Information Retrieval, 7, 7–31, 2004
© 2004 Kluwer Academic Publishers. Manufactured in The Netherlands.

Cross-Language Evaluation Forum: Objectives, Results, Achievements

MARTIN BRASCHLER martin.braschler@eurospider.com
Eurospider Information Technology AG, Schaffhauserstr. 18, 8006 Zürich, Switzerland; Université de Neuchâtel, Institut interfacultaire d'informatique, Pierre-à-Mazel 7, CH-2001 Neuchâtel, Switzerland

CAROL PETERS carol.peters@isti.pi.cnr.it
ISTI-CNR, Area di Ricerca, 56124 Pisa, Italy

Received August 26, 2003; Revised August 26, 2003; Accepted September 3, 2003

Abstract. The Cross-Language Evaluation Forum (CLEF) is now in its fourth year of activity. We summarize the main lessons learned during this period, outline the state-of-the-art of the research reported in the CLEF experiments and discuss the contribution that this initiative has made to research and development in the multilingual information access domain. We also make proposals for future directions in system evaluation aimed at meeting emerging needs.

Keywords: evaluation campaign, cross-language information retrieval, evaluation methodology, CLIR state-of-the-art

1. Introduction

The Cross-Language Evaluation Forum is in its fourth year of activity – the results of the CLEF 2003 campaign have now been judged and were presented in the annual workshop in August. However, the history of CLEF can be traced back to 1997 with the first track for the evaluation of cross-language systems at the Text REtrieval Conference (TREC) (Schäuble and Sheridan 1998). We are thus now in the position to attempt an assessment of the results achieved in nearly seven years of activity.

The declared high-level objectives of CLEF are threefold:

1. to provide an infrastructure for the testing and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts,
2. to construct test-suites of reusable data that can be employed by system developers for benchmarking purposes,
3. to create an R&D community in the cross-language information retrieval (CLIR) sector.

The aim of this paper will be to describe how CLEF has worked towards realizing these objectives, what has been achieved so far, and what lessons we feel have been learnt from this experience. It is often claimed (see e.g. Smeaton and Harman 1997) that evaluation campaigns really do play an active part in advancing system development. A question we will thus be asking is what has been the impact of CLEF on cross-language system development in this period. However, before we begin to assess this impact, we start by

examining the status of R&D in this sector and the relationship between the research world and the application communities.

Although first experiments with cross-language information retrieval date back to the early 70s (Salton 1970), recognition of CLIR as a separate area of research probably began at SIGIR 1996 with the organization of a Workshop on "Cross-Linguistic Information Retrieval" (Grefenstette 1998); several other activities and events followed on closely—showing that there was much interest in the research community in this "new" area. The growing popularity of the Internet and the consequent wide availability of (free) networked information sources for an increasingly vast public have clearly fuelled academic interest in CLIR. The World Wide Web phenomenon, coupled with increasing globalization of corporations and organizations, has led to strong demand for tools that permit the user to find information wherever and however it is stored, regardless of language boundaries. The demand has also been stimulated by more non-English-speaking end users going online and corporations finding themselves competing in a world-wide marketplace driven by foreign language information. There would thus appear to be a strong potential for effective and efficient systems that allow users to search document collections in multiple languages and retrieve relevant information in a form that is useful to them, even when they have little or no linguistic competence in the target languages. However, such systems are not easy to develop and although there has been much work on system development since 1996, as evidenced by the papers in this special issue, as yet there has been no strong commercial take-up. The intention of this paper is to summarize the lessons learnt from the first three CLEF evaluation campaigns (CLEF 2000–CLEF 2002), the results of the fourth still under analysis at the time of writing, and to address the question of what remains to be done in order to bridge the gap between research and application, i.e. between system developer and system user. The paper also provides the necessary background information for a full understanding of the experiments reported in the other papers included in this special issue dedicated to CLEF.

The rest of this paper is organised as follows. Section 2 describes the background to CLEF, from its origins to the present, whereas Section 3 presents details on the evaluation methodology, the test collections, and the techniques used for results calculation and analysis. In Section 4, we analyse the main findings from this series of campaigns and investigate whether CLEF has a direct impact on the development of the systems of the groups that participate. In the final section, we summarise our achievements so far and outline our ideas for future directions.

2. Seven years of activity

2.1. CLIR at TREC—1997–1999

The surge of interest in the CLIR research area after 1996 led to the organization of a first track for CLIR system evaluation in 1997 at the TREC-6 conference (Schäuble and Sheridan 1998). The following year, the NTCIR Workshop began cross-language evaluation for systems working on Asian languages (Kando 2003).

In the first year of the CLIR track at TREC, the participating groups performed a bilingual task—searching target document collections in French, German or English with queries

formulated in a different language. Results were to be returned in the form of lists of document identifiers, ranked by decreasing probability of their relevance to the query. Thirteen groups took part, using all possible combinations of source and target languages, with the consequence that it was very difficult to compare results over different systems. The track took a big step forward in the following year with the introduction of a new task in which the systems had to search a multilingual collection of documents in four languages—the new language was Italian—for relevant items, using a single query language. Of the nine groups participating in the CLIR track in TREC-7, five groups tried this task. Its introduction forced developers to study the most appropriate indexing, transfer and retrieval strategies when handling collections in multiple languages simultaneously, ranking results across collections and languages in a single list—a complex exercise (Braschler et al. 1999). The multilingual task was repeated in 1999 with eight groups attempting it. Bilingual tasks continued to be offered as secondary exercises. An overview of the CLIR tracks at TREC is given in Harman et al. (2001).

The experience at TREC showed that an activity of this type, which involved handling, processing and understanding text in many languages, needed the collaboration of native speakers of the languages involved. For this reason, from 1998 on, the CLIR track at TREC was organised on a distributed basis, with sites in different countries being responsible for handling the work on the document collections in each language. At the end of 1999, it was decided that it would be more appropriate to centre the coordination of the evaluation activity for European languages in Europe while TREC would shift its attention to other language groups (Chinese in 2000, Arabic in 2001 and 2002). TREC, CLEF and NTCIR agreed to coordinate their activities and their schedule to avoid, as far as possible, overlapping of dates, in order to facilitate groups that wished to participate in more than one of the cross-language campaigns.

2.2. Cross-language system evaluation moves to Europe

The move to Europe and the launching of an independent project, known as the Cross-Language Evaluation Forum (CLEF), has made it possible to build on and extend the results achieved within TREC. The multilingual environment provided by Europe has facilitated the addition of new languages and has stimulated participation. European businesses and institutions are accustomed to working in a multilingual context and need tools that help them to do this. The European Union currently recognises 13 official languages—this number is expected to become 25 in 2004 with the addition of ten new member states (Pieters 2002). It is one of the guiding policies of the EU that linguistic and cultural diversity should be safeguarded and access to knowledge should be guaranteed in all the languages of the Union. Support for CLEF was thus sought from the European Commission on the grounds that an initiative that involves an increasing number of European languages should be organised on the European rather than the national level.[1]

The first campaigns of CLEF had the main goals:

- to accommodate as many European languages as possible;
- to encourage the participation of European groups (disappointingly low during the three years of activity coordinated in the US);

Table 1. Growth in number of participants and experiments over the years.

Year	# Participants (for all tracks)	# Participants (core tracks only)	# Experiments (core tracks only)
TREC-6 (1997)	13	13	(95)*
TREC-7 (1998)	9	9	27
TREC-8 (1999)	12	12	45
CLEF 2000	20	20	95
CLEF 2001	34	31	198
CLEF 2002	37	34	282
CLEF 2003	42	33	415

*In TREC-6, only bilingual retrieval was offered, which resulted in a large number of runs combining different pairs of languages (Schäuble and Sheridan 1998). Starting with TREC-7, multilingual runs were introduced (Braschler et al. 1999), which usually consist of multiple runs for the individual languages that are merged. The number of experiments for TREC-6 is therefore not directly comparable to later years.

