CLEF Ad-hoc: A Perspective on 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 – 56124 Pisa – Italy
carol.peters@isti.cnr.it

Abstract. *MultiLingual Information Access (MLIA)* is a key topic in Digital Libraries. In the last decade the *Cross-Language Evaluation Forum (CLEF)* has stimulated system research and development in this field through the organization of large-scale evaluation campaigns. We discuss the achievements of CLEF in the light of the evolution of one of its main tasks, the Ad Hoc track, which studies multilingual document retrieval, and include details of experimentation with collections provided by The European Library.

1 Introduction

MultiLingual Information Access (MLIA) is a key topic in Digital Libraries. This is the reason why the European Commission, and more specifically the unit for Digital Libraries has sponsored the activity of the *Cross-Language Evaluation Forum (CLEF)*[3] over the last decade. CLEF is the major evaluation initiative for the experimentation and testing of MLIA systems operating on European languages.

This paper shows how CLEF has stimulated system research and development in this field by focusing on the evolution of one of its core tasks, the Ad Hoc track, which studies techniques for multilingual document retrieval on document collections in multiple languages and on different genres: news data and library catalog cards from the archives of The European Library. Our efforts over the years have resulted in a wide coverage of the building blocks for multilingual system development (e.g. tools, components, resources and lexicons).. We feel that it is now time to shift the focus of our activity to acquiring a deeper understanding of the underlying issues.

This change in direction is helped by the launching in 2008 by the European Commission of the TrebleCLEF Coordination Action[4] which intends to promote research, development, implementation and industrial take-up of multilingual, multimodal information access functionality [1]:

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

[4] See http://www.trebleclef.eu/

- by continuing to support the annual CLEF system evaluation campaigns with tracks and tasks designed to meet the requirements of the user and application communities;
- by constituting a scientific forum for the MLIA community of researchers enabling them to meet and discuss results and new directions;
- by providing a central reference point for anyone interested in studying or implementing MLIA functionality.

In the following sections we briefly introduce the CLEF evaluation campaigns, describe the evolution of the Ad-hoc track, and provide an outlook for the future.

2 The CLEF Evaluation Campaigns

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[5]. 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 [2]. 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.

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 *NII-NACSIS Test Collection for IR Systems (NTCIR)*[6] 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.

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 (i.e. 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. Most tracks offer several different tasks and these tasks normally vary each year,

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

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

according to the interests of the track coordinators[7] and participants. Figure 1 shows when tracks have been introduced and when they have been terminated.

3 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 through the exploration of a comprehensive set of CLIR-related topics:

- **experimental collections**: test collections are built for as many European languages as possible, attempting to cover diverse language typologies;
- **tasks**: groups are 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., is supported;
- **linguistic components**: the development and/or application of linguistic tools, such as stemmers, lemmatizers, decompounders, part of speech taggers, and so on, is fostered;
- **translation approaches**: groups are 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 are 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, are 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 are developed and employed.

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. As can be seen from Table 1, the different ad-hoc tasks present varying degrees of difficulty: there are more basic tasks, such as the monolingual tasks or the bilingual English tasks,

[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/`.

