

*Commenced Publication in 1973*

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

*Lancaster University, UK*

Takeo Kanade

*Carnegie Mellon University, Pittsburgh, PA, USA*

Josef Kittler

*University of Surrey, Guildford, UK*

Jon M. Kleinberg

*Cornell University, Ithaca, NY, USA*

Friedemann Mattern

*ETH Zurich, Switzerland*

John C. Mitchell

*Stanford University, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

Oscar Nierstrasz

*University of Bern, Switzerland*

C. Pandu Rangan

*Indian Institute of Technology, Madras, India*

Bernhard Steffen

*University of Dortmund, Germany*

Madhu Sudan

*Massachusetts Institute of Technology, MA, USA*

Demetri Terzopoulos

*New York University, NY, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Moshe Y. Vardi

*Rice University, Houston, TX, USA*

Gerhard Weikum

*Max-Planck Institute of Computer Science, Saarbruecken, Germany*

Carol Peters Julio Gonzalo  
Martin Braschler Michael Kluck (Eds.)

# Comparative Evaluation of Multilingual Information Access Systems

4th Workshop of the  
Cross-Language Evaluation Forum, CLEF 2003  
Trondheim, Norway, August 21-22, 2003  
Revised Papers

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

Julio Gonzalo

Universidad Nacional de Educación a Distancia (UNED)

Departamento de Lenguajes y Sistemas Informáticos

c/ Juan del Rosal, 16, 28040 Madrid, Spