duplicate or near duplicate documents retrieved in response to a query are hidden from the user. Ideally the top results of a ranked list will contain diverse items representing different sub-topics within the results. This task focused on the study of successful clustering to provide diversity in the top-ranked results. The target collection contained images with captions in English and German; queries were in English.
- A medical image retrieval task: this is a domain-specific retrieval task in a domain where many ontologies exist; the target collection was a subset of the Goldminer collection containing images from English articles published in Radiology and Radiographics with captions and html links to the full text articles. Queries were provided in English, French and German.
- A visual concept deception task: the objective was to identify language-independent visual concepts that would help in solving the photo retrieval task. A training database was released with approximately 1,800 images classified according to a concept hierarchy. This data was used to train concept detection/annotation techniques. For each of the 1,000 images in the test database, participating groups were required to determine the presence/absence of the concepts.
- An automatic medical image annotation task: automatic image annotation or image classification can be an important step when searching for images from a database of radiographs. The aim of the task was to find out how well current language-independent techniques can identify image modality, body orientation, body region, and biological system on the basis of the visual information provided by the images.
- A Wikipedia image retrieval task: this was an ad hoc image search task where the information structure can be exploited for retrieval. The aim was to investigate retrieval approaches in the context of a larger scale and heterogeneous collection of images (similar to those encountered on the Web) that are searched for by users with diverse information needs.

As can be seen from Figure 3, ImageCLEF has become the most popular of all tracks, even though (or maybe because) it is the track that deals the least with language and linguistic issues. One of the secrets of its popularity is that image search has a number of well-known applications. In CLEF we focus on one of the most important: medical image processing and analysis. A major focus is on the combination of text and image features to improve search. In this respect, it is interesting to note that one of the findings of the coordinators of the medical tasks this year was that from an examination of mixed media runs that had corresponding text-only runs, it was clear that combining good textual retrieval techniques with questionable visual retrieval techniques can negatively affect system performance. This demonstrates the difficulty of usefully integrating both textual and visual information, and the fragility that such combinations can introduce into retrieval systems [7]

Multilingual Web Retrieval (WebCLEF) In the past three years this track has focused on evaluation of systems providing multi- and cross-lingual access to web data. WebCLEF 2008 repeated the track setup of the 2007 edition. In 2008, a multilingual information synthesis task was offered, where, for a given topic, participating systems were asked to extract important snippets from web pages (fetched from the live web and provided by the task organizers). The systems had to focus on extracting, summarizing, filtering and presenting information relevant to the topic, rather than on large scale web search and retrieval per se. The focus was on refining the assessment procedure and evaluation measures. WebCLEF 2008 had lots of similarities with (topic-oriented) multi-document summarization and with answering complex questions. An important difference is that at WebCLEF, topics could come with extensive descriptions and with many thousands of documents from which important facts have to be mined. In addition, WebCLEF worked with web documents, which can be very noisy and redundant.

Although the Internet would seem to be the obvious application scenario for a CLIR system, WebCLEF has had a rather disappointing participation. For this reason, we have decided to drop this track – at least for the coming year.

Cross-Language Geographical Retrieval (GeoCLEF) The purpose of GeoCLEF is to test and evaluate cross-language geographic information retrieval for topics with a geographic specification. How best to transform into a machine readable format the imprecise description of a geographic area found in many user queries is still an open research problem. As in previous years, GeoCLEF 2008 examined geographic search of a text corpus. Some topics simulated the situation of a user who poses a query when looking at a map on the screen.

In GeoCLEF 2006 and 2007, it was found that keyword based systems often do well on the task and the best systems worked without any specific geographic resource. In 2008 the best monolingual systems used specific geo reasoning; there was much named-entity recognition (often using Wikipedia) and NER topic parsing. Geographic ontologies were also used (such as GeoNames and World Gazeteer), in particular for query expansion. However, as in previous years, in the cross-language tasks, the best systems used no specific geo components; standard approaches like BM25 and blind relevance feedback worked well. It has been decided to terminate GeoCLEF this year; however, a new track in 2009, LogCLEF, will continue to study information retrieval problems from the geographical perspective.

Cross-Language Video Retrieval (VideoCLEF) VideoCLEF used a video corpus containing episodes of a dual language television program in Dutch and English. Three tasks were offered: (1) Automatic assignment of subject tags (i.e., classification), (2) Automatic translation of metadata for visualization, and (3) Automatic selection of semantically representative keyframes. The dual language programming of Dutch TV offered a unique scientific opportunity, presenting the challenge of how to exploit speech features from both languages.

Five research groups participated in this track. Participants were supplied with archival metadata including title and description, shot boundaries, shot-level keyframes and automatic speech recognition (ASR) transcripts in both Dutch and English. The video content was chosen to reflect the cultural heritage domain and the subject labels used in the automatic classification task were selected to be representative of cultural heritage themes.

The results of the classification task (i.e., automatic assignment of subject labels to videos) demonstrated that classification of dual language video content is in no way trivial. It was, however, possible to reach satisfactory classification performance, particularly on individual classes The translation task was a success. We were able to conclude that if non-Dutch speakers wish to access content in a Dutch-language archive to find embedded English-language interviews, translation of the metadata is NOT a bottleneck. Finally, the semantic keyframe selection task also yielded very encouraging results. Shot-level keyframes were automatically selected on the basis of the spoken content of the shot in order to represent the entire television documentary. The automatically selected keyframes were competitive with human selected keyframes.

Multilingual Information Filtering (INFILE@CLEF) INFILE (INformation, FILtering & Evaluation) was a cross-language adaptive filtering evaluation track sponsored by the French National Research Agency. INFILE extended the last filtering track of TREC 2002 in the following ways:

- Monolingual and cross-language tasks were offered using a corpus of 100,000 Agence France Press (AFP) comparable newswire stories for Arabic, English and French;
- Evaluation was performed by an automatic interrogation of test systems with a simulated user feedback. A curve of the evolution of efficiency was computed along with more classical measures tested in TREC.

Details on the technical infrastructure, the organisation and an analysis of the results of all these tracks can be found in the track overview reports in the CLEF 2008 Working Notes[12].

2.3 Agenda for CLEF 2009

As we write, the final agenda for CLEF 2009 is being decided. There will be eight main evaluation tracks - six have been inherited from last year [13] and two new ones have been added: Intellectual Property (CLEF-IP) and Log File Analysis (LogCLEF). These are described below.

Intellectual Property (CLEF–IP) The CLEF–IP track in 2009 will utilize a collection of more than 1M patent documents mainly derived from sources of the European Patent Office. The collection will cover English French and German with at least 100,000 documents in each language. Queries and relevance judgements will produced by two methods. The first is using queries produced by Intellectual Property Experts and reviewed by them in a fairly conventional way. The second is an automatic method using patent citations from seed patents. Search results will be reviewed to ensure that the majority of test and training queries produce results in more than one language. We will primarily report results retrieving documents across all three languages. In 2009 we will stick to the Cranfield evaluation model: in subsequent years we expect to offer refined retrieval process models and assessment tools[14].

[12] See `http://www.clef-campaign.org/`

[13] As stated in the previous section, the Domain-Specific, WebCLEF and GeoCLEF tracks were terminated in 2008.

[14] See `http://www.ir-facility.org/the_irf/clef-ip09.htm`

Log File Analysis (LogCLEF) LogCLEF deals with the analysis of queries as expression of user behavior. The goal is the analysis and classification of queries in order to improve search systems. It builds on some of the work of GeoCLEF 2007. LogCLEF has two tasks:

- Log Analysis and Geographic Query Identification (LAGI): The recognition of the geographic component within a query stream is a key problem for geographic information retrieval (GIR). Geographic queries require specific treatment and often a geographically oriented output (e.g. a map). The task is to (1) classify geographic queries and (2) identify their geographic and non-geographic elements. A real search engine log file and logs from The European Library (TEL) will be used.
- Log Analysis for Digital Societies (LADS): This task will use logs from The European Library (TEL) and intends to analyze user behavior with a focus on multilingual search. The task is open to different approaches. Potential targets are query reformulation, multilingual search behavior and community identification[15]

CLEF 2009 will also include an experimental pilot task: GridCLEF. Our ideas for this activity are described below in Section 4.1

The preliminary Call for Participation for CLEF 2009 can be found at `http://www.trebleclef.eu/`

2.4 Growth of Test Collections

CLEF campaigns mainly adopt a comparative evaluation approach in which system performances are compared according to the Cranfield methodology, which makes use of experimental collections. The CLEF test collections are thus made up of documents, topics and relevance assessments. The topics are created to simulate particular information needs from which the systems derive the queries to search the document collections. System performance is evaluated by judging the results retrieved in response to a topic with respect to their relevance, and computing the relevant measures, depending on the methodology adopted by the track.

