

Introduction to the CLEF 2003 Working Notes

Carol Peters

Istituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR), Pisa, Italy

carol.peters@isti.cnr.it

These Working Notes contain descriptions of the experiments conducted within CLEF 2003 – the fourth in a series of annual system evaluation campaigns organised by the Cross-Language Evaluation Forum¹. The results of the experiments will be presented and discussed in the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway. The final papers - revised and extended as a result of the discussions at the Workshop - together with a comparative analysis of the results will appear in the CLEF 2003 Proceedings, to be published by Springer in their Lecture Notes for Computer Science series.

CLEF organises a series of evaluation tracks designed to test different aspects of mono- and cross-language information retrieval system development. The intention is to encourage systems to move from monolingual text retrieval to the implementation of a full multilingual multimedia search service. The main features of the 2003 campaign are briefly outlined here below in order to provide the necessary background to the experiments reported in this volume.

1. Tracks and Tasks in CLEF 2003

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 methodologies for their testing in the multilingual information access context. The core tracks are coordinated by the members of the CLEF consortium, whereas the additional tracks are organized by interested associated groups on a voluntary basis, under the CLEF umbrella. Tracks can consist of a number of separate tasks. CLEF 2003 offered eight separate tracks evaluating the performance of systems for:

- multilingual information retrieval
- bilingual information retrieval
- monolingual (non-English) information retrieval
- domain-specific information retrieval
- interactive cross-language information retrieval
- mono- and cross-language question answering
- cross-language retrieval in image collections
- cross-language spoken document retrieval

1.1 Core Tracks

Multilingual/Bilingual/Monolingual Information Retrieval: The main track in CLEF is the multilingual one. One of the principal objectives of CLEF is to encourage developers to build systems capable of using a single query in a preferred query language to retrieve relevant documents in all languages in a multilingual collection, listing the results in a single, ranked list. In CLEF 2003, for the first time 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.

The objective of the 2003 bilingual track was to encourage the tuning of systems running on challenging language pairs that did not include English; however we also wanted to ensure comparability of results. For this reason, unlike CLEF 2002 where the bilingual track allowed participants to query any available target collection using queries in any permitted topic language, bilingual runs in CLEF 2003 were only accepted for one or more

¹ From October 2001, CLEF is supported by the European Commission under the IST programme (IST-2000-31002). The consortium members are: ISTI-CNR, Pisa; IZ Sozialwissenschaften, Bonn; ELRA/ELDA, Paris; Eurospider, Zurich; UNED, Madrid; NIST, USA. For more information, see: <http://www.clef-campaign.org>.

of the following source -> target languages pairs: Italian -> Spanish, German -> Italian, French -> Dutch, Finnish -> German. Exceptionally, newcomers only (i.e. groups that had not previously taken part in a CLEF cross-language track) 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 a Russian task in the bilingual track, permitting any language to be used for the queries.

CLEF also provides the opportunity for monolingual system testing and tuning, and for building test suites in European languages other than English. There were eight different target collections for monolingual system testing in CLEF 2003.

For each of these tracks (mono-/bi-/multilingual), the participating systems constructed their queries (automatically or manually) from a common set of topics, created to simulate user information needs. Each topic consists of three parts: a brief title statement; a one-sentence description; a more complex narrative specifying the relevance assessment criteria. Topic sets were produced by native speakers in the nine document languages and additionally for Chinese. A Japanese topic set will be made available in the future for use by participants in further experiments. As in previous years, a condition was that, for each task attempted, a mandatory run using the title and description fields had to be submitted. The objective is to facilitate comparison between the results of different systems.

Relevance assessment was also performed in all cases by native speakers. The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead approximate recall techniques are calculated using pooling techniques. The results submitted by the participating groups were used to form a pool of documents for each topic and for each language by collecting the highly ranked documents from all submissions. The results were then analysed and run statistics produced and distributed. These are given in Appendix B.

Domain-Specific Information Retrieval: The rationale for this track is to study CLIR on other types of collections, serving different kinds of information needs. The information provided by domain-specific scientific documents is highly targeted and contains much terminology. The domain-specific track offered mono- and bilingual tasks on the GIRT4 collection of social science documents. 25 topics were prepared in three languages: English, German and Russian. The topics were created and relevance assessments were performed by domain experts.

More information on the organisation of CLEF evaluation campaigns and the creation of the test collections or the core tracks can be found in Martin Braschler and Carol Peters: CLEF 2002 Methodology and Metrics, *Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign*. Lecture Notes in Computer Science, Vol. 2785, Springer 2003.

1.2 Additional Tracks

Interactive CLIR (iCLEF): This was the third year in which an interactive track was offered at CLEF. Two kinds of experiments were performed: experiments in cross-language document selection where the aim was to compare the degree to which different translation strategies can help the retrieval of relevant documents, and full cross-language search experiments where the user attempts to find as many relevant documents as possible with the help of a complete interactive CLIR system.

Multilingual Question Answering (QAatCLEF): Question answering was introduced as a new activity in CLEF 2003, and represents an important innovation as successful QA systems need to integrate both information retrieval and natural language processing methodologies. 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 experimental systems for cross-language QA. The track thus offered both monolingual and bilingual tasks: monolingual systems were evaluated within the framework of three non-English European languages, Dutch, Italian and Spanish while, in the cross-language task, Dutch, French, German, Italian or Spanish queries could be used on an English document collection. 200 questions were offered for each task.

Cross-Language Retrieval in Image Collections (ImageCLEF): This track was offered for the first time in CLEF 2003 as a pilot experiment. The task was 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. 50 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.

Cross-Language Spoken Document Retrieval (CL-SDR): The introduction of this track follows a first experimentation in 2002 within the framework of the DELOS Network of Excellence for Digital Libraries (<http://delos-noe.iei.pi.cnr.it/>). The aim this year in CLEF has been to evaluate CLIR systems on noisy automatic

transcriptions of spoken documents with known story boundaries. 100 topics (for both development and evaluation) were made available in English, French, German, Italian, Spanish and Dutch.

Details on the technical infrastructure, the collections used, and the organisation of these tracks can be found in the track overview reports in this volume, collocated at the beginning of the relevant sections.

2. Document Collections

The main CLEF multilingual corpus consists of sets of documents in different European languages but with common features (e.g., same genre and time period, comparable content). The collection was considerably expanded for CLEF 2003 and now contains well over 1.5 million documents in nine languages: Dutch, English, Finnish, French, German, Italian, Russian, Spanish and Swedish. Russian is an important new addition as it is the first collection in the CLEF corpus that does not use the Latin-1 encoding system; it has been formatted in XML and encoded in UTF-8. The CLEF corpus includes both newswires and national newspapers and most collections cover the period 1994-1995. Exceptions are the Finnish collection, which goes from November 1994 through 1995, and the Russian dataset which consists of newspaper documents for just 1995.

The domain-specific collection used in CLEF 2003 consisted of a new much larger version of the GIRT (German Indexing and Retrieval Test) social science database already used in previous editions of CLEF. This collection of 151319 documents consists of parallel corpora in English and in German. Controlled vocabularies in German-English and German-Russian were also made available to the participants in this track.

3. Participation

Participation in CLEF 2003 was slightly up with respect to the previous year with 43 groups submitting results for one or more of the different tracks (compared with 37 in 2002): 10 from N.America; 30 from Europe, and 3 from Asia. Of these, 9 were from industry and 34 from academia. Once again the number of system runs submitted for the core tracks increased dramatically (approximately 450 against the nearly 300 received in 2002). The breakdown of participation of groups per track is as follows: multilingual-8: 7; multilingual-4: 14; bilingual: 13; monolingual: 22; GIRT: 4; iCLEF: 5; QAatCLEF: 8; ImageCLEF: 4; CL-SDR: 4. As in previous years, participating groups consist of a nice mix of new-comers (15) and veteran groups (28) coming back for a second, third or even fourth time. Another important trend that is again noticeable this year is the progression of many of the returning groups to a more complex task this year, from monolingual to bilingual, from bilingual to multilingual.

CLEF 2003 posed a number of challenges to the participants, in particular with respect to the multilingual and bilingual tracks. We have been pleased by the results. We view the fact that 7 groups were brave enough to try the multilingual-8 task and that a considerable number of participants were willing to test their bilingual retrieval performance on “unusual” language pairs as very encouraging. Our aim is to stimulate groups to develop flexible systems, capable of handling many different language combinations rather than a favourite two or three. Another very positive aspect of CLEF2003 has been the number of new tracks and tasks that have been offered as pilot experiments. The aim has been to try out new ideas and develop new evaluation methodologies, suited to the emerging requirements of both system developers and users with respect to today’s digital collections and to encourage work on many European languages rather than just those most widely used. CLEF is thus gradually pushing its participants towards the ultimate goal: the development of truly multilingual systems capable of processing collections in diverse media.

4. Working Notes and Workshop

The Working Notes provide a first description of the different experiments run by the participating groups. The volume is divided into two main parts: Core Tracks and Additional Tracks, with further subdivisions into sections largely corresponding to the different tracks. The papers have thus been collocated in the section considered most appropriate, even though many papers describe more than one type of experiment. A final section contains reports on evaluation issues and other evaluation initiatives. There are two appendixes. Appendix A contains the run statistics for the question answering task. Appendix B gives a list of the participants and a summary of the characteristics of all runs for the core tracks together with overview graphs for the different tasks and individual statistics for each run.