- to provide facilities for monolingual system testing and tuning in European languages other than English, which was already well covered by TREC;
- to stimulate systems to move from monolingual searching to the implementation of a full multilingual retrieval service;
- to study the needs of both system developers and system users in order to promote the introduction of new tasks, designed to meet newly identified requirements.

The results have been encouraging from the start. Separate tracks to test monolingual, bilingual and multilingual systems were provided with the aim of allowing groups to work their way up gradually from mono- to multilingual retrieval. Additional tracks have been added to supplement the core tracks. The test collection has continued to grow. Participation of both academic and industrial groups, and especially of European groups, has increased rapidly (see Table 1). In the following section, we describe the organization of these evaluation campaigns and explain the underlying motivations for certain choices and decisions.

3. Organization of the CLEF evaluation campaign

CLEF aims at providing a forum that is not so much a competition, but rather a place where people can objectively evaluate their systems and ideas. It is left open to the participants whether they want to perfect a tried-and-true approach, or try "revolutionary" new ideas, even though some of these may quickly disappear. In this sense, the intentions of participants, and the amount of effort they invest in their CLEF experiments differ vastly. The campaign as a whole benefits from this fact, as newcomers and creative, "daring" groups find as much a place for their work as veteran participants that have built elaborate, finely tuned systems.

3.1. Comparative evaluation and the Cranfield paradigm

For the core activities, CLEF adopts a corpus-based, automatic scoring method, based on ideas first introduced in the Cranfield experiments (Cleverdon 1997) in the late 1960s. This methodology is widely used and accepted in the information retrieval community. Its properties have been thoroughly investigated and are well understood. This approach is also used by the popular series of TREC conferences (Harman 1995), which are the "gold standard" for this form of evaluation campaigns. "End-users" are not directly involved in the evaluation when following this "Cranfield" paradigm. While user-based evaluation aims to directly measure the user's satisfaction with a particular system, the CLEF core activities have been limited to system evaluation. One of the reasons for adopting system evaluation lies in the costs and complexities incurred by conducting user-based evaluations. While much is to be said for involving end users in the evaluation of systems (and the "interactive track" of CLEF indeed tries to work in this direction), besides significantly lowering the cost of conducting the evaluation, abstracting the evaluation process can help to control some of the parameters that affect retrieval performance, and thus may increase the power of comparative experiments. In CLEF, following the Cranfield paradigm, good system performance is equated with good retrieval effectiveness (in terms of returning lists of documents). There are a host of assumptions underlying the "laboratory setting" used for system evaluation in CLEF, and we will briefly discuss some of their implications. For a more detailed discussion of the Cranfield paradigm see (Voorhees 2002).

The CLEF campaigns use a combination of a set of retrievable documents, a set of formulations of "information needs" and measures for evaluating the effectiveness of the system in answering these information needs on the basis of the set of documents. The measures used by CLEF, mainly recall and precision,[2] are based on the "relevance" of a document to the corresponding information need, which is defined on topical similarity as determined by expert relevance assessors. For the sake of practicality, CLEF must make multiple assumptions with regard to the concept of relevance it adopts: first, that all "relevant documents" contribute equally to the performance measures, second, the relevance of a document is independent of the relevance of other documents, and third, that all potential users agree on the relevance of a document with respect to an information need. For some performance measures, such as recall, it is further assumed that all relevant documents in the collection are known. Relevance assessments were chosen to be binary: a document is either relevant or irrelevant. CLEF emphasizes comparative evaluation. An important consequence of the abstractions used in this methodology is that absolute scores of individual experiments are not meaningful in isolation.

Of course, the distinguishing feature of CLEF is that it applies this evaluation paradigm in a multilingual setting. This means that the criteria normally adopted to create a test collection consisting of suitable documents, sample queries and relevance assessments must be adapted to satisfy the particular requirements of this context. In the rest of this section, we will thus discuss how the CLEF evaluation campaigns have been designed and set up to meet the special needs of multilingual information retrieval. In Section 3.2 we describe how the evaluation tasks have been studied to meet the needs of the multilingual system developers community, and in Section 3.3 we describe the measures taken to create a

test collection that can be used in a multilingual, multicultural context and can be trusted to give consistent, unbiased results. This is particularly important with respect to the relevance assessments whose objectivity must be guaranteed. Finally, in Section 3.4, we comment on the completeness of the relevance assessments as used in CLEF.

3.2. Evaluation tracks

Over the years, the range of activities offered to participants in the initial TREC CLIR and subsequent CLEF campaigns has expanded and been modified in order to meet the needs of the research community. Consequently, the campaigns are structured into several distinct tracks. Some of these tracks are in turn structured into multiple tasks. Almost from the very beginnings in TREC, the main focus of the campaigns has been the multilingual retrieval track, in which systems must use queries in one language to retrieve items from a test collection that contains documents written in a number of different languages (four in CLEF 2000, five for CLEF 2001 and 2002, and either four or eight in CLEF 2003). Participants are actively encouraged to work on this, the hardest task offered. A stated goal of the CLEF campaign is to allow groups to gradually move from "easier" tracks/tasks in their first participation to eventually work with as many languages as possible, joining the multilingual track. To this end, bilingual and monolingual tracks are also offered. These smaller tracks serve additional important purposes, for example in terms of helping to better understand the characteristics of individual languages, and to fine-tune procedures.

In CLEF, we distinguish between the *core tracks*, which are those that are offered regularly each year (the monolingual, bilingual, multilingual and domain-specific tracks) and *additional tracks*, which tend to be organized on a more experimental basis and have the objective of identifying new requirements and appropriate methodology for their testing. The core tracks are coordinated by the members of the CLEF consortium, whereas the additional tracks are organized by interested associated groups, under the CLEF umbrella, on a voluntary basis. Here below we describe the tracks and tasks offered by the campaigns from CLEF 2000 through CLEF 2003. For each of the core tracks, the participating systems construct their queries (automatically or manually) from a common set of statements of information needs (known as topics) and search for relevant documents in the collections provided, listing the results in a ranked list. With the exception of the paper by Oard et al. which discusses work done in the Interactive CLEF track, the articles in this special issue describe experiments on the core tracks.

CORE TRACKS

Multilingual Information Retrieval. This is the main task in CLEF. It requires searching a multilingual collection of documents for relevant items, using a selected query language. Multilingual information retrieval is a complex task, testing the capability of a system to handle a number of different languages simultaneously, ordering them according to relevance. In CLEF 2000, the multilingual collection for this track contained English, German, French, and Italian documents. In CLEF 2001 and 2002, it also contained Spanish

texts. For CLEF 2003, two distinct multilingual tasks were offered: multilingual-4 and multilingual-8. The collection for multilingual-4 contained English, French, German and Spanish documents. Multilingual-8 involved searching a collection containing documents in eight languages: Dutch, English, Finnish, French, German, Italian, Spanish and Swedish. For each campaign, a common set of topics (i.e. structured statements of information needs from which queries are extracted) has been prepared in up to twelve languages: Dutch, English, Finnish, French, German, Italian, Spanish, Swedish, Russian, Portuguese, Japanese and Chinese. Topics have been offered in other languages on demand.