A number of different document collections were used in CLEF 2008 to build the test collections:

- CLEF multilingual corpus of more than 3 million news documents in 14 European languages. This corpus is divided into two comparable collections: 1994-1995 - Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish, Swedish; 2000-2002 - Basque, Bulgarian, Czech, English, Hungarian. The Basque data was new this year. Parts of this collections were used in the AdHoc, QuestionAnswering, GeoCLEF and MorphoChallenge tracks.
- Data from The European Library (TEL): approximately 3 million library catalog records in English, French and German, used in the Ad Hoc track.
- Hamshahri Persian newspaper corpus; nearly 170,000 documents used in the Ad Hoc track;
- The GIRT-4 social science database in English and German (over 300,000 documents) and two Russian databases: the Russian Social Science Corpus (approx. 95,000 documents) and the Russian ISISS collection for sociology and economics (approx. 150,000 docs). The RSSC corpus was not used this year. Cambridge Sociological Abstracts in English (20,000 docs). These collections were used in the domain-specific track.
- Online Flickr database, used in the iCLEF track
- The ImageCLEF track used collections for both general photographic and medical image retrieval:
  - IAPR TC-12 photo database of 20,000 still natural images (plus 20,000 corresponding thumbnails) with captions in English, and German;
  - ARRS Goldminer database - nearly 200,000 images published in 249 selected peer-reviewed radiology journals
  - IRMA collection in English and German of 12,000 classified images for automatic medical image annotation
  - INEX Wikipedia image collection, approximately 150,000 images associated with unstructured and noisy textual annotations in English
- Videos in Dutch and English of documentary television programs, approximately 30 hours, used in the VideoCLEF track.
- Agence France Press (AFP) comparable newswire stories in Arabic, French and English for the INFILE track

The number, format and languages of the topic sets created varied according to the needs of the particular track. In all cases, ground truth creation was done using human resources.

[15] See `http://www.uni-hildesheim.de/logclef/`

These test suites form valuable and reusable resources. They are created according to rigorous guidelines and are tested to confirm their stability. An official CLEF Test Suite consisting of the data created for the monolingual, bilingual, multilingual and domain-specific text retrieval tracks for the CLEF 2000-2003 Campaigns is now publicly available. It consists of multilingual document collections in eight languages; step-by-step documentation on how to perform a system evaluation; tools for results computation; multilingual sets of topics; multilingual sets of relevance assessments; guidelines for participants (in English); tables of the results obtained by the participants; publications[16]. We are now planning to release additional test collections to cover later years.

2.5 The Evaluation Infrastructure

The current approach to experimental evaluation is mainly focused on creating comparable experiments and evaluating their performance whereas researchers would also greatly benefit from an integrated vision of the scientific data produced, together with analyses and interpretations, and from the possibility of keeping, re-using, and enriching them with further information. The way in which experimental results are managed, made accessible, exchanged, visualized, interpreted, enriched and referenced is an integral part of the process of knowledge transfer and sharing towards relevant application communities.

The University of Padua has thus developed *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)*[17], a sophisticated evaluation infrastructure which provides a set of tools capable of managing high-level tasks, such as topic creation, experiment submission, pool assessment, relevance assessment, statistical analysis on the experiments, etc. [4]. DIRECT supports a data curation approach within CLEF as an extension to the traditional methodology in order to better manage, preserve, interpret and enrich the scientific data produced, and to effectively promote the transfer of knowledge. A detailed description of the DIRECT architecture and functionality is provided in [1].

DIRECT has been successfully employed in the last four CLEF evaluation campaigns (2005, 2006, 2007 & 2008) managing the technical infrastructure for a number of tracks: Ad Hoc, Domain-Specific and Geo-CLEF, and providing procedures to handle:

- the track set-up, harvesting and processing of the document collections, management of topic creation procedures
- the registration of participants to tracks, submission of experiments, collection of metadata about experiments, and their validation;
- the creation of document pools and the management of relevance assessment;
- the provision of common statistical analysis tools for both organizers and participants in order to allow the comparison of the experiments;
- the provision of common tools for summarizing, producing reports and graphs on the measured performances and conducted analyses.

An extension to DIRECT to manage the technical infrastructure for ImageCLEF is now under discussion.

2.6 Main results

Over the years, CLEF has played an important role in stimulating research activity in new, previously unexplored areas, such as cross-language question answering, image and geographic information retrieval and in encouraging the study and implementation of evaluation methodologies for diverse types of cross-language IR systems. This has lead to the building up of a strong multidisciplinary research community and to the creation of important, reusable test collections for system benchmarking[18]. The research activities promoted by CLEF have provided valuable quantitative and qualitative evidence with respect to best practice in cross-language system development. For example, CLEF evaluations have provided evidence along the years as to which methods give the best results in certain key areas, such as multilingual indexing, query translation, resolution of translation ambiguity, results merging [3]. DIRECT is a valuable tool when assessing the results over the years as it stores the data over the years and enables us to make all kinds of in-depth analyses.

However, although CLEF has done much to promote the development of multilingual IR systems, the focus has been on building and testing research prototypes rather than developing fully operational systems. The challenge that we are now attempting to tackle is how to best transfer these research results to the market place. How we are now trying to face this challenge is described in the following section.

3 TrebleCLEF

For many years, CLEF has thus been a forum where researchers can perform experiments, discuss results and exchange ideas; most of the results have

[16] The Evaluation Package is now available in the European Language Resources Association (ELRA) catalogue (ref. ELRAE0008). Information can be found at: `http://catalog.elra.info/`.

[17] `http//direct.dei.unipd.it/`

[18] The 2000-2003 test collections are now publicly available on the *Evaluation and Language resources Distribution Agency (ELDA)* catalog, see `http://www.elda.org/`.

been published but the extensive CLEF-related literature is mainly intended for the academic community. Contacts with interested application communities have been notably lacking.

In fact, evaluation campaigns have their limitations. They tend to focus on aspects of system performance that can be measured easily in an objective setting (e.g. precision and recall) and to ignore others that are equally important for overall system development. Thus, while in CLEF, much attention has been paid to improving performance in terms of the ranking of results through the refining of query expansion procedures, term weighting schemes, algorithms for the merging of results, equally important criteria of speed, stability, usability have been mainly ignored. Clearly, in any real world *MultiLingual Information Access (MLIA)* system, the results must be presented in an understandable and useful fashion. The user interface implementation thus needs to be studied very carefully, according to the particular user profile. Such aspects tend to be neglected in traditional evaluation campaigns.

At the beginning of 2008 we thus launched a new activity which aims at building on and extending the results already achieved by CLEF. This activity, called TrebleCLEF , aims at stimulating the development of operational MLIA systems rather than research prototypes.

TrebleCLEF is promoting research, development, implementation and industrial take-up of multilingual, multimodal information access functionality in the following ways:

- by continuing to support the annual CLEF system evaluation campaigns with tracks and tasks designed to stimulate R&D to meet the requirements of the user and application communities, with particular focus on the following key areas:
  - user modeling, e.g. the requirements of different classes of users when querying multilingual information sources;
  - language-specific experimentation, e.g. looking at differences across languages in order to derive best practices for each language, best practices for the development of system components and best practices for MLIA systems as a whole;
  - results presentation, e.g. how can results be presented in the most useful and comprehensible way to the user.
- by constituting a scientific forum for the MLIA community of researchers enabling them to meet and discuss results, emerging trends, new directions:
  - providing a scientific digital library to manage accessible the scientific data and experiments produced during the course of an evaluation campaign. This library would also provide tools for analyzing, comparing, and citing the scientific data of an evaluation campaign, as well as curating, preserving, annotating, enriching, and promoting the re-use of them;
- by acting as a virtual centre of competence providing a central reference point for anyone interested in studying or implementing MLIA functionality and encouraging the dissemination of information:
  - making publicly available sets of guidelines on best practices in MLIA (e.g. what stemmer to use, what stop list, what translation resources, how best to evaluate, etc., depending on the application requirements);
  - making tools and resources used in the evaluation campaigns freely available to a wider public whenever possible; otherwise providing links to where they can be acquired;
  - organising workshops, and/or tutorials and training sessions.

Thus with TrebleCLEF we hope to bridge the gap between research activities promoted in CLEF and the application of the results in a real-world context. The first major results of this activity will be seen in 2009 with the publication of three Best Practices studies:

- Best Practices in Language Resources for Multilingual Information Access
- Recommendations for Best Practices in System and User-oriented Multilingual Information Access
- Best Practices for Test Collection Creation, Evaluation Methodologies and Language Processing Technologies

In addition, we will be holding a Summer School in June 2009 aimed at providing a grounding in the core topics that constitute the multidisciplinary area of Multilingual Information Access and a MLIA Technology Day will be held in autumn 2009 in order to disseminate the results of our studies to the application communities. In this way, we hope to provide a real contribution not only to MLIA research but also in the MLIA application communities.

4 Ad Hoc as a Case Study

As has been stated, the Ad Hoc track has always been considered as the core track in CLEF and is the

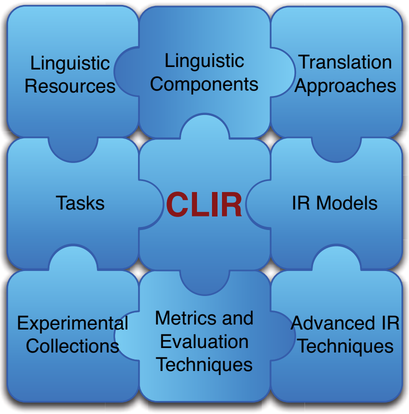


Figure 4. CLIR areas explored by the Ad Hoc track over the years.

starting point for many groups as they begin to be interested in developing functionality for the multilingual information access.

Our objective in this track has been to promote R&D in monolingual, bilingual and multilingual text retrieval. As can be seen in Table 1, the different adhoc tasks present varying degrees of difficulty: there are more basic tasks, such as the monolingual tasks or the bilingual English tasks, designed to encourage inexperienced groups to experiment and increase their knowhow; there are intermediate tasks, such as the bilingual task with unusual pair of languages, where groups can try to apply more advanced techniques or experiment their own consolidated techniques in a more challenging scenario; finally, there are advanced tasks, such as the multilingual and robust tasks, where groups have to address difficult issues and discover innovative solutions. In this way, over the years, we have offered different entry points to the fields of CLIR and MLIA in order to support the creation and growth of a research community with diversified expertise. In addition, as shown in Figure 4, we have promoted R&D in the multilingual field through the exploration of a comprehensive set of CLIR-related topics:

- **experimental collections**: test collections have been built for as many European languages as possible, attempting to cover diverse language typologies; in 2008 we added the first non-European language in order to add another dimension of complexity;
- **tasks**: groups have been stimulated to experiment with retrieval over unusual pairs of languages and retrieval from collections of multiple languages with diverse characteristics, such as long documents, sparse information, and so on;
- **linguistic resources**: the development and/or use of language resources, such as stop lists, dictionaries, lexicons, aligned and parallel corpora, etc., has been supported;
- **linguistic components**: the development and/or application of linguistic tools, such as stemmers, lemmatizers, decompounders, part of speech taggers, and so on, has been fostered;
- **translation approaches**: groups have been encouraged to experiment with different approaches for crossing language barriers, such as *Machine Translation (MT)*, and dictionary-based, parallel corpora-based, or conceptual network-based translation mechanisms;
- **IR models**: different models have been studied and applied – boolean, vector space, probabilistic, language models, and so on – to improve retrieval performances across languages;
- **advanced IR techniques**: advanced techniques, such as data fusion and merging or relevance feedback, have been adopted to address issues such as the need for query expansion to improve translation or the fusion of multilingual results;
- **metrics and evaluation techniques**: metrics to analyse system behaviour in a multilingual setting and compare performances across languages and tasks have been developed and employed.

Much of the effort of CLEF over the years has been devoted to the investigation of key questions such as “What is CLIR?”, “What areas should it cover?” and “What resources, tools and technologies are needed?” As mentioned in Section 1, CLEF began when CLIR was just starting to be recognized as an independent sub-discipline and thus promoted much pioneering work in the field.

In recent editions of CLEF, the ad-hoc track has started to explore finer-grained questions in the CLIR scenario. An example of this is the CLEF 2008 TEL task which explores effective approaches when searching collections of multilingual document surrogates – in this case catalog records – to determine whether the documents described by such surrogates are relevant to a given information need. This task focuses its attention on a specific data type, i.e. sparse and semi-structured catalog records, in an inherently multilingual collection. These features are particularly challenging from a retrieval point of view, since the catalog records have to be suitably processed and expanded and the intrinsic multilinguality of the collection has to be catered for with techniques that go beyond the traditional fusion strategies adopted in previous multilingual tasks. It is not expected that new language resources or linguistic tools will be produced in this task but rather that already existing ones will be exploited. The TEL task represents an example of a task that focuses more on retrieval issues than on language or linguistic aspects and is further evidence that cross-language information retrieval is much more than simple machine translation plus information retrieval. Exploiting the CLIR acronym, we could say that it is a CL**IR** task, meaning that it stresses the importance of the retrieval techniques in a multilingual setting.

The opposite example is provided by the CLEF 2008 Robust *Word Sense Disambiguated (WSD)* task which focuses on the benefits that a deeper and more sophisticated linguistic analysis can produce in a multilingual setting, especially when hard topics are being handled and the aim is to achieve robust performances across the set of topics. From a retrieval point of view, the necessary techniques are well-known and it is not expected that participants produce new IR components. On the other hand, the development and adoption of word sense disambiguation algorithms and their introduction into a consolidated retrieval pipeline puts attention on the linguistic part of the process. In this case, we could say that this is a **CL**IR task.

These two examples show that we are now ready to move from a breadth-wise exploration of the CLIR field to a deeper investigation of each specific area with the objective of acquiring a more profound understanding of the basic mechanisms.

This new approach is taken a step further in the Grid@CLEF Pilot task[19] which is described in the next section.

4.1 Grid@CLEF

This task has been proposed for CLEF 2009 with the following goals in mind:

- to look at differences across a wide set of languages;
- to identify best practices for each language;
- to help other countries to develop their expertise in the IR field and create IR groups;
- to provide a repository, in which all the information and knowledge derived from the experiments undertaken can be managed and made available via the DIRECT system.

Individual researchers or small groups do not usually have the possibility of running large-scale and systematic experiments over a large set of experimental collections and resources. Figure 5 depicts the performances, e.g. mean average precision, of the composition of different IR components across a set of languages as a kind of surface area which we intend to explore with our experiment. The average CLEF participants, shown in Figure 5(a), may only be able to sample a few points on this surface since, for example, they usually test just a few variations of their own or customary IR model with a stemmer for two or three languages. Instead, the expert CLEF participant, represented in Figure 5(b), may have the expertise and competence to test all the possible variations of a given component across a set of languages, as [11] does for stemmers, thus investigating a good slice of the surface area.

However, even though each of these cases produces valuable research results and contributes to the advancement of the discipline, they are both still far removed from a clear and complete comprehension of the features and properties of the surface represented in Figure 5. A far deeper sampling would be needed for this.

It is our hypothesis that a series of systematic grid experiments can re-use and exploit the valuable resources and experimental collections made available by CLEF in order to gain more insights about the effectiveness of, for example, the various weighting schemes and retrieval techniques with respect to the languages. This knowledge could then be disseminated to both the research and the application communities.

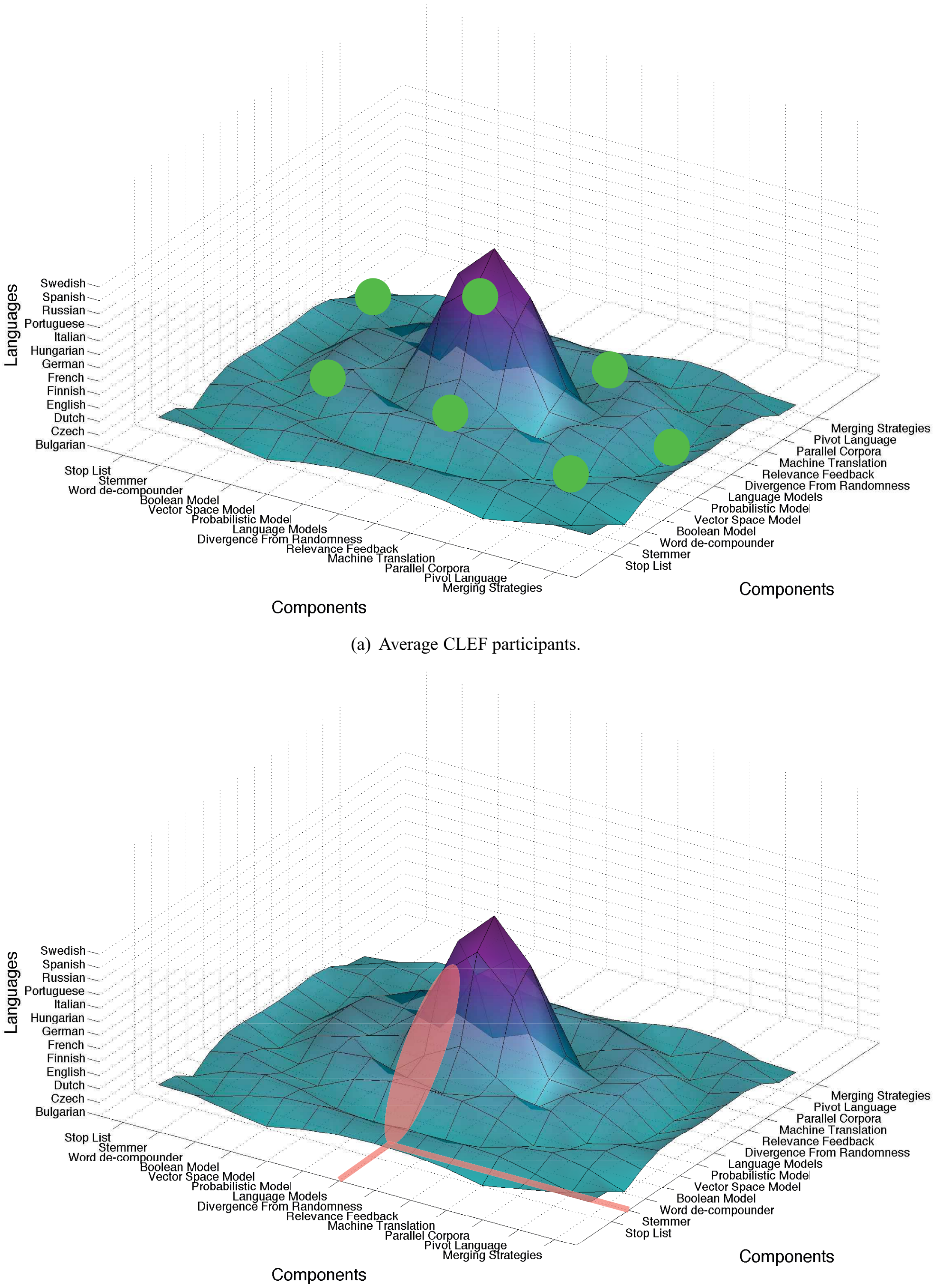
In order to do this, we must deal with the interaction of three main entities, as shown in Figure 6:

- **Component**: in charge of carrying out one of the steps of the IR process;
- **Language**: will affect the performance and behaviour of the different components of an *Information Retrieval System (IRS)* depending on its specific features, e.g. alphabet, morphology, syntax, and so on.
- **Task**: will impact on the performances of IRS components according to its distinctive characteristics;

We assume that the contributions of these three main entities to retrieval performance tend to overlap; nevertheless, at present, we do not have enough knowledge about this process to say whether, how, and to what extent these entities interact and/or overlap – and how their contributions can be combined, e.g. in a linear fashion or according to some more complex relation.

The above issue is in direct relationship with another long-standing problem in the IR experimentation: the impossibility of testing a single component independently of a complete IRS. [10, p. 12] points out that "if we want to decide between alternative indexing strategies for example, we must use these strategies *as part of a complete information retrieval system*, and examine its overall performance (with each of the alternatives) directly". This means that we have to proceed by changing only one component at

[19] `http://ims.dei.unipd.it/gridclef/`



(a) Average CLEF participants.

(b) Expert CLEF participant.

Figure 5. Coverage achieved by different kinds of participants.

time and keeping all the others fixed, in order to identify the impact of that component on retrieval effectiveness; this also calls for the identification of suitable baselines with respect to which comparisons can be made.

It is our aim to be able to re-use the existing CLEF collections and exploit the specific competence of each participant. Therefore, in order to run these grid experiments, we need to set up a framework in which participants can exchange the intermediate output of the components of their systems and create a run by using the output of the components of other participants.

For example, if the expertise of participant A is in building stemmers and decompounders while participant B's expertise is in developing probabilistic IR models, we would like to make it possible for participant A to apply his stemmer to a document collection, pass the output to participant B, who tests his probabilistic IR model, thus obtaining a final run which represents the test of participant A' stemmer + participant B probabilistic IR model.

Basically, there are two possible alternatives to achieve this type of collaboration:

- build a common software framework into which participants can plug their components and which takes care of executing the whole run;
- build a common message framework (XML-based) where participants publish the output of their components and receive the input of other components in order to create a complete run.

The first approach implies building a software framework and adopting a particular technology, e.g. Java. This would involve considerable effort even if existing solutions, such as for example *TERabyte RetrIEveR (TERRIER)*[20], are adapted for the purpose. This could cause participants to feel forced to produce (new) code in a different way than usual, and might present problems with the integration of legacy code. The second approach poses challenges with respect to the representation of a component's output and the orchestration of the various messages among participants, but could turn out to be the more flexible or, at least, pose less burden on the participants.

In order to carry out these experiments, we plan to lever on previous and partially similar experiences. For example, during the *Reliable Information Access (RIA)* Workshop [5], organized by the US *National Institute of Standards and Technology (NIST)* in 2003, an experiment called `swapterms` was conducted. `swapters` tested the performances of relevance feedback when one system uses n terms determined by another system in order to expand the query. In another example at the NTCIR-7 *Advanced Cross-Lingual Information Access (ACLIA)*[21], an IR system could use the results of analyses performed by an external *Question Answering (QA)* system in order to search documents for a given question; similarly, a QA system could use the retrieval results of an external IR system in order to answer a question. Both these experiences represent cases of exchange of information among components of different systems and represent a valuable input for grid experiments.

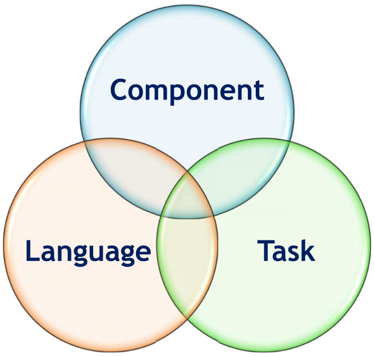


Figure 6. The three main entities involved in grid experiments.

