Carol Peters Paul Clough
Julio Gonzalo Gareth J. F. Jones
Michael Kluck Bernardo Magnini (Eds.)

LNCS 3491

Multilingual Information Access for Text, Speech and Images

5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004
Bath, UK, September 2004, Revised Selected Papers

CLEF

Springer

Carol Peters   Paul Clough
Julio Gonzalo   Gareth J.F. Jones
Michael Kluck   Bernardo Magnini (Eds.)

Multilingual Information Access for Text, Speech and Images

5th Workshop of the
Cross-Language Evaluation Forum, CLEF 2004
Bath, UK, September 15-17, 2004
Revised Selected Papers

Springer

Volume Editors

Carol Peters
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche (ISTI-CNR)
Via G. Moruzzi 1, 56124 Pisa, Italy
E-mail: carol.peters@isti.cnr.it

Paul Clough
University of Sheffield, Department of Information Studies
Western Bank, Sheffield, S10 2TN, UK
E-mail: p.d.clough@sheffield.ac.uk

Julio Gonzalo
Universidad Nacional de Educación a Distancia
Departamento de Lenguajes y Sistemas Infomáticos
Juan del Rosal, 16, 28040 Madrid, Spain
E-mail: julio@lsi.uned.es

Gareth J.F. Jones
Dublin City University, School of Computing
Glasnevin, Dublin 9, Ireland, UK
E-mail: gareth.jones@computing.dcu.ie

Michael Kluck
Informationszentrum Sozialwissenschaften (IZ)
Lennéstr. 30, 53113 Bonn, Germany
E-mail: michaelkluck@web.de

Bernardo Magnini
Centro per la Ricerca Scientifica e Tecnologica (ITC-Irst)
Via Sommarive 18, 38050 Povo (TN), Italy
E-mail: magnini@itc.it

Library of Congress Control Number: 2005928950

CR Subject Classification (1998): H.3, I.2, H.4

ISSN 0302-9743
ISBN-10 3-540-27420-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-27420-9 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11519645 06/3142 5 4 3 2 1 0

Preface

The fifth campaign of the Cross-Language Evaluation Forum (CLEF) for European languages was held from January to September 2004. Participation in the CLEF campaigns has increased each year and CLEF 2004 was no exception: 55 groups submitted results for one or more of the different tracks compared with 42 groups in the previous year. CLEF 2004 also marked a breaking point with respect to previous campaigns. The focus was no longer mainly concentrated on multilingual document retrieval as in previous years but was diversified to include different kinds of text retrieval across languages (e.g., exact answers in the question-answering track) and retrieval on different kinds of media (i.e., not just plain text but collections containing image and speech as well). In addition, increasing attention was given to issues that regard system usability and user satisfaction with tasks to measure the effectiveness of interactive systems or system components being included in both the cross-language question answering and image retrieval tasks with the collaboration of the coordinators of the interactive track.

The campaign culminated in a two-and-a-half-day workshop held in Bath, UK, 15–17 September, immediately following the 8th European Conference on Digital Libraries. The workshop was attended by nearly 100 researchers and system developers. In addition to presentations by participants in the campaign, talks included reports on the activities of the NTCIR evaluation initiative for Asian languages, and on industrial experience in building cross-language applications. The final session consisted of a panel in which the members attempted to analyze the current organization of the CLEF campaigns in depth, discussing whether CLEF is working on the right problems, choosing its investments wisely, and giving sufficient attention to the user perspective. Suggestions for the CLEF 2005 campaign included multilingual Web retrieval and a cross-language Geographic Information Retrieval track.

CLEF 2004 was conducted as an activity of the DELOS Network of Excellence on Digital Libraries, within the framework of the Information Society Technologies programme of the European Commission. These post-campaign proceedings were prepared with the assistance of the Center for the Evaluation of Language and Communication Technologies (CELCT), Trento, Italy. The support of DELOS and CELCT is gratefully acknowledged. We should also like to thank the other members of the CLEF Steering Committee for their assistance in the coordination of this event.

April 2005

Carol Peters
Paul Clough
Julio Gonzalo
Gareth J.F. Jones
Michael Kluck
Bernardo Magnini

CLEF 2004 Coordination

CLEF is coordinated by the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, ISTI-CNR, Pisa.