Bilingual Information Retrieval. In this track, any query language can be used to search a single target document collection. Many newcomers to CLIR system evaluation prefer to begin with the simpler bilingual track before moving on to tackle the more complex issues involved in truly multilingual retrieval. CLEF 2000 offered the possibility for a bilingual search on an English target collection, using any other language for the queries. CLEF 2001 offered two distinct bilingual tracks with either English or Dutch target collections. In response to considerable pressure from the participants, in CLEF 2002 we decided to extend the choice to all of the target document collections, with the single limitation that only newcomers to a CLEF cross-language evaluation task could use the English target document collection. This decision had the advantage of encouraging experienced groups to experiment with "different" target collections, rather than concentrating on English, but it had the strong disadvantage that the results were harder to assess in a comparative evaluation framework. There were simply too many topic-target language combinations, only receiving a few experiments each. Consequently, for CLEF 2003, we offered a very different choice. The main objective in the 2003 bilingual track was to encourage the tuning of systems running on challenging language pairs that do not include English, but also to ensure comparability of results. For this reason, runs were only accepted for one or more of the following source → target languages pairs: Italian → Spanish, German → Italian, French → Dutch, Finnish → German. Newcomers only (i.e. groups that have not previously participated in a CLEF cross-language task) could choose to search the English document collection using a European topic language. At the last moment, we acquired a Russian collection and thus also included Russian as a target collection in the bilingual task, permitting any language to be used for the queries.

Monolingual (non-English) IR. Until recently, most IR system evaluation focused on English. However, many of the issues involved in IR are language dependent. CLEF provides the opportunity for monolingual system testing and tuning, and for building test suites in other European languages apart from English. Each of the CLEF campaigns has thus provided the opportunity for monolingual system testing and tuning on any of the target collections available, with the exception of the English collection.

Domain-Specific Mono- and Cross-Language Information Retrieval. The rationale for this task is to study CLIR on other types of collections, serving a different kind of information need. The information that is provided by domain-specific scientific documents is far more targeted than news stories and contains much terminology. It is claimed that the users of this type of collection are typically interested in the completeness of results. This means that they are generally not satisfied with finding just some relevant documents in a collection that

may contain many more. Developers of domain-specific cross-language retrieval systems need to be able to tune their systems to meet this requirement. See (Kluck and Gey 2001) for a discussion of this point.

ADDITIONAL TRACKS

Interactive CLIR (iCLEF). The aim of the tracks listed so far is to measure system performance mainly in terms of its effectiveness in document ranking. However, this is not the only issue that interests the user. User satisfaction with an IR system is based on a number of factors, depending on the functionality of the particular system. For example, the ways in which a system can help the user when formulating a query or the ways in which the results of a search are presented are of great importance in CLIR systems where it is common to have users retrieving documents in languages with which they are not familiar. An interactive track that has focused on both user-assisted query formulation and document selection has been implemented with success since CLEF 2001 (see Oard et al. 2004).

Multilingual Question Answering (QA at CLEF). This is a completely new track introduced for the first time at CLEF 2003. It consists of several tasks and offers the possibility to test monolingual question answering systems running on Spanish, Dutch and Italian texts, and cross-language systems using questions in Dutch, French, German, Italian and Spanish to search an English document collection. This track is an important innovation for CLEF as successful question answering systems need to integrate IR technology together with sophisticated natural language processing (NLP) procedures. The aim of this track is both to stimulate monolingual work in the question answering area on languages other than English and to encourage the development of the first experimental systems for cross-language QA.

Cross-Language Spoken Document Retrieval (CL-SDR). The current growth of multilingual digital material in a combination of different media (e.g. image, speech, video) means that there is an increasing interest in systems capable of automatically accessing the information available in these archives. For this reason, the DELOS Network of Excellence for Digital Libraries[3] supported a preliminary investigation aimed at evaluating systems for cross-language spoken document retrieval in 2002. The aim was to establish baseline performance levels and to identify those areas where future research was needed. The results of this pilot investigation were first presented at the CLEF 2002 Workshop and are reported in Jones and Federico (2003). Cross-language spoken document retrieval has been offered as a pilot experiment in CLEF 2003.

Cross-Language Retrieval in Image Collections (Image CLEF). This track was offered for the first time in CLEF 2003 as a pilot experiment. The aim is to test the effectiveness of systems to retrieve as many relevant images as possible on the basis of an information need expressed in a language different from that of the document collection. Queries were made available in five languages (Dutch, French, German, Italian and Spanish) to search a British-English image collection. Searches could make use of the image content, the text captions or both.

These last two tracks show the effort that is now being made by CLEF to progress from text retrieval tasks to tasks that embrace multimedia.

3.3. The test collections

The CLEF test collection for the core tracks is formed of sets of documents in different European languages but with common features (same genre and time period, comparable content); a single set of topics rendered in a number of languages; relevance judgments determining the set of relevant documents for each topic. A separate test collection has been created for systems tuned for domain-specific tasks.

3.3.1. Document collections. The main document collection now consists of well over 1.5 million documents in nine languages—Dutch, English, Finnish, French, German, Italian, Russian, Spanish and Swedish. This collection has been expanded gradually over the years. The 2000 collection consisted of newspaper, news magazine and news agency articles mainly from 1994 in four languages: English, French, German and Italian. Two languages were added in 2001: Spanish because of its global importance, and Dutch, partly to meet the demands of a considerable number of Dutch participants—a very active community in early CLEF, but also because it provided a challenge for those who wanted to test the adaptability of their systems to a new, less widespread language. Swedish and Finnish were introduced for the first time in CLEF 2002 for different reasons. Swedish was chosen as a representative of the Nordic languages, whereas Finnish was included both because it was a representative of a different language group (the Uralic languages) and also because its complex morphology makes it a particularly challenging language from the text processing viewpoint. Russian was an important addition in 2003 as it is the first collection in the CLEF corpus that does not use the Latin-1 (ISO-8859-1) encoding system.

The domain-specific collection consists of the GIRT database of German social science documents, with controlled vocabularies for English-German and German-Russian. The GIRT texts were first used in the TREC CLIR tracks and have been expanded for CLEF. In 2003 a third, even more extensive version of the GIRT database was introduced, consisting of more than 150,000 documents in an English-German parallel corpus**.** In CLEF 2002, this track also used the Amaryllis scientific database of approximately 150,000 French bibliographic documents, and a controlled vocabulary in English and French.

Table 2 gives details of the source and dimensions of the main multilingual document collection used in CLEF. Most papers in this special issue make reference to the collection as of 2002. The table gives the overall size of each subcollection, the year of its addition, number of documents contained, and three key figures indicating some typical characteristics of the individual documents: the median length in bytes, tokens and features. Tokens are "word" occurrences, extracted by removing all formatting, tagging and punctuation, and the length in terms of features is defined as the number of distinct tokens occurring in a document.

3.3.2. Topics. For the core tasks, the participating groups derive their queries in their preferred language from a set of topics that were created to simulate user information needs. Following the TREC philosophy, each topic consists of three parts: a brief title statement; a one-sentence description; a more complex narrative specifying the relevance assessment

Table 2. Sources and dimensions of the main CLEF document collection.

Collection	Added in	Size (MB)	No. of docs	Median size of docs. (Bytes)	Median size of docs. (Tokens)[a]	Median size of docs. (Features)
Dutch: Algemeen Dagblad 94/95	2001	241	106483	1282	166	112
Dutch: NRC Handelsblad 94/95	2001	299	84121	2153	354	203
English: LA Times 94	2000	425	113005	2204	421	246
English: Glasgow Herald 95	2003	154	56472	2219	343	202
Finnish: Aamulehti late 94/95	2002	137	55344	1712	217	150
French: Le Monde 94	2000	158	44013	1994	361	213
French: ATS 94	2001	86	43178	1683	227	137
French: ATS 95	2003	88	42615	1715	234	140
German: Frankfurter Rundschau 94	2000	320	139715	1598	225	161
German: Der Spiegel 94/95	2000	63	13979	1324	213	160
German: SDA 94	2001	144	71677	1672	186	131
German: SDA 95	2003	144	69438	1693	188	132
Italian: La Stampa 94	2000	193	58051	1915	435	268
Italian: AGZ94	2001	86	50527	1454	187	129
Italian: AGZ 95	2003	85	48980	1474	192	132
Russian: Izvestia 95[b]	2003	68	16761			
Spanish: EFE 94	2001	511	215738	2172	290	171
Spanish: EFE 95	2003	577	238307	2221	299	175
Swedish: TT 94/95	2002	352	142819	2171	183	121

SDA/ATS/AGZ: Schweizerische Depeschenagentur (Swiss News Agency), EFE: Agencia EFE S.A (Spanish News Agency), TT: Tidningarnas Telegrambyrå (Swedish News Agency).