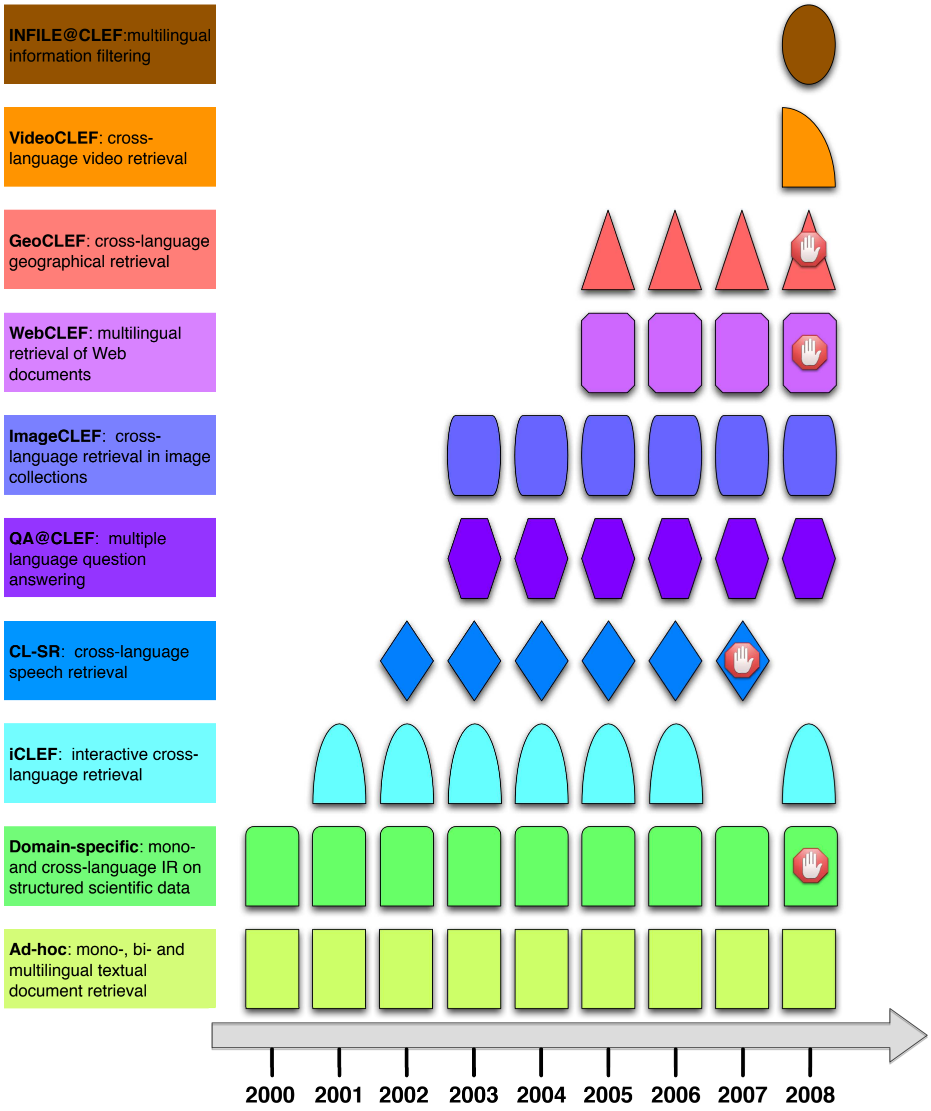
INFILE@CLEF:multilingual information filtering
VideoCLEF: cross-language video retrieval
GeoCLEF: cross-language geographical retrieval
WebCLEF: multilingual retrieval of Web documents
ImageCLEF: cross-language retrieval in image collections
QA@CLEF: multiple language question answering
CL-SR: cross-language 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

Fig. 1. CLEF 2000 – 2008 tracks. Full details of the activities and results of each track can be found on the CLEF website at `http://www.clef-campaign.org/`.

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	

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.

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. [3] 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.

In 2008 there was a big change in focus in this track and we started 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. To this end, 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)[8] 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 catalog 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.

The TEL task is an example of how the ad-hoc track has started to explore finer-grained questions in the CLIR scenario. Indeed the sparsity of the data and their intrinsic multilingualigy 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 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 impor-

[8] See `http://www.theeuropeanlibrary.org/`

tance. This task focuses on the creation of new experimental collections to develop both new linguistic resources (lexicons, dictionaries, and so on) and new linguistic components (stop lists, stemmers, part of speech taggers, and so on). From a retrieval point of view, the necessary techniques are well-known and it is not expected that participants produce new IR components. In this case, we could say that this is a **CL**IR task.

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. This task 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. As in the previous case, 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. Again, In this case, we define this as a **CL**IR task.

This deeper investigation will be taken a step further in the Grid@CLEF Pilot task[9] which has been proposed for CLEF 2009 with the following goals in mind [4]:

- 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.

Indeed, 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. 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 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 result of testing participant A stemmer + participant B probabilistic IR model.

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

The Pilot Grid task in CLEF 2009 will provide a framework for a first set of experiments which will allow us 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.

4 Conclusions

In this paper, we have discussed the evolution of the CLEF by focussing on the activities carried out in the Ad Hoc track. 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?" since CLEF began when CLIR was just starting to be recognized as an independent sub-discipline and thus promoted much pioneering work in the field.

Now, we are in the position of conducting a much deeper investigation of the core issues of the field and we will focus on:

- in-depth analyses on how the various components of MLIA systems (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.

Acknowledgements

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. Braschler, M., Di Nunzio, G.M., Ferro, N., Gonzalo, J., Peters, C., Sanderson, M.: From CLEF to TrebleCLEF: promoting Technology Transfer for Multilingual Information Retrieval. In Second DELOS Conference - Working Notes (2007)
2. Harman, D.K., Braschler, M., Hess, M., Kluck, M., Peters, C., Schaüble, P., Sheridan, P.: CLIR Evaluation at TREC. In Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum (CLEF 2000), LNCS 2069, Springer (2001) 7–23
3. Braschler, M.: Combination Approaches for Multilingual Text Retrieval. Information Retrieval **7**(1/2) (2004) 183–204
4. Ferro, N., Harman, D.: Dealing with MultiLingual Information Access: Grid Experiments at TrebleCLEF. In Post-proceedings of the Fourth Italian Research Conference on Digital Library Systems (IRCDL 2008), ISTI-CNR at Gruppo ALI (2008) 29–32