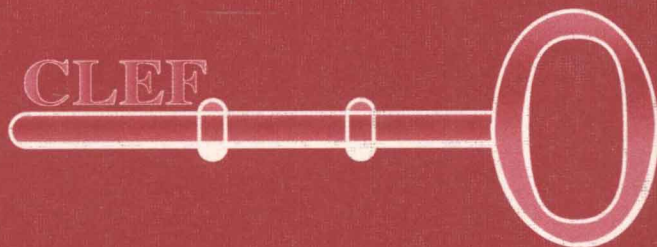Carol Peters (Ed.)

Cross-Language Information Retrieva and Evaluation

Workshop of the Cross-Language Evaluation Forum, CLEF 2000
Lisbon, Portugal, September 2000
Revised Papers

CLEF

Springer

Carol Peters (Ed.)

Cross-Language Information Retrieval and Evaluation

Workshop of the Cross-Language Evaluation Forum, CLEF 2000
Lisbon, Portugal, September 21-22, 2000
Revised Papers

Springer

Preface

The first evaluation campaign of the Cross-Language Evaluation Forum (CLEF) for European languages was held from January to September 2000. The campaign culminated in a two-day workshop in Lisbon, Portugal, 21–22 September, immediately following the fourth European Conference on Digital Libraries (ECDL 2000). The first day of the workshop was open to anyone interested in the area of Cross-Language Information Retrieval (CLIR) and addressed the topic of CLIR system evaluation. The goal was to identify the actual contribution of evaluation to system development and to determine what could be done in the future to stimulate progress. The second day was restricted to participants in the CLEF 2000 evaluation campaign and to their experiments. This volume constitutes the proceedings of the workshop and provides a record of the campaign.

CLEF is currently an activity of the DELOS Network of Excellence for Digital Libraries, funded by the EC Information Society Technologies to further research in digital library technologies. The activity is organized in collaboration with the US National Institute of Standards and Technology (NIST). The support of DELOS and NIST in the running of the evaluation campaign is gratefully acknowledged.

I should also like to thank the other members of the Workshop Steering Committee for their assistance in the organization of this event.

April 2001   Carol Peters

CLEF 2000 Workshop Steering Committee

Martin Braschler, Eurospider, Switzerland
Julio Gonzalo Arroyo, UNED, Madrid, Spain
Donna Harman, NIST, USA
Michael Hess, University of Zurich, Switzerland
Michael Kluck, IZ Sozialwissenschaften, Bonn, Germany
Carol Peters, IEI-CNR, Pisa, Italy
Peter Schäuble, Eurospider, Switzerland

Table of Contents

Introduction

Carol Peters

Istituto di Elaborazione della Informazione, CNR
Area della Ricerca di San Cataldo, 56124 Pisa, Italy
carol@iei.pi.cnr.it

The objective of the Cross-Language Evaluation Forum (CLEF) is to develop and maintain an infrastructure for the testing and evaluation of information retrieval systems operating on European languages, in both monolingual and cross-language contexts, and to create test-suites of reusable data that can be employed by system developers for benchmarking purposes. The first CLEF evaluation campaign started in early 2000 and ended with a workshop in Lisbon, Portugal, 22-23 September 2000.

This volume constitutes the proceedings of the workshop and also provides a record of the results of the campaign. It consists of two parts and an appendix. The first part reflects the presentations and discussions on the topic of evaluation for cross-language information retrieval systems during the first day of the workshop, whereas the second contains papers from the individual participating groups reporting their experiments and analysing their results. The appendix presents the evaluation techniques and measures used to derive the results and provides the run statistics. The aim of this Introduction is to present the main issues discussed at the workshop and also to provide the reader with the necessary background to the experiments through a description of the tasks set for CLEF 2000. In conclusion, our plans for future CLEF campaigns are outlined.

1 Evaluation for CLIR Systems

The first two papers in Part I of the proceedings describe the organization of cross-language evaluation campaigns for text retrieval systems. CLEF is a continuation and expansion of the cross-language system evaluation activity for European languages begun in 1997 with the track for Cross-Language Information Retrieval (CLIR) in the Text REtrieval Conference (TREC) series. The paper by Harman et al. gives details on how the activity was organized, the various issues that had to be addressed, and the results obtained. The difficulties experienced during the first year, in which the track was coordinated centrally at NIST (US National Institute for Standards and Technology) led to the setting up of a distributed coordination in four countries (USA, Germany, Italy and Switzerland) with native speakers being responsible for the preparation of topics (structured statements of possible information needs) and relevance judgments (assessment of the relevance of the ranked lists of results submitted by participating systems). A natural consequence of this distributed coordination was the

decision, in 1999, to transfer the activity to Europe and set it up independently as CLEF. The infrastructure and methodology adopted in CLEF is based on the experience of the CLIR tracks at TREC.

The second paper by Kando presents the NTCIR Workshops, a series of evaluation workshops for text retrieval systems operating on Asian languages. The 2000-2001 campaign conducted by NTCIR included cross-language system evaluation for Japanese-English and Chinese-English. Although both CLEF and NTCIR have a common basis in TREC there are interesting differences between the methodology adopted by the two campaigns. In particular, NTCIR employs multigrade relevance judgments rather than the binary system used by CLEF and inherited from TREC. Kando motivates this decision and discusses the effects.

The CLEF campaign provides participants with the possibility to test their systems on both general-purpose texts (newspapers and newswires) and domain-specific collections. The third paper by Kluck and Gey examines the domain-specific task, begun in TREC and continued in CLEF, and describes the particular document collection used: the GIRT database for social sciences.

The rest of the papers in the first part of this volume focus on some of the main issues that were discussed during the first day of the workshop. These included the problem of resources, the transition from the evaluation of cross-language text retrieval systems to systems running on other media, the need to consider the user perspective rather than concentrating attention solely on system performance, and the importance of being able to evaluate single system components rather than focusing on overall performance. A further point for discussion was the addition of new languages to the multilingual document collection.

The problem of resources has always been seen as crucial in cross-language system development. In order to be able to match queries against documents, some kind of lexical resource is needed to provide the transfer mechanism, e.g. bilingual or multilingual dictionaries, thesauri, or corpora. In order to be able to process a number of different languages, suitable language processing tools are needed, e.g. language-specific tokenizers, stemmers, morphologies, etc.. It is generally held that the quality of the resource used considerably affects system performance. This question was discussed at length during the workshop. The paper by Gonzalo presents a survey on the different language resources used by the CLEF 2000 participants. Many of the resources listed were developed by the participants themselves, thus showing that an evaluation exercise of this type is not only evaluating systems but also the resources used by the systems. The need for more pooling and sharing of resources between groups in order to optimize effort emerges clearly from this survey. Gonzalo concludes with some interesting proposals for the introduction of additional tasks, aimed at measuring the effect of the resources used on overall system performance, in a future campaign.

The papers by Oard and by Jones both discuss CLIR from the user perspective. Oard focuses on the document selection question: how the users of a CLIR system can correctly identify the - for them - most useful documents from a ranked list of results when they cannot read the language of the target collection. He advocates the advantages of an interactive CLIR evaluation and makes a proposal as to how an evaluation of this type could be included in CLEF. Jones also supports the extension of evaluation exercises in order to assess the usefulness of techniques that can assist the user with

relevance judgment and information extraction. In this respect, he mentions the importance of document summarization – already included in the NTCIR evaluation programme. In addition, Jones talks about work in cross-language multimedia information retrieval and suggests directions for future research. He asserts that specifically-developed standard test collections are needed to advance research in this area.

In the final paper in Part I, Gey lists several areas in which research could lead to improvement in cross-language information retrieval including resource enrichment, the use of pivot languages and phonetic transliteration. In particular, he discusses the need for post-evaluation failure analysis and shows how this could provide important feedback resulting in improved system design and performance. CLEF provides the research community with the necessary infrastructure for studies of this type.

2   The CLEF 2000 Experiments

There were several reasons behind the decision to coordinate the cross-language system evaluation activity for European languages independently and to move it to Europe. One was the desire to extend the number of languages covered, another was the intention to offer a wider range of retrieval tasks to better meet the needs of the multilingual information retrieval research community.

As can be seen from the descriptions of the experiments in Part II of this volume, CLEF 2000 included four separate evaluation tracks:
• multilingual information retrieval
• bilingual information retrieval
• monolingual (non-English) information retrieval
• cross-language domain-specific information retrieval

The main task – inherited from TREC - required searching a multilingual document collection, consisting of national newspapers in four languages (English, French, German and Italian) of the same time period, in order to retrieve relevant documents. Forty topics were developed on the basis of the contents of the multilingual collection – ten topics for each collection - and complete topic sets were produced in all four languages. Topics are structured statements of hypothetical user needs. Each topic consisted of three fields: a brief title statement; a one-sentence description; a more complex narrative specifying the relevance assessment criteria. Queries are constructed using one of more of these fields. Additional topic sets were then created for Dutch, Finnish, Spanish and Swedish, in each case translating from the original. The main requirement was that, for each language, the topic set should be as linguistically representative as possible, i.e. using the terms that would naturally be expected to represent the set of topic concepts in the given language. The methodology followed was that described in the paper by Harman et al..

A bilingual system evaluation task was also offered, consisting of querying the English newspaper collection using any topic language (except English). Many newcomers to cross-language system evaluation prefer to begin with the simpler bilingual task before moving on to tackle the additional issues involved in truly multilingual retrieval.

One of the aims of the CLEF activity is to encourage the development of tools to manipulate and process languages other than English. Different languages present different problems. Methods that may be efficient for certain language typologies may not be so effective for others. Issues that have to be catered for include word order, morphology, diacritic characters, language variants. For this reason, CLEF 2000 included a track for French, German and Italian monolingual information retrieval.

The cross-language domain-specific task has been offered since TREC-7. The rationale of this subtask is to test retrieval on another type of document collection, serving a different kind of information need. The implications are discussed in the paper by Kluck and Gey in the first part of this volume.

The papers in Part II describe the various experiments by the participating groups with these four tasks. Both traditional and innovative approaches to CLIR were experimented, and different query expansion techniques were tried. All kinds of source to target transfer mechanisms were employed, including both query and document translation. Commercial and in-house resources were used and included machine translation, dictionary and corpus-based methods. The strategies used varied from traditional IR to a considerable employment of natural language processing techniques. Different groups focused on different aspects of the overall problem, ranging from the development of language-independent tools such as stemmers to much work on language-specific features like morphology and compounding. Many groups compared different techniques in different runs in order to evaluate the effect of a given technique on performance. Overall, CLEF 2000 offered a very good picture of current issues and approaches in CLIR.

The first paper in this part by Martin Braschler provides an overview and analysis of all the results, listing the most relevant achievements and comparing them with those of previous years in the CLIR track at TREC. As one of the main objectives of CLEF is to produce evaluation test-suites that can be used by the CLIR research community, Braschler also provides an analysis of the test collection resulting from the CLEF 2000 campaign, demonstrating its validity for future system testing, tuning and development activities. The appendix presents the evaluation results for each group, run by run.

3   CLEF in the Future

The CLEF 2001 campaign is now under way. The main tasks are similar to those of the first campaign. There are, however, some extensions and additions. In particular the multilingual corpus has been considerably enlarged and Spanish (news agency) and Dutch (national newspaper) collections for 1994 have been added. The multilingual task in CLEF 2001 involves querying collections in five languages (English, French, German, Italian and Spanish) and there will be two bilingual tracks: searching either the English or the Dutch collections. Spanish and Dutch have also been included in the monolingual track. There will be seven official topic languages, including Japanese. Additional topics will be provided in a number of other European languages, including Finnish, Swedish and Russian, and also in Chinese and Thai.

CLEF 2000 concentrated on the traditional metrics of recall and precision – however these have limitations in what they tell us about the usefulness of a retrieval system to the user. CLEF 2001 will thus also include an experimental track designed to test interactive CLIR systems and to establish baselines against which future research progress can be measured. The introduction of this track is a direct result of discussions which began in the workshop with the presentations by Oard and by Jones, and of the proposal by Oard reported in Part I of this volume.