[a]The number of tokens extracted from each document can vary slightly across systems, depending on the respective definition of what constitutes a token. Consequently, the number of tokens and features given in this table are approximations and may differ from actual implemented systems.

[b]Figures for Russian are not comparable due to a different encoding system.

criteria. The title contains the main keywords, the description is a "natural language" expression of the concept conveyed by the keywords, and the narrative adds additional syntax and semantics, stipulating the conditions for relevance assessment. Queries can be constructed from one or more fields. Here below we give the English version of a typical topic from CLEF 2002.

<top>
<num> C091 </num>
<EN-title> AI in Latin America </EN-title>
<EN-desc> Amnesty International reports on human rights in Latin America. </EN-desc>
<EN-narr> Relevant documents should inform readers about Amnesty International reports regarding human rights in Latin America, or on reactions to these reports. </EN-narr>
</top>

The motivation behind using structured topics is to simulate query "input" for a range of different IR applications, ranging from very short to elaborate query formulations, and representing keyword-style input as well as natural language formulations. The latter potentially allows sophisticated systems to make use of morphological analysis, parsing, query expansion and similar features. In the cross-language context, the transfer component must also be considered, whether dictionary or corpus-based, a fully-fledged MT system or other. Different query structures may be more appropriate for testing one or the other methodology.

The creation of a topic set in a multilingual context necessitates a very rigorous procedure in order to ensure consistency and coherency of the topic sets in the different languages. CLEF topics are developed on the basis of the contents of the multilingual document collection. For each language, native speakers propose a set of topics covering events of local, European and general importance. The topics are then compared over the different sites to ensure that a high percentage of them will find some relevant documents in all collections, although the ratio can vary considerably. The fact that the same topics are used for the mono-, bi-, and multilingual tracks is a significant constraint.[4] While in the multilingual task it is of little importance if a given topic does not find relevant documents in all of the collections, in both the bilingual and monolingual tracks, where there is a single target collection, a significant number of the queries must retrieve relevant documents.

Other criteria must be met to provide a full range of cross-language testing possibilities for the participating systems. For example, it is important to include names of locations (translatable or not), of people (where some kind of robust matching may be necessary, e.g. Eltsin, Ieltsin, or Yeltsin), important acronyms, some terminology (testing lexical coverage), syntactic and semantic equivalents. The goal is to achieve a natural, balanced topic set accurately reflecting real world user needs while at the same time testing a system's processing capabilities to the full.

Once the topics have been selected, they are prepared in all the collection languages by skilled translators translating into their native language. They can then be translated into additional languages, depending on the demand from the participating systems. In all cases, the aim is to produce natural language renderings of the concepts expressed rather than literal translations.

The main CLEF topic set currently consists of 200 topics for eight different languages (Dutch, English, Finnish, French, German, Italian, Spanish, Swedish), and a subset of them for additional topic languages (Chinese, Japanese, Portuguese, Russian, Thai). 40 topics were created for CLEF 2000, 50 topics each for 2001 and 2002, and 60 for CLEF 2003. Separate topic sets have been developed for the GIRT task in German, English and Russian and in French and English for Amaryllis, The CLEF topic generation process and the issues involved are described in detail in Womser-Hacker (2002) and Mandl and Womser-Hacker (2003).

The size of the topic set in each campaign is dictated by limited evaluation resources. With the kind of effort that is possible within the CLEF campaigns, it is impractical to produce exhaustive relevance assessments for larger topic sets. Using such a topic set as a (small) sample of all potential information needs that end users might have has implications on the interpretation of the experimental results. Indeed, statistical analysis of the results in the multilingual tracks of the 2001 and 2002 campaigns has shown that performance changes

need to be rather large to find statistically significant differences between experiments (Braschler 2002, 2003). The best remedy is to use a larger topic set for experiments, which CLEF facilitates by keeping portions of the document collection unchanged from campaign to campaign, effectively allowing post-campaign experiments to use multiple topic sets from several campaigns. A recent discussion of the influences of different topic set sizes can be found in Voorhees and Buckley (2002).

3.3.3. Relevance judgments. The relevance assessments are produced in the same distributed setting and by the same groups that work on the topic creation. CLEF uses methods adapted from TREC to ensure a high degree of consistency in the relevance judgments. All assessors follow the same criteria when judging the documents. An accurate assessment of relevance of retrieved documents for a given topic implies a good understanding of the topic. This is much harder to achieve in the distributed scenario of CLEF where understanding is influenced by language and cultural factors. Rules are established to ensure, as far as possible, that the decisions taken as to relevance are consistent over sites, and over languages.

The practice of assessing the results on the basis of the "Narrative" means that only using the "Title" and/or "Description" parts of the topic implicitly assumes a particular interpretation of the user's information need that is not (explicitly) contained in the actual query run in the experiment. The fact that the information contained in the title and description fields could have additional possible interpretations has influence only on the absolute values of the evaluation measures, which in general are inherently difficult to interpret. However, comparative results across systems are usually stable when considering different interpretations. These considerations are important when using the topics to construct very short queries to evaluate a system in a web-style scenario.

The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead, approximate recall figures are calculated by using pooling techniques. The results submitted by the participating groups are used to form a "pool" of documents for each topic and for each language by collecting the highly ranked documents from all the submissions. The implications of using this pooling procedure on the validity of the results published by CLEF are discussed in the following section. Table 3 gives an overview of the number of topics and documents used in the campaigns for the core tracks, and of the size of the corresponding document pools.

Table 3. Number of documents assessed for the CLEF campaigns (core tracks).

Collection	# Lang.	# Docs.	Size in MB	# Docs. assessed	# Topic.	# Assessed/topic
CLEF 2003	9	1,611,178	4124	∼188,000	60*	∼3100
CLEF 2002	8	1,138,650	3011	140,043	50*	∼2900
CLEF 2001	6	940,487	2522	97,398	50	1948
CLEF 2000	4	368,763	1158	43,566	40	1089

*Only 30 topics assessed for Finnish in 2002 and only 37 topics assessed for Russian in 2003.

3.4. Results analysis

As was mentioned when introducing the Cranfield paradigm used by the CLEF campaigns in Section 3.2, performance measures reported for the CLEF experiments depend heavily on relevance assessments. Using a pooling methodology, i.e. only judging documents for relevance that have been highly ranked in at least one of the experiments considered, while essential to make relevance assessment feasible for large collections, incurs the risk that a substantial portion of relevant documents goes undetected, i.e. that the relevance assessments are not sufficiently complete. Since such a situation would cast serious doubts on the validity of the conclusions derived on the basis of the CLEF results, tests investigating pool quality have been run since the first CLEF campaign in 2000. The situation is particularly critical with respect to the reusability of the test collections produced by CLEF: an incomplete pool may put experimenters whose systems did not contribute to the pool during the campaigns at a disadvantage.

We have adopted an idea originally put forward by Zobel (1998): to get an indication of the completeness of the pool, individual participants are in turn removed from the pool, each time re-evaluating results. If results remain stable throughout this process, evidence is gained that the pool is sufficiently complete so that new participants would add little in terms of more relevant documents to the pool. That is, the pool contains the vast majority of relevant documents. We have calculated the mean and maximum difference observed when using reduced relevance assessments for the multilingual document pool for all three campaigns held to date. For 2002, we have also calculated the respective numbers for the subcollections formed by the individual languages. In all cases, we have found that the CLEF collections compare favourably to other test collections produced by similar campaigns, with a maximum difference for the multilingual experiments of 5.99% observed for the 2000 campaign, which has been steadily lowered to 1.76% in 2002 thanks to even larger pools based on more different systems and experiments being used (see Table 4). The mean differences are in the order of 1% or less, meaning that conclusions with respect to one system outperforming another system are only affected in cases where differences are too small to be of any likely statistical significance. For the subcollections formed by individual languages, somewhat larger differences are observed, mainly for the newer languages that have been introduced in the later campaigns (see Table 5). This is to be expected due to fewer participants and systems contributing to the respective pools. Still, the numbers are sufficiently favourable to make it unlikely that conclusions based on the test collections are

Table 4. Multilingual track of the CLEF 2000, CLEF 2001 and CLEF 2002 campaigns. Key values of the pool quality analysis: mean and maximum change in average precision when removing the pool contribution of one participant, and associated standard deviation.

Campaign	Mean difference	Max. difference	Std. dev. difference
CLEF 2000	0.0013 (0.80%)	0.0059 (5.99%)	0.0012 (1.15%)
CLEF 2001	0.0023 (1.02%)	0.0076 (4.50%)	0.0051 (2.37%)
CLEF 2002	0.0008 (0.48%)	0.0030 (1.76%)	0.0018 (1.01%)

Table 5. CLEF 2002, subcollections for individual languages. Key values of the pool quality analysis: mean and maximum change in average precision when removing the pool contribution of one participant, and associated standard deviation.

Track	Mean difference	Max. difference	Std. dev. difference
DE German	0.0025 (0.71%)	0.0095 (5.78%)	0.0054 (1.71%)
EN English	0.0023 (1.14%)	0.0075 (3.60%)	0.0051 (2.60%)
ES Spanish	0.0035 (0.87%)	0.0103 (2.52%)	0.0075 (1.86%)
FI Finnish	0.0021 (0.82%)	0.0100 (4.99%)	0.0049 (2.05%)
FR French	0.0019 (0.54%)	0.0050 (1.86%)	0.0038 (1.08%)
IT Italian	0.0008 (0.22%)	0.0045 (0.93%)	0.0016 (0.46%)
NL Dutch*	0.0045 (1.26%)	0.0409 (9.15%)	0.0116 (3.09%)
SV Swedish	0.0082 (3.32%)	0.0306 (10.19%)	0.0182 (7.51%)

*One experiment that was an extreme outlier in terms of performance was removed before calculation of the Dutch figures to avoid a non-representative skew in the numbers.

invalidated, provided that care is taken in considering the inherent limitations of system evaluations such as those conducted by CLEF.

3.5. Workshops

Each CLEF evaluation campaign culminates in a two-day workshop. The objective of the workshops is to bring together the groups that have participated in that year's campaign so that they can report on the results of their experiments. These workshops are an essential part of the CLEF experience; they provide the opportunity for researchers working on common problems to get together and exchange ideas and opinions which not only regard current approaches and techniques but also future directions for research in this field. Participants also have the chance to make proposals for new tasks to be introduced in future campaigns. The workshops play a strong role in the creation of a CLIR research community around the CLEF activity. It is easy to witness their impact on successive campaigns as we see groups experimenting with approaches they have seen presented in previous years and also sharing tools and resources. Copies of the Working Notes and the presentations given at the workshop are made publicly available on the CLEF website.

4. Current state of CLEF systems

In the three CLEF campaigns that have been completed to date (and in the three CLIR tracks held at the TREC conference before that), numerous ideas and methods have been used by the participants. Some of these methods have quickly faded away after one campaign, whereas others have been eagerly adopted and expanded on by other participants. This sharing of ideas is a very important aspect of the CLEF activities, which we will expand on later in this section. We start with the perhaps rather ambitious goal of describing the blueprint for "a successful CLEF system"—of course any such attempt is by necessity

limited by what has been learnt by the participants in the campaigns that have taken place so far. Even so, the sharing of ideas implies a certain amount of convergence toward systems that use one of the (or the only) blueprints that have proven successful so far. Clearly, such convergence could lead to the risk of a "monoculture" of CLIR systems; this evidences the value of participants that dare to "think outside the box" and try new approaches. Luckily, the CLEF campaign has seen a number of experiments by such participants. We think this may be helped by the fact that CLEF is set up not to be a purely competitive forum, but rather as a place for both introducing and fine-tuning ideas.

4.1. A successful blueprint for a multilingual retrieval system

When analyzing the results of the CLEF 2002 campaign (the CLEF 2003 campaign has not yet concluded), it becomes apparent that there are strong similarities between the systems of the three participants that submitted the top performing experiments for the multilingual retrieval task. Our interpretation of this fact is that through participation in earlier campaigns and learning from the experiences reported from them, these three groups have found a "blueprint" for what constitutes—in terms of the state-of-the-art—a type of CLIR system that is successful for the CLEF task. We will now try to outline our understanding of this blueprint. These three systems have been built by Université de Neuchâtel (Savoy 2004), University of California at Berkeley Group 2 (Chen and Gey 2004) and Eurospider (Braschler 2004) and are each described in articles included in this special issue. All three systems are based on a strong foundation in monolingual retrieval for some or all of the languages in the multilingual track in CLEF 2002 (English, French, German, Italian, Spanish). This does not necessarily mean that an extraordinary amount of linguistic knowledge or language-specific processing is used, but that the methods employed for monolingual retrieval are robust and well-tuned, leading to performance in the monolingual retrieval tracks by these groups which either outperforms other participants or is very close to top performance. All three groups used stemming for all five languages and decompounding for German words. However, they did not use sophisticated morphological analyzers or tools such as part-of-speech taggers. For term weighting, well-known weighting schemes that have previously been shown to be successful in English language evaluations such as TREC were applied (BM25, Lnu.ltn, Berkeley ranking). Blind feedback (i.e. automatic query expansion by terms collected from the documents ranked at the top after initial retrieval) was used by all three groups. This "formula" for monolingual retrieval (robust stemming, well-known weighting schemes, and blind feedback) was only outperformed for Italian by two Italian groups (Amati et al. 2003, Bertoldi and Federico 2003), which may hint at potential for some more language-specific fine-tuning.

In terms of strategies used to cross the language barrier, the approaches of these three groups also show parallels. All use the combination of more than one type of translation resource, and in some cases also more than one translation resource of the same type. Université de Neuchâtel uses a range of machine translation systems, supplemented by an electronic dictionary. UC Berkeley also uses a dictionary, plus some of the same machine translation systems, but combines these resources with a corpus-based translation resource built from parallel texts. Eurospider combines machine translation using different systems

with a "similarity thesaurus" derived from suitable training data. All three systems use query translation, although Eurospider also combines this with document translation through machine translation.

For multilingual retrieval, several alternatives for the handling of all the languages exist. They can be handled simultaneously, or they can be handled one at a time, through a succession of bilingual retrieval steps, and then subsequently merged into one, multilingual result. All three groups have used the latter approach for query translation (when using document translation, this problem does not arise, since retrieval on the translated document collection is monolingual). However, the effective merging of the various bilingual results has proven to be a difficult challenge for all participants in the last two campaigns, and it appears that no robust, well-performing methods have been found. While all use several (mostly simple) methods for merging, the merging performance of these three groups appears to be still far below the theoretical optimum.

In summary, and looking at these three systems, we conclude that one possible blueprint for building a system that is successful for the CLEF 2002 multilingual track could well consist of:

- effective monolingual retrieval for most or all of the languages involved. This is achieved through use of robust stemming, well-known weighting schemes and blind feedback.
- a combination of translation information derived from multiple types of translation resources. The right parameterization and combination of these elements leads to effective retrieval results on the CLEF test collection.
- using query translation for a series of bilingual retrieval steps for individual language pairs. The results are then merged into one, multilingual result set.