The following institutions contributed to the organization of the different tracks of the CLEF 2004 campaign:

- Center for the Evaluation of Language and Communication Technologies (CELCT), Trento, Italy
- Centro per la Ricerca Scientifica e Tecnologica, Istituto Trentino di Cultura, Trento, Italy
- College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, USA
- Department of Computer Science and Information Systems, University of Limerick, Ireland
- Department of Information Studies, University of Sheffield, UK
- Department of Information Studies, University of Tampere, Finland
- Eurospider Information Technology AG, Zürich, Switzerland
- Evaluations and Language Resources Distribution Agency, ELDA, Paris, France
- German Research Centre for Artificial Intelligence, DFKI, Saarbrücken, Germany
- Information and Language Processing Systems, University of Amsterdam, Netherlands
- Informationszentrum Sozialwissenschaften, Bonn, Germany
- Lenguajes y Sistemás Informáticos, Universidad Nacional de Educación a Distancia, Madrid, Spain
- Linguateca, Sintef, Oslo, Norway
- Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Sofia, Bulgaria
- National Institute of Standards and Technology, Gaithersburg MD, USA
- School of Computing, Dublin City University, Ireland
- University and University Hospitals of Geneva, Switzerland

CLEF 2004 Steering Committee

- Maristella Agosti, University of Padua, Italy
- Eija Airio, University of Tampere, Finland
- Martin Braschler, Eurospider Information Technologies, Switzerland
- Hsin-Hsi Chen, National Taiwan University, Taipei, Taiwan
- Khalid Choukri, Evaluations and Language Resources Distribution Agency, ELDA, Paris, France
- Paul Clough, University of Sheffield, UK
- David A. Evans, Clairvoyance Corporation, USA
- Marcello Federico, ITC-irst, Trento, Italy
- Christian Fluhr, CEA-LIST, Fontenay-aux-Roses, France
- Norbert Fuhr, University of Duisburg, Germany
- Frederic C. Gey, UC Berkeley, USA
- Julio Gonzalo, LSI-UNED, Madrid, Spain
- Donna Harman, National Institute of Standards and Technology, USA
- Gareth Jones, Dublin City University, Ireland
- Franciska de Jong, University of Twente, Netherlands
- Noriko Kando, National Institute of Informatics, Tokyo, Japan
- Jussi Karlgren, Swedish Institute of Computer Science, Sweden
- Michael Kluck, Informationszentrum Sozialwissenschaften Bonn, Germany
- Natalia Loukachevitch, Moscow State University, Russia
- Bernardo Magnini, ITC-irst, Trento, Italy
- Paul McNamee, Johns Hopkins University, USA
- Henning Müller, University and University Hospitals, Geneva, Switzerland
- Douglas W. Oard, University of Maryland, USA
- Maarten de Rijke, University of Amsterdam, Netherlands
- Jacques Savoy, University of Neuchâtel, Switzerland
- Peter Schäuble, Eurospider Information Technologies, Switzerland
- Richard Sutcliffe, University of Limerick, Ireland
- Hans Uszkoreit, German Research Center for Artificial Intelligence (DFKI), Germany
- Felisa Verdejo, LSI-UNED, Madrid, Spain
- José Luis Vicedo, University of Alicante, Spain
- Ellen Voorhees, National Institute of Standards and Technology, USA
- Christa Womser-Hacker, University of Hildesheim, Germany

Table of Contents

Part I. Ad Hoc Text Retrieval Tracks

Mainly Cross-Language Experiments

Monolingual Experiments

Part II. Domain-Specific Document Retrieval

Part III. Interactive Cross-Language Information Retrieval

Part IV. Multiple Language Question Answering

Part V. Cross-Language Retrieval in Image Collections

Part VI. Cross-Language Spoken Document Retrieval

Part VII. Issues in CLIR and in Evaluation

What Happened in CLEF 2004?

Carol Peters

Istituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR), Pisa, Italy
carol.peters@isti.cnr.it

Abstract. The organization of the CLEF 2004 evaluation campaign is described and details are provided concerning the tracks, test collections and participation. Information on new activities for CLEF 2005 is also given.

1 Introduction

This volume reports the results of the fifth in a series of annual system evaluation campaigns organised by the Cross-Language Evaluation Forum (CLEF)[1]. The main objectives of CLEF are (i) to provide an infrastructure that facilitates testing of all kinds of multilingual information access systems – from monolingual retrieval for multiple languages to the implementation of complete multilingual multimedia search services, and (ii) to construct test-suites of reusable data that can be used for benchmarking purposes. These objectives are achieved through the organisation of evaluation campaigns that culminate each year in a workshop in which the groups that participated in the campaign can report and discuss their experiments. An additional aim of CLEF is to encourage contacts between the R&D and the application communities and promote the industrial take-up of research results.