Two main issues must be considered when planning future CLEF campaigns: the addition of more languages, and the inclusion of new tasks.

The extension of language coverage, discussed considerably at the workshop, depends on two factors: the demand from potential participants and the existence of sufficient resources to handle the requirements of new language collections. It was decided that Spanish and Dutch met these criteria for CLEF 2001. CLEF 2002 and 2003 will be mainly funded by a contract from the European Commission (IST-2000-31002) but it is probable that, in the future, it will be necessary to seek support from national funding agencies as well if more languages are to be included. The aim will be to cover not only the major European languages but also some representative samples of minority languages, including members from each major group: e.g. Germanic, Romance, Slavic, and Ugro-Finnic languages. Furthermore, building on the experience of CLEF 2001, we intend to continue to provide topics in Asian languages.

CLEF 2000 concentrated on cross-language text retrieval and on measuring overall system performance. However, in the future, we hope to include tracks to evaluate CLIR systems working on media other than text. We are now beginning to examine the feasibility of organizing a spoken CLIR track in which systems would have to process and match spoken queries in more than one language against a spoken document collection. Another important innovation would be to devise methods that enable the assessment of single system components, as suggested in the paper by Gonzalo.

CLIR system development is still very much in the experimental stage and involves expertise from both the natural language processing and the information retrieval fields. The CLEF 2000 Workshop provided an ideal opportunity for a number of key players, with very different backgrounds, to come together and exchange ideas and compare results on the basis of a common experience: participation in the CLEF evaluation campaign. CLEF is very much a collaborative effort between organizers and participants with the same common goal: the improvement of CLIR system performance. The discussions at the workshop have had considerable impact on the organization of the 2001 campaign. The success of future campaigns will depend on the continuation and strengthening of this collaboration.

More information on the organization of the current CLEF campaign and instructions on how to contact us can be found at: http://www.clef-campaign.org/.

Acknowledgements

thank the other members of the CLEF Coordinating Group for all their efforts aimed at making both the campaign and the workshop a great success:

Martin Braschler, Eurospider, Switzerland
Julio Gonzalo Arroyo, UNED, Madrid, Spain
Donna Harman, NIST, USA
Michael Hess, University of Zurich, Switzerland
Michael Kluck, IZ Sozialwissenschaften, Bonn, Germany
Peter Schäuble, Eurospider, Switzerland
Felisa Verdejo Maillo, UNED, Madrid, Spain
Ellen Voorhees, NIST, USA
Christa Womser-Hacker, University of Hildesheim, Germany

I must also express my gratitude to the ECDL 2000 Conference organisers for their assistance in the local organisation of the CLEF Workshop, and in particular Caroline Hagège and Nuno Mamede, (Local Coordinators) and Eulália Carvalho, and José Luis Borbinha (ECDL Chair).

It is not easy to set up an infrastructure that meets the needs of a large number of languages. I should like to thank the following organisations who voluntarily engaged translators to provide topic sets in Dutch, Finnish and Swedish, working on the basis of the set of source topics:

- the DRUID project for the Dutch topics;
- the Department of Information Studies, University of Tampere, Finland, engaged the UTA Language Centre for the Finnish topics;
- SICS Human Computer Interaction and Language Engineering Laboratory for the Swedish topics.

The support of all the data providers and copyright holders is also gratefully acknowledged, and in particular:

- The Los Angeles Times, for the English data collection;
- Le Monde S.A. and ELDA: European 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 data.
- Schweizerische Depeschenagentur (SDA) and Associated Press (AP) for the newswire data of the training collection.

Without their help, this evaluation activity would be impossible.

Last, but not least, I thank Julio Gonzalo for his help and encouragement in the preparation of this volume.