A total of eleven groups submitted experiments to the multilingual track in CLEF 2002, and many alternative ideas to the ones outlined above were proposed, some radically different, and also successful. We will address some of these approaches in the following sections. Even so, elements of the same blueprint can be detected in some further experiments by other participants in both the multilingual and bilingual tracks, such as Océ (Brand and Brünner 2003), University of Exeter (Lam-Adesina and Jones 2003) and others. The challenge remains to prove that the success of this specific blueprint is generalizable to other test collections and operational settings.

4.2. Other approaches

In 2000, combination systems, i.e. systems combining more than one type of translation resource, were already used to some extent by groups from Eurospider (Braschler and Schäuble 2001), from Johns Hopkins University (JHU-APL) (McNamee et al. 2001), the Twenty-One group at TNO-TPD and University of Twente (Hiemstra et al. 2001) and the Laboratoire RALI at Université de Montreal (Nie et al. 2001). However, in general, combination approaches were not widely used during the first CLEF campaign in 2000, indicating that participants have adopted them over the years on the basis of experiences reported on in earlier campaigns. The typical CLEF 2000 system was based on only one type of translation resource, most likely machine-readable dictionaries (MRD). There was considerable work on using corpus-based techniques for disambiguating different translation alternatives, but

less effort on using such corpus-based techniques directly for translation. Then, as now, query translation was by far the preferred approach. In 2001, a massive increase in interest in corpus-based approaches for translation could be observed (This, and other fluctuations in the type of resources used by participants for their experiments, is documented in Section 4.5 of this paper). The number of groups using some type of corpus-based resource, either exclusively or in combination with other alternatives, doubled. Even so, machine-readable dictionaries remained the most popular choice, which we believe has a lot to do with newcomers often adopting them as their first choice when starting out. Indeed, the fluctuation of groups using dictionary-based approaches is unusually high compared to the alternatives of corpus-based resources and machine translation. The 2001 campaign seems to have been the campaign where participants experimented with the largest overall number of different resources, while scaling back somewhat for 2002, keeping the resources and methods that worked well for them in 2001. A possible interpretation is that for many groups, 2000 was a starting point for their work in CLIR within an evaluation setting, meaning that they started out with rather simple systems. In 2001, these groups frequently adopted many of the ideas proposed by other groups, trying a large range of alternatives. Based on their experiences in that campaign, they then kept the pieces that worked well for them, and concentrated on making them work even better in 2002. The emergence of well-performing systems with strong parallels for the multilingual track would seem to be consistent with this possible pattern.

4.3. Learning curve

Apart from adapting and enhancing other group's ideas, over the years participants have also moved from simpler to more complex systems, and from easier (monolingual) to harder (bilingual, or even multilingual) tasks. There are a number of groups that have progressed from only participating in monolingual retrieval in CLEF 2000 to producing full-blown multilingual experiments in 2002. We are very happy to see this effect, as it is an indication that participation in CLEF can stimulate groups to expand or enhance their systems. Examples of groups that have progressed like this are the groups from Istituto Trentino di Cultura (ITC-irst) (Bertoldi and Federico 2004), from University of Amsterdam (Hollink et al. 2004) and from University of Tampere (bilingual to multilingual) (Hedlund et al. 2004), which are all represented by papers in this special issue.

4.4. Thinking outside the box—Some more unusual approaches

As previously stated, the CLEF campaigns are not set up to be a competition. While many participants seek to perfect their systems with respect to the effectiveness they attain in CLEF tracks and tasks, this is by no means the only valuable use of the evaluation resources provided by the CLEF consortium. A substantial number of groups want to try new, unproven ideas, regardless of the possibility that these methods may not be successful on the CLEF task. Not only are such contributions invaluable in stimulating new strands of research and avoiding a "monoculture" of look-alike CLIR systems, they also tend to help to increase the quality of the evaluation resources produced by the campaign, as they increase the breadth of results that get judged. We cannot in this restricted space do justice to all the creative ideas

that have been proposed so far in the first three years of CLEF campaigns, but nevertheless we try to give a suggestion of the diversity of CLEF experiments by highlighting some of the work that fell outside of the norm for the respective year.

1. no translation. It seems a fair assumption that for successful cross-language retrieval, the system must translate either the query, the documents, or both to bridge the language gap. Contrary to this, the group at Johns Hopkins University (JHU-APL) has produced CLIR experiments that use no translation at all, instead matching on character *n*-grams shared between words in the respective languages. The method works best when the two languages are closely related etymologically, and when the query can be massively expanded prior to retrieval, in order to obtain as many *n*-gram matches as possible. For more details see McNamee and Mayfield (2004).
2. random indexing. The Swedish Institute for Computer Science (SICS) has demonstrated a new corpus-based approach that uses bit vectors of comparatively small length to represent relations between terms. They have used this technique both to calculate term-term similarities across languages on multilingual training data (Sahlgren and Karlgren 2002) and for query expansion in monolingual retrieval (Sahlgren et al. 2003).
3. lexical triangulation. When no translation resources are available for direct translation between two languages, one alternative is the use of a "pivot" intermediary language, translating from the source language to the pivot (often English), and then translating from the pivot to the target language. The drawback of this method is the amplification of translation ambiguities due to the multiple translation steps. A group from the University of Sheffield (Gollins and Sanderson 2001) demonstrated a technique that tries to minimize this additional ambiguity by using multiple pivot languages, and combining the information derived from them.
4. translation selection (interactive). Most participants view the main tracks of the CLEF campaign as a "batch processing setting", deriving queries entirely automatically from the topics, and then using no manual intervention during the retrieval process. The introduction of the interactive track has allowed interested groups to investigate some of the additional problems that arise when humans enter the "equation", such as how to facilitate document selection for users that have little or no knowledge of the language the documents are written in. Even before the interactive track started, a group from New Mexico State University (Ogden and Du 2000) investigated the question of whether a monolingual user can aid the query translation phase of cross-language retrieval, by disambiguating the query terms during interaction with the system, thus allowing more precise translation.
5. automorphology. One of the obstacles for scaling multilingual retrieval systems is the complexity involved in adding more languages to the system. Even systems that do not use elaborate linguistic knowledge and language-specific processing usually need at least some resources for new languages, such as a stemmer, or stopword lists. A group from University of Chicago (Goldsmith et al. 2001) investigated the question of how to automatically derive morphological information from a document corpus, without human supervision. They used their method as a stemming component for monolingual retrieval in CLEF.

4.5. Summarizing the use of different translation resource types in CLEF

So far, we have discussed the different approaches used in CLEF, the blueprints derived from them, and the learning curves demonstrated by some groups as we can derive them from the reports published during the past campaigns. In the following, we will attempt to summarize our discussion by analyzing the adoption of different types of translation resources in CLEF. For this purpose, we have divided the translation resources in four rough categories:

1. (Manually produced) Machine-readable dictionaries (MRD). All translation resources that consist of (manually assembled) lists of words and phrases and their associated translations. Manually produced thesauri are also included in this category.
2. Corpus-based approaches. All translation resources that have been automatically derived from suitable multilingual training data, such as parallel or comparable corpora. This includes methods such as latent semantic indexing, similarity thesauri, statistical translation models, translation probabilities, etc.
3. Machine Translation (MT). Full machine translation, i.e. the (attempt at) grammatically correct translation of entire sentences and documents.
4. No Translation at all. Approaches that use no direct translation resources, such as n-gram matching or cognate matching.

Not all methods and approaches can easily be classified according to this scheme. Some methods use a primary resource, such as a manually generated bilingual dictionary, coupled with a secondary resource, such as corpus-based translation probabilities. In these cases, when a type of translation resource is used only in a very limited fashion, we will note these secondary resources separately. Groups may use multiple types of primary and secondary resources in each campaign, either in a single experiment or in different experiments.

In an informal sense, we have for long time observed that participants learn from each other's experiences and adopt successful ideas. In addition to tallying the uses of different resource types, our goal was thus to investigate whether we can find evidence of "flows" between these four types of translation resources, i.e. if participants abandon one type of translation resource in order to move to a different type in the following campaign.