The Pilot Grid task in CLEF 2009 will provide us with an opportunity to begin to set up a suitable framework in order to carry out a first set of experiments which will allow us to acquire an initial set of measurements and to start to explore the interaction among IR components and languages. This initial knowledge will allow us to tune the overall protocol and framework, to understand what directions are more promising, and to scale the experiments up to a finer-grain comprehension of the behaviour of IR components across languages.

5 Conclusions

In this paper, we have attempted to show how CLEF has progressed over the years from a small research action studying cross-language textual document retrieval for a few of the most widely used European languages to a large-scale activity that aims at promoting R&D in the multidisciplinary area of multilingual, multimodal information retrieval by providing an infrastructure which allows communities of researchers to collaborate together on a large-scale and also by disseminating those results to the outside world.

We have already taken important first steps in this direction both through the organization of a series of dissemination activities in TrebleCLEF, as mentioned above, and also in the design of the CLEF evaluation activities. The CLEF 2008 and 2009 tracks have been studied not just to meet the requirements of research but also to better reflect the needs of the users. Important examples are the ongoing collaboration with The

[20] `http://ir.dcs.gla.ac.uk/terrier/index.html`

[21] `http://aclia.lti.cs.cmu.edu/wiki/Home`

European Library in the Ad Hoc track, the user studies being undertaken by the Interactive track, the log file analysis of LogCLEF, the work with radiographic images in ImageCLEFmed, and the Intellectual Property track to begin in CLEF 2009.

In the coming years, it is our intention to focus much of our efforts in three directions:

- in-depth analyses on how the various components of an MLIA system (stemmers, IR models, relevance feedback, translation techniques) behave with respect to languages;
- the organization of evaluation exercises modeled on the results of MLIA user profiling studies;
- transfer of the research results to the relevant applications.

In this way, we feel that we will be able to exploit to the full the considerable volume of data and knowhow which has been built up in the last decade thanks to the efforts of the approximately 180 research groups that have participated in CLEF so far.

Acknowledgements

CLEF is organised on a distributed basis with different research groups being responsible for the running of the various tracks. We should like to express out gratitude to all those who have been involved in the coordination of this activity over the years. The authors would also like to thank Giorgio Maria Di Nunzio for his generous assistance in the preparation of this paper. The work reported has been partially supported by the TrebleCLEF Coordination Action, FP7 ICT programme for Digital Libraries and Technology-enhanced Learning. Grant agreement: 215231.

References

[1] M. Agosti, G. M. Di Nunzio, and N. Ferro. A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns. In T. Sakay, M. Sanderson, and D. K. Evans, editors, *Proc. 1st International Workshop on Evaluating Information Access (EVIA 2007)*, pages 62–73. National Institute of Informatics, Tokyo, Japan, 2007.

[2] M. Braschler. Combination Approaches for Multilingual Text Retrieval. *Information Retrieval*, 7(1/2):183–204, 2004.

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

[4] G. M. Di Nunzio and N. Ferro. DIRECT: a System for Evaluating Information Access Components of Digital Libraries. In A. Rauber, S. Christodoulakis, and A. Min Tjoa, editors, *Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, pages 483–484. Lecture Notes in Computer Science (LNCS) 3652, Springer, Heidelberg, Germany, 2005.

[5] D. Harman and C. Buckley. The NRRC Reliable Information Access (RIA) Workshop. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 528–529. ACM Press, New York, USA, 2004.

[6] D. K. Harman, M. Braschler, M. Hess, M. Kluck, C. Peters, P. Schaübie, and P. Sheridan. CLIR Evaluation at TREC. In Peters [8], pages 7–23.

[7] H. Müller, J. Kalpathy-Cramer, . Kahn, C.E., W. Hatt, S. Bedrick, and W. Hersh. Overview of the ImageCLEFmed 2008 Medical image Retrieval Task. In F. Borri, A. Nardi, and C. Peters, editors, *Working Notes for the CLEF 2008 Workshop*. `http://www.clef-campaign.org/2008/working_notes/GeoCLEF-2008-overview-notebook-paperWNfinal.pdf` [last visited 2008, September 10], 2008.

[8] C. Peters, editor. *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum (CLEF 2000)*. Lecture Notes in Computer Science (LNCS) 2069, Springer, Heidelberg, Germany, 2001.

[9] C. Peters. Introduction. In *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum (CLEF 2000)* [8], pages 1–6.

[10] S. E. Robertson. The methodology of information retrieval experiment. In K. Spärck Jones, editor, *Information Retrieval Experiment*, pages 9–31. Butterworths, London, United Kingdom, 1981.

[11] J. Savoy. Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross–Language Evaluation Forum (CLEF 2001) Revised Papers*, pages 27–43. Lecture Notes in Computer Science (LNCS) 2406, Springer, Heidelberg, Germany, 2002.