The main features of the 2004 campaign are briefly outlined below in order to provide the necessary background to the experiments reported in these post-campaign proceedings.

2 Tracks and Tasks in CLEF 2004

In recent years, CLEF distinguished between the core tracks, which were those offered regularly each campaign (the monolingual, bilingual, multilingual and domain-specific tracks), and additional tracks, which were organised on an experimental basis with the objective of identifying new requirements and appropriate methodologies for their testing in a cross-language context. This distinction no longer held in 2004. The interactive track, run since 2001, was finally recognised as part of the main activity, and the great success of the pilot tracks in CLEF 2003, and in particular the cross-language question answering and image retrieval activities, led to their inclusion as regular tracks. This meant that the scope of CLEF 2004 was considerably widened with respect to previous years, with much attention being given to tasks involving information extraction (question answering) and retrieval from multimedia.

[1] CLEF 2004 is included in the activities of the DELOS Network of Excellence on Digital Libraries, funded by the Sixth Framework Programme of the European Commission. DELOS is an "old" friend of CLEF, having promoted the first two campaigns in 2000 and 2001. For information on DELOS, see www.delos.info.

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 1 – 9, 2005.

CLEF 2004 thus offered six tracks designed to evaluate the performance of systems for:

- mono-, bi- and multilingual document retrieval on news collections (Ad hoc)
- mono- and cross-language domain-specific retrieval (GIRT)
- interactive cross-language retrieval (iCLEF)
- multiple language question answering (QA@CLEF)
- cross-language retrieval on image collections (ImageCLEF)
- cross-language spoken document retrieval (CL-SDR)

The organisation of each of these tracks and the results obtained are described and commented in the track overviews at the beginning of each section of this volume.

3 CLEF 2004 Test Collections

CLEF campaigns adopt a comparative evaluation approach in which system performance is measured using appropriate test collections. The test collections consist of sets of sample query statements often called "topics", document collections, and relevance judgments determining the set of relevant documents in a collection for a given query statement. All language dependent tasks such as topic/question creation and relevance assessment are performed in a distributed setting by native speakers. Rules are established and a tight central coordination is maintained in order to ensure consistency and coherency of topic and relevance judgment sets over the different collections, languages and tracks.

Five separate document collections were used in CLEF 2004:

- CLEF multilingual comparable corpus
- GIRT social science database
- St Andrews historical photographic archive
- CasImage radiological medical database
- Speech transcriptions supplied by TREC

The main CLEF collection is the multilingual comparable corpus of newspaper and news agency documents. In 2004, this collection was used by the Ad hoc, iCLEF and QA@CLEF tracks. The multilingual corpus increases in size and coverage every year as new collections and languages are added. In 2004 it contained nearly 1.8 million news documents from the same time period (1994-1995) in ten languages: Dutch, English, Finnish, French, German, Italian, Portuguese[2], Russian, Spanish and Swedish. Table 1 gives the main specifics of this collection and Table 2 shows which part of this collection was used in which track in 2004.

However, news media have characteristics which may not hold true for other genres: e.g. wide use of proper nouns (names and places), association of date stamps, particular style of writing and a rapid evolution of general-purpose vocabulary. Certain features may facilitate access and retrieval, others may hinder it. For this reason, a separate document collection is included for systems tuned for domain-specific tasks.

[2] Portuguese was a new addition in 2004; Hungarian and Bulgarian newspaper collections are being added for 2005.

Table 1. Sources and dimensions of the main CLEF 2004 multilingual document collection

Collection	Added in	Size (MB)	No. of Docs	Median Size of Docs. (Bytes)	Median Size of Docs. (Tokens)[3]	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 Rundschau94	2000	320	139715	1598	225	161
German: Der Spiegel 94/9	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: AGZ 94	2001	86	50527	1454	187	129
Italian: AGZ 95	2003	85	48980	1474	192	132
Portuguese: Público 94	2004	164	51751	NA	NA	NA
Portuguese: Público 95	2004	176	55070	NA	NA	NA
Russian: Izvestia 95	2003	68	16761	NA	NA	NA
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 newspaper)
NA = Not Available at this moment

[3] 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.