When looking at the numbers in Table 6, we can see that machine-readable dictionaries were the most used type of translation resource in all three campaigns covered by the analysis. A more in-depth analysis of this fact reveals that a large number of newcomers tend to adapt MRDs every year, boosting this number (Table 7). This may well be due to the fact that newcomers find it easiest to start out with dictionaries for their initial CLIR work (MT possibly being too much a "black box" to be integrated into their experiments). Even so, groups using (only) MRDs are also the most likely not to return for further campaigns, possibly indicating that these groups had limited agendas with regard to CLIR research.

Corpus-based approaches have seen an extraordinary surge between 2000 and 2001, doubling the number of participants using them. This is a good example of participants

Table 6. Use of different types of translation resources (all participants). The table shows the number of participants using a specific type of resource either as primary resource or as a secondary resource (figure in brackets).

Resource type	CLEF 2000	CLEF 2001	CLEF 2002
Machine-readable dictionary (MRD)	13 (1)	15 (1)	11 (0)
Corpus-based	4 (2)	8 (4)	6 (2)
Machine translation	5 (0)	7 (0)	6 (0)
No translation	0 (0)	0 (0)	1 (0)

Table 7. Use of different types of translation resources over the years (newly adopted resources). The table shows the number of participants newly adopting a specific type of resource either as primary resource (first figure) or as a secondary resource (figure in brackets).

Resource type	CLEF 2000	CLEF 2001	CLEF 2002
Machine-readable dictionary (MRD)	–	9 (1)	4 (0)
Corpus-based	–	4 (2)	1 (1)
Machine translation	–	4 (0)	2 (0)
No translation	–	0 (0)	1 (0)

taking up (successful) ideas from the earlier campaign, and trying to expand on them. The rise in the number of groups using such approaches was also facilitated by the emergence of combination approaches as an attractive option of CLIR.

It is interesting to note that in 2001 the highest total number of different types of resources was used by the participants. Apparently, after the initial 2000 campaign participants started investigating as many translation resource types as possible, determining in 2001 what works well for them. The 2002 campaign then seems to have brought some consolidation.

As mentioned, we have also attempted to detect flows between different types of translation resources, i.e. evidence of participants switching between translation technologies. The conclusion after our analysis has to be that there is in fact very little "flow" between translation resources from one campaign to the next. This may seem contrary to our informal observation that there is substantial uptake of ideas between groups. We attribute this phenomenon however to a trend we distinguish towards the use of combination systems: participants tend to add new types of translation resources to their systems rather than replacing them. This is not directly captured by tallying the use of individual resources. Furthermore, participants may shift the focus of their effort to different types of translation resources between years, even if they do not completely abandon them. Again, simply counting the number of groups using the different types of resources cannot uncover such subtle shifts. In fact, it is very hard to accurately derive the information necessary to track such minor shifts for as large a pool of participants as have participated in all the CLEF campaigns. Thus, while acknowledging the shortcomings of our analysis, we believe to

have found evidence that participants tend to shift the focus of their work instead of abandoning selected translation resources. Apart from possibly being a result of more robust and flexible combination approaches being used, this may also be an indication of the difficulty in acquiring more and better suited translation resources: resources are too valuable to throw away.

5. Directions for the future

In this paper, we have described the organization of the CLEF evaluation campaigns and the main results achieved so far. It is now the moment to ask whether we have gone far enough and whether our initial objectives as stated in the Introduction have been achieved:

- The infrastructure for the testing of multilingual information retrieval systems has been set up and has been shown to be operating well, scaling up to the more than 400 experiments submitted for the 2003 campaign.
- Test-suites of reusable data for a large number of European languages have been created and assessed for reliability; the end product of the EC-funded CLEF activity is envisaged to be test-suites that are publicly available rather than being restricted to CLEF participants as in the present.
- A strong CLIR research community, involving both academic and industrial participants, which recognizes CLEF as a forum for cross-language research and development activities, has been created.
- Research into improving the performance of systems for monolingual information retrieval for European languages other than English has been stimulated.
- A successful blueprint for the building of effective multilingual text retrieval systems appears to be in existence.

The question is thus whether there is any real need to continue with the CLEF campaigns.

To some extent, it is fair to say that research in the cross-language information retrieval sector is currently at a crossroads. In a recent workshop at SIGIR 2002[5] the question asked was whether the CLIR problem can now be considered as solved. The answers given were mixed: the basic technology for cross-language text retrieval systems is now in place as is clearly evidenced by the papers in this special issue. But if this is so, why has this technology not been adopted by any of the large Web search engines and why do most commercial information services not offer CLIR as a standard service? Although there is a strong market potential, the actual systems are still not ready to meet the needs of the generic user. For a commercial CLIR system to be successful, it needs to be versatile, efficient when working on-line, accommodate many ("all") languages, present its results in a sufficiently user-friendly fashion, and possibly handle multimedia. It is clear that much work remains to be done to address these points and bridge the present gap between the CLIR R&D community and the application world.

What role should CLEF play in this scenario? It seems evident that in the future, we must go further in the extension and enhancement of CLEF evaluation tasks, moving

gradually from a focus on cross-language text retrieval and the measurement of document rankings to the provision of a comprehensive set of tasks covering all major aspects of multilingual, multimedia system performance with particular attention to the needs of the end-user. This is not unlike what has happened in recent years with the TREC conferences.

Some of the further challenges left to address by CLEF include:

- What kind of evaluation methodologies should be developed to address more advanced information requirements?
- How can we cover the needs of all European languages—including minority ones?
- What type of coordination and funding model should be adopted—centralized or distributed?
- How can we best reduce the gap between research and application communities?

The most recent campaign, CLEF 2003, has attempted to move in the direction outlined above. It offered four additional tracks on top of the traditional core tracks for mono-, bi-multilingual and domain-specific systems. The aim has not only been to diversify the tasks offered but also to stimulate system developers to work in those areas that should find take-up by the application world. For this reason, in addition to continuing with the very successful interactive track, CLEF 2003 also included an activity for multilingual question answering (QA). The development of a successful question answering system implies not only a finely tuned IR engine but also sophisticated NLP functionality. Considerable work on QA system development has already been undertaken for English (with evaluation tracks at TREC) and has recently begun for Asian languages (with evaluation for systems running on Japanese at NTCIR), CLEF is attempting to encourage similar experiments for other European languages and also the development of QA systems that search across languages. Such systems recognise that users frequently want to retrieve certain specific pieces of information, and only that information, without having to wade through a large number documents to find it.

We are also beginning to take the first steps towards the evaluation of systems that run on collections of documents in more than one medium. Two pilot experiments were thus set up to test cross-language spoken document retrieval systems and cross-language retrieval on an image collection in the CLEF 2003 campaign. Both activities actually test the performance of particular types of text retrieval systems: automatic speech transcriptions in the first case, image captions in the second. Our aim is to stimulate system developers into examining and handling the problems involved in searching over multilingual, multimedia collections where language dependent and language independent factors interplay. These new tracks were organised as a result of proposals made at the CLEF 2002 workshop and consequent to the interest expressed in the ensuing discussion. Preliminary findings have been presented at the CLEF 2003 workshop.

We believe that if CLEF is to continue in the future it will be necessary to continue in this direction—being increasingly aware of and adapting to the needs of commercial system developers and offering a variety of tasks designed to meet these needs.

For more information on the CLEF evaluation campaigns, see http://www.clef-campaign.org.

Acknowledgments

The authors would like to acknowledge the help, support and advice of numerous individuals and organizations which has been invaluable in the organization of the CLEF campaigns. Many aspects of these campaigns have been modelled after successful blueprints that were previously tested and developed within the TREC conferences. CLEF itself started its "life" as a track organized within the TREC framework. We are particularly grateful to Donna Harman and Ellen Voorhees from NIST, organizers of TREC, for their tireless support. CLEF is an activity conducted by the CLEF project consortium, of which the two authors are but two members. We are greatly indebted to all the other consortium members, both individuals and their organizations. Lastly, we want to acknowledge all our data providers; too numerous to list here. Without their generous support the CLEF evaluation activity would be impossible.