The aim of the Workshop is to give all the groups that have participated in the CLEF 2003 evaluation campaign the opportunity to get together in order to compare approaches and to exchange ideas. The work of the groups participating in this year’s campaign will be presented in paper and poster sessions. Additional talks will include a report of the activities of the NTCIR evaluation initiative for Asian languages, and a report on cross-language information retrieval work at Moscow State University. The final session will discuss perspectives for future evaluation activities within the CLEF framework. The Workshop should thus provide an ample overview of the

current state-of-the-art and the latest research directions in the multilingual information retrieval area. The presentations at the Workshop will be posted on the CLEF website at <http://www.clef-campaign.org>.

We very much hope that the Workshop will prove an interesting, worthwhile and enjoyable experience to all those who participate.

Acknowledgements

It would be impossible to run the CLEF evaluation campaigns and organise the annual workshops without considerable assistance from many groups, working mainly on a voluntarily basis. We have numerous people and organisations to thank for their help in the running of CLEF 2003.

Here below I list some of them:

The Workshop Steering Committee

Martin Braschler, Eurospider Information Technology, Switzerland
Khalid Choukri, Evaluations and Language resources Distribution Agency, Paris, France
Marcello Federico, Centro per la Ricerca Scientifica e Tecnologica, Istituto Trentino di Cultura, Italy
Julio Gonzalo Arroyo, Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, Madrid, Spain
Donna Harman, National Institute of Standards and Technology, USA
Gareth Jones, University of Exeter, UK
Noriko Kando, National Institute of Informatics, Japan
Michael Kluck, IZ Sozialwissenschaften, Bonn, Germany
Bernardo Magnini, Centro per la Ricerca Scientifica e Tecnologica, Istituto Trentino di Cultura, Italy
Douglas W. Oard, University of Maryland, USA
Mark Sanderson, University of Sheffield, UK
Peter Schäuble, Eurospider Information Technology, Switzerland
Ellen Voorhees, National Institute of Standards and Technology, USA

CLEF Consortium Management

Francesca Borri, ISTI-CNR, Pisa, Italy

Associated Members of the CLEF Consortium

- Department of Information Studies, University of Tampere, Finland – responsible for work on the Finnish collection
- Human Computer Interaction and Language Engineering Laboratory, SICS, Kista, Sweden - responsible for work on the Swedish collection
- University of Twente, Centre for Telematics and Information Technology, The Netherlands - responsible for work on the Dutch collection
- Universität Hildesheim, Institut für Angewandte Sprachwissenschaft - Informationswissenschaft, Germany – responsible for checking and revision of the multilingual topic set
- College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA – co-organisers of iCLEF
- Centro per la Ricerca Scientifica e Tecnologica, Istituto Trentino di Cultura, Italy, main coordinators of the multilingual question answering track and co-organisers of the cross-language spoken document retrieval track
- University of Exeter, co-organisers of the cross-language spoken document retrieval track
- University of Sheffield, coordinators of the cross-language image retrieval track

In addition, I should like to thank colleagues from the Natural Language Processing Lab, Department of Computer Science and Information Engineering, National Taiwan University, for preparing topics in Chinese, and the Moscow State University, Russia, for their assistance in obtaining the Russian collection

We also gratefully acknowledge the support of all the data providers and copyright holders, and in particular:

- The Los Angeles Times, for the American English data collection;

- Le Monde S.A. and ELDA: Evaluations and Language resources Distribution Agency, for the French data.
- Frankfurter Rundschau, Druck und Verlagshaus Frankfurt am Main; Der Spiegel, Spiegel Verlag, Hamburg, for the German newspaper collections.
- InformationsZentrum Sozialwissenschaften, Bonn, for the GIRT database.
- Hypersystems Srl, Torino and La Stampa, for the Italian newspaper data.
- Agencia EFE S.A. for the Spanish newswire data.
- NRC Handelsblad, Algemeen Dagblad and PCM Landelijke dagbladen/Het Parool for the Dutch newspaper data.
- Aamulehti Oyj for the Finnish newspaper documents
- Tidningarnas Telegrambyrå for the Swedish newspapers
- The Herald 1995, SMG Newspapers, for the British English newspaper data
- Schweizerische Depeschagentur, Switzerland, for the French, German and Italian Swiss news agency data.
- Russika-Izvestia for the Russian collection
- St Andrews University Library for the image collection
- NIST for access to the TREC-8 and TREC-9 SDR transcripts.

Without their help, this evaluation activity would be impossible.

Last and not least, I should like to express our gratitude to both Francesca Borri in Pisa and the ECDL conference organisers in Trondheim for their assistance in the local organisation of the CLEF 2003 Workshop.

Carol Peters
August 2003