CLIR Evaluation at TREC

Donna Harman[1], Martin Braschler[2], Michael Hess[3], Michael Kluck[4],
Carol Peters[5], Peter Schäuble[2], and Páraic Sheridan[6]

[1] National Institute of Standards and Technology, Gaithersburg, Md, USA.
donna.harman@nist.gov
[2] Eurospider Information Technology AG, Zürich, Switzerland
braschler@eurospider.com schauble@eurospider.com
[3] Dept. of Computer Science, University of Zurich, Switzerland
hess@ifi.unizh.ch
[4] InformationsZentrum Sozialwissenschaften (IZ), Bonn, Germany
mkl@bonn.iz-soz.de
[5] Istituto di Elaborazione della Informazione - CNR., Pisa, Italy
carol@iei.pi.cnr.it
[6] MNIS-Textwise Labs, Syracuse, N.Y. USA.
paraic@textwise.com

Abstract. Starting in 1997, the National Institute of Standards and Technology conducted 3 years of evaluation of cross-language information retrieval systems in the Text REtrieval Conference (TREC). Twenty-two participating systems used topics (test questions) in one language to retrieve documents written in English, French, German, and Italian. A large-scale multilingual test collection has been built and a new technique for building such a collection in a distributed manner was devised.

1   Introduction

The increasing globalization of information has led to an heightened interest in retrieving information that is in languages users are unable search effectively. Often these users can adequately read retrieved documents in non-native languages, or can use existing gisting systems to get a good idea of the relevance of the returned documents, but are not able to create appropriate search questions. Ideally they would like to search in their native language, but have the ability to retrieve documents in a *cross-language* mode.

The desire to build better cross-language retrieval systems resulted in a workshop on this subject at the Nineteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval in 1996. Whereas many of the participants at this conference were concerned with the lack of sufficient parallel text to form a basis for research, one of the papers presented at that workshop provided the hope of avoiding the use of parallel corpora by the use of *comparable* corpora.

This paper, by Páraic Sheridan, Jean Paul Ballerini and Peter Schäuble of the Swiss Federal Institute of Technology (ETH), [1], used stories from the Swiss

news agency Schweizerische Depeschen Agentur (SDA) that were taken from the same time period. These newswire stories are not translations but are produced independently in each language (French, German and Italian) in the various parts of Switzerland. Whereas the stories do not overlap perfectly, there is in fact a high overlap of stories (e.g. international events) which are of interest in all parts of Switzerland. The paper detailed the use of this collection of stories to produce a test collection that enabled the evaluation of a series of cross-language retrieval experiments [2].

In 1997 it was decided to include cross-language information retrieval (CLIR) system evaluation as one of the tracks at the Sixth Text REtrieval Conference (TREC) held at the National Institute of Standards and Technology (NIST) [3] [http://trec.nist.gov]. The aim was to provide researchers with an infrastructure for evaluation that would enable them to test their systems and compare the results achieved using different cross-language strategies. This track was done in cooperation with the Swiss Federal Institute of Technology, who not only obtained permission for TREC to use the SDA data, but also provided considerable guidance and leadership to the track.

The main goals of the CLIR track in TREC were:

1. to create the infrastructure for testing cross-language information retrieval technology through the creation of a large-scale multilingual test collection and a common evaluation setting;
2. to investigate effective evaluation procedures in a multilingual context; and
3. to provide a forum for the exchange of research ideas.

There were CLIR tracks for European languages in TREC-6, TREC-7, and TREC-8. The TREC proceedings for each year (available on-line at [http://trec.nist.gov], contain overviews of the track, plus papers from all groups participating in the CLIR track that year. The rest of this paper summarizes the CLIR work done in those three years, with those summaries derived from the various track overviews [4], [5], [6]. To conserve space, the numerous individual papers are not included in the references but can be found in the section for the cross-language track in the appropriate TREC proceedings. A table listing all participants for a given TREC is given in each result section to faciliate the location of the individual papers. Note that there are additional publications from these groups including further results and analyses, and the references in the track overviews should be checked to obtain these.

2   TREC-6 CLIR Track Task Description

The TREC-6 Cross-Language Information Retrieval (CLIR) track required the retrieval of either English, German or French documents that are relevant to topics written in a different language. Participating groups could choose any cross-language combination, for example English topics against German documents or French topics against English documents. In order have a baseline retrieval performance measurement for each group, the results of a monolingual retrieval

experimental run in the document language were also to be submitted. For instance, if a cross-language experiment was run with English topics retrieving German documents, then the result of an equivalent experiment where German topics retrieve German documents must also have been submitted. These results would be considered comparable since the topics are assumed to be proper translations across the languages.

The different document collections used for each language are outlined in Table 1. The Associated Press collection consists of newswire stories in English, while the French SDA collection is a similar collection of newswire stories from the Swiss news agency (Schweizerische Depeschen Agentur). The German document collection has two parts. The first part is composed of further newswire stories from the Swiss SDA while the second part consists of newspaper articles from a Swiss newspaper, the 'Neue Zuercher Zeitung' (NZZ). The Italian data is included in this table for completeness although it was not used in TREC-6.

The newswire collections in English, French and German were chosen to overlap in timeframe (1988 to 1990) for two reasons. First, since a single set of topics had to be formulated to cover all three document languages, having the same timeframe for newswire stories increased the likelihood of finding a greater number of relevant documents in all languages. The second reason for the overlapping timeframe was to allow groups who use corpus-based approaches for cross-language retrieval to investigate what useful corpus information they could extract from the document collections being used. One of the resources provided to CLIR track participants was a list of 83,698 news documents in the French and German SDA collections which were likely to be comparable based on an alignment of stories using news descriptors assigned manually by the SDA reporters, the dates of the stories, and common cognates in the texts of the stories.

Document Collections			
Doc. Language	**Source**	**No. Documents**	**Size**
English	AP news, 1988-1990	242,918	760MB
German	SDA news, 1988-1990	185,099	330MB
	NZZ articles, 1994	66,741	200MB
French	SDA news, 1988-1990	141,656	250MB
Italian	SDA news, 1989-1990	62,359	90MB

Table 1. Document Collections used in the CLIR track.

The 25 test topic descriptions were provided by NIST in English, French and German, using translations of topics originally written mostly in English (see Figure 1 for an example topic, including all its translations). Participating groups who wished to test other topic languages were permitted to create translations of the topics in their own language and use these in their tests, as long as the translated topics were made publicly available to the rest of the track

participants. The final topic set therefore also had translations of the 25 topics in Spanish, provided by the University of Massachusetts, and Dutch, provided by TNO in the Netherlands.

```
<num> Number: CL9
<E-title> Effects of logging

<E-desc> Description:
What effects has logging had on desertification?

<E-narr> Narrative:
Documents with specific mention of local government's or international
agencies' efforts to stop deforestation are relevant. Also relevant
are documents containing information on desertification and its
side effects such as climate change, soil depletion, flooding, and
hurricanes caused by excessive logging.

<num> Number: CL9
<F-title> Les effets de la déforestation

<F-desc> Description:
Quels sont les effets de la déforestation sur la désertification?

<F-narr> Narrative:
Tous les documents qui donnent des analyses spécifiques sur les mesures
des gouverments locaux ou des agences internationales pour frêner
la déforestation sont pertinants. Les articles qui contiennent des
renseignements sur la désertification et ses effets secondaires comme
les changements de climat, l'épuisement de la terre, les inondations et
les ouragans sont également applicables.

<num> Number: CL9
<G-title> Auswirkungen von Abholzung

<G-desc> Description:
Welche Auswirkungen hat das Abholzen auf die Ausbreitung der Wüste?

<G-narr> Narrative:
Alle Artikel über Bemühungen von Regierungen ebenso wie von
internationalen Agenturen die Wüstenausbreitung zu bremsen, sind
wesentlich. Ebenso relevant sind Artikel über Ausbreitung der Wüsten
und ihre Mitwirkungen, wie zum Beispiel Klimawechsel, Verarmung der
Erde und Orkane die auf übermässige Abholzung zurückzuführen sind.
```

Fig. 1. Sample CLIR topic statement from TREC-6, showing all languages.