Notes

1. CLEF 2000 and CLEF 2001 were sponsored by the European Commission under the Information Society Technologies programme and within the framework of the DELOS Network of Excellence for Digital Libraries (IST-1999-12262); CLEF 2002 and 2003 have been funded as an independent project (IST-2000-31002). The consortium members are ISTI-CNR, Italy (coordinators); ELDA, France; Eurospider Information Technology, Switzerland; IZ-Bonn, Germany; LSI-UNED, Spain; NIST, USA.
2. Recall measures the proportion of relevant documents retrieved with respect to the total number of relevant documents in the entire document collection. Precision measures the proportion of relevant documents retrieved with respect to the total number of documents retrieved from the collection.
3. More information on DELOS can be found at http://delos-noe.iei.pi.cnr.it/.
4. This is essential for economic reasons and important in order to create a reliable test collection. Relevance assessment is a very resource consuming task. By using the same topics, the relevance assessment for the three tracks can be done simultaneously and the document pools per topic for each language collection are sufficiently large.
5. See http://ucdata.berkeley.edu/sigir-2002/ for details on the workshop and, in particular, the position paper by Douglas W. Oard.

References

Amati G, Carpineto C and Romano G (2003) Italian monolingual information retrieval with PROSIT. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785. Spinger Verlag, pp. 257–264.

Bertoldi N and Federico M (2003) ITC-irst at CLEF 2002: Using N-best query translations for CL-IR. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785. Spinger Verlag, pp. 49–58.

Bertoldi N and Federico M (2004) Statistical models for monolingual and bilingual information retrieval. Information Retrieval, 7:51–70.

Brand R and Brünner M (2003) Océ at CLEF 2002. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785. Spinger Verlag, pp. 59–65.

Braschler M, Krause J, Peters C and Schäuble P (1999) Cross-language information retrieval (CLIR) track overview. In: Proceedings of the Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242, pp. 25–32.

Braschler M and Schäuble P (2001) Experiments with the Eurospider retrieval system for CLEF 2000. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000. Revised Papers, pp. 140–148.

Braschler M (2002) CLEF 2001—Overview of results. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Revised Papers.

Braschler M (2003) CLEF 2002—Overview of results. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785. Spinger Verlag, pp. 9–27.

Braschler M (2004) Combination approaches for multilingual text retrieval. Information Retrieval, 7:181–202.

Chen A and Gey FC (2004) Multilingual information retrieval using machine translation, relevance feedback and word decompounding. Information Retrieval, 7:147–180.

Cleverdon C (1977) The Cranfield tests on index language devices. In: Sparck-Jones K and Willett P, Eds. Readings in Information Retrieval: Morgan Kaufmann, pp. 47–59.

Goldsmith JA, Higgins D and Soglasnova S (2001) Automatic language-specific stemming in information retrieval. In: Peters C, Ed. Cross-Language Information Retrieval and Evaluation. Lecture Notes in Computer Science 2069. Springer Verlag, pp. 273–283.

Gollins T and Sanderson M (2001) Sheffield university: CLEF 2000 submission—Bilingual track: German to English. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000. Revised Papers, pp. 245–252.

Grefenstette G (1998), Ed. Cross-Language Information Retrieval. Kluwer Academic Publishers.

Harman D (1995) The TREC conferences. In: Kuhlen R and Rittberger M, Eds. Hypertext, Information Retrieval, Multimedia: Synergieeffekte Elektronischer Informationssysteme, Proceedings of HIM '95. Universitätsverlag Konstanz, pp. 9–28.

Harman D, Braschler M, Hess M, Kluck M, Peters C, Schäuble P and Sheridan P (2001) CLIR Evaluation at TREC. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000. Revised Papers, pp. 7–23.

Hedlund T, Airio E, Kekustalo H, Lehtokangas R, Pirkola A and Järvelin K (2004) Dictionary-based cross-language information retrieval: Learning experience from CLEF. Information Retrieval, 7: 97–117.

Hiemstra D, Kraaij W, Pohlmann R and Westerveld T (2001) Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000. Revised Papers, pp. 102–115.

Hollink V, Kamps J, Monz C and de Rijke M (2004) Monolingual retrieval for European languages. Information Retrieval, 7: 31–50.

Jones GJF and Federico M (2003) Cross-language spoken document retrieval pilot track report. In: Peters C, E Braschler M, Gonzalo J and Kluck M, Eds. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785. Spinger Verlag, pp. 446–457.

Kando N (2003) CLIR at NTCIR Workshop 3. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785. Spinger Verlag, pp. 485–504.

Kluck M and Gey FC (2001) The domain-specific task of CLEF—Specific evaluation strategies in cross-language information retrieval. In: Peters C, Eds. Cross-Language Information Retrieval and Evaluation. Lecture Notes in Computer Science 2069. Springer Verlag, pp. 48–56.

Lam-Adesina AM and Jones GJF (2003) Exeter at CLEF 2002: Experiments with machine translation for monolingual and bilingual retrieval. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785. Spinger Verlag, pp. 127–146.

Mandl T and Womser-Hacker C (2003) Linguistic and statistical analysis of the CLEF topics. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785. Spinger Verlag, pp. 505–511.

McNamee P, Mayfield J and Piatko C (2001) A language-independent approach to European text retrieval. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000. Revised Papers, pp. 129–139.

McNamee P and Mayfield J (2004) Character N-gram tokenization for European text retrieval. Information Retrieval, 7: 71–95.

Nie J-Y, Simard M and Foster G (2001) Multilingual information retrieval based on parallel texts from the Web. In: Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000. Revised Papers, pp. 188–201.

Oard DW, Gonzalo J, Sanderson M, López-Ostenero F and Wang J (2004) Interactive cross-language document selection. Information Retrieval, 7:203–226.

Ogden B and Du B (2000) Can monolingual users create good multilingual queries without machine translation? In: Peters C, Ed. First Results of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign. Working Notes for the CLEF 2000 Workshop, ERCIM-00-W01, pp. 133–134.

Pieters D (2002) The Languages of the European Union. Publication of the European Commission,IBSN 90-807420-3-1. Available at http://europa.eu.int/futurum/documents/offtext/espdiscuss10_en.pdf.

Sahlgren M and Karlgren J (2002) Vector-based semantic analysis using random indexing for cross-lingual query expansion. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Revised Papers, pp. 169–176.

Sahlgren M, Karlgren J, Cöster R and Järvinen T (2003) SICS at CLEF 2002: Automatic query expansion using random indexing. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, LNCS 2785. Spinger Verlag, pp. 311–320.

Savoy J (2004) Combining multiple strategies for effective monolingual and cross-language retrieval. Information Retrieval, 7:119–146.

Schäuble P and Sheridan P (1998) Cross-language information retrieval (CLIR) track overview. In: Proceedings of the Seventh Text Retrieval Conference (TREC-7). NIST Special Publication 500-422, pp. 25–32.

Smeaton AF and Harman D (1977) The TREC experiments and their impact on Europe. Journal of Information Science, 23:169–174.

Salton G (1970) Automatic processing of foreign language documents. Journal of the American Society for Information Science, 21(3):187–194.

Voorhees E (2002) The Philosophy of Information Retrieval Evaluation. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Revised Papers, pp. 355–370.

Voorhees E and Buckley C (2002) The effect of topic set size on retrieval experiment error. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 316–323.

Womser-Hacker C (2002) Multilingual Topic Generation within the CLEF 2001 experiments. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Revised Papers, pp. 389–393.

Zobel J (1998) How reliable are the results of large-scale information retrieval experiments? In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.