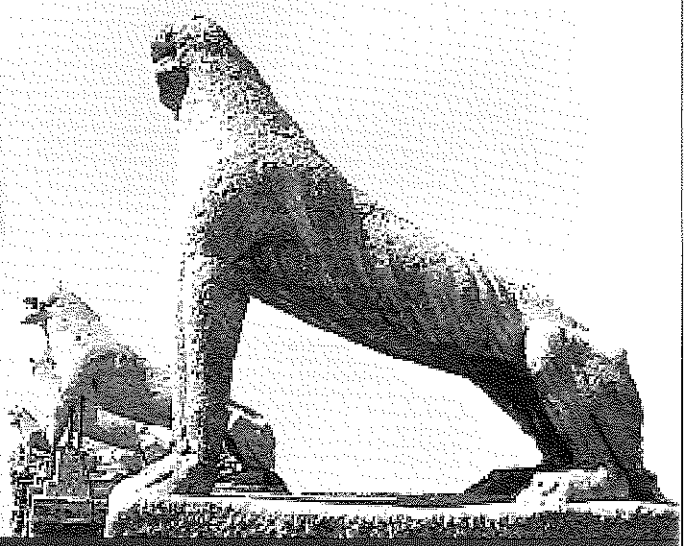
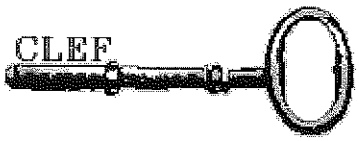
European Research Consortium
for Informatics and Mathematics
ERCIM
www.ercim.org



DELOS Network of Excellence on Digital Libraries Workshop Series

CROSS-LANGUAGE
EVALUATION FORUM



First Results of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign

Working Notes for the CLEF 2000 Workshop

22 September, Lisbon, Portugal

ERCIM-00-W01

First Results of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign

Working Notes for the CLEF 2000 Workshop
22 September, Lisbon, Portugal

edited by Carol Peters

CONTENTS

Foreword

These Working Notes present the preliminary results of CLEF 2000 – the first campaign of the Cross-Language Evaluation Forum, one of the activities of the DELOS Network of Excellence for Digital Libraries[1]. They contain descriptions of the systems and strategies used by the research groups participating in one or more of the set retrieval tasks. The final papers and a comparative analysis of the results will be published in the CLEF 2000 Proceedings, due to appear by the end of this year. The Working Notes provide the background information for the presentations and discussions by the participants in the CLEF 2000 Workshop, 22 September, Lisbon, Portugal.

The Cross-Language Evaluation Forum is a continuation and expansion of the cross-language system evaluation activity begun in 1997 at the Text REtrieval Conference (TREC) series, in the track for Cross-Language Information Retrieval (CLIR). The CLIR track was proposed and run with success at TREC for three years with a growing number of participating groups. In 1999, however, it was decided to move cross-language system evaluation for European languages to Europe. There were several reasons for this. The first experiences at TREC had demonstrated the necessity for this kind of activity to be organised on a distributed basis: it is important for topics and results assessments for a multilingual document collection to be managed by native speakers of each language involved. Thus, from 1998 on, the CLIR track was run by groups working in four countries (USA, Switzerland, Germany and Italy). However, the desire to further extend the number of languages catered for meant that Europe was seen as a more appropriate platform for future coordination. Moreover, it was felt that rendering the activity independent would make it possible to focus on a wider range of retrieval tasks. In fact, as can be seen from the papers in this collection, CLEF 2000 has 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 required searching a multilingual document collection, consisting of national newspapers in four languages (English, French, German and Italian) for the same time period, in order to retrieve relevant documents. 40 topics were developed on the basis of the contents of the multilingual collection and topic sets were produced in all four languages. 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 query concepts in the given language.

A bilingual system evaluation task was also offered. This consisted of querying the Los Angeles Times collection using any topic language (other than English). In later years, the target collection may well be in another language: French, German or Spanish, for example.

In order to be successful with multilingual retrieval, a good understanding of the questions involved in monolingual information retrieval is necessary. Different languages present different problems.

---

[1] CLEF is conducted as an EU-US collaboration. The US partner is the National Institute for Standards and Technology (NIST), Gaithersburg, MD. The European partners are all members of the DELOS Network of Excellence: Eurospider Information Technology, Switzerland; InformationsZentrum Sozialwissenschaften, Bonn, Germany; Istituto di Elaborazione della Informazione - CNR, Pisa, Italy; IEEC-UNED, Madrid, Spain; University of Zurich, Switzerland. For more information, see: http://www.ercim.org/delos

Methods that may be highly 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. One of the aims of the CLEF activity is to encourage the development of tools to manipulate and process languages other than English. For this reason, another of the CLEF 2000 tasks has regarded monolingual (non-English) information retrieval for systems developed to run on French, German and Italian.

The cross-language domain-specific task which covers a vertical domain (social sciences), 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.

It appears that this range of tasks was appealing as the response to the CLEF 2000 Call for Participation was very good: 28 groups signed up to take part in one or more of the tasks. In the end, 20 groups actually participated: 8 from N.America; 12 from Europe. A total of 90 runs were received; runs were submitted for all tasks (multilingual, bilingual, domain-specific and monolingual non-English) and for all topic languages. In consideration of the strict time constraints on this first year's activity, we are very satisfied with this result. The Working Notes provide a first description of the different experiments run by the participating groups. The results of the experiments can be found in the Appendix. Listed are a summary of the characteristics of all runs, followed by overview graphs for the different tasks and individual statistics for every run.

We should like to thank the ECDL 2000 Conference organisers for all their assistance in the organisation of the CLEF Workshop, and in particular Caroline Hagège and Nuno Mamede, (Local Coordinators) and Eulália Carvalho, and José Luis Borbinha (ECDL Chair). The aim is to give all the groups that have participated in the CLEF evaluation campaign the opportunity to get together in order to discuss and compare their approaches and to exchange ideas and experiences. It will also provide the opportunity for an open discussion on the organisation and scheduling of future CLEF evaluation campaigns.

We very much hope that the CLEF 2000 Workshop will prove a useful, enjoyable and memorable experience for everyone participating, and will be the first of a long series.

The 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

Acknowledgements

The topic sets were all prepared by independent groups, i.e. by groups not participating in the system evaluation tasks. The main topic sets (DE, EN, FR, IT) and the Spanish topics were prepared by the project partners. Here, we should like to express our gratitude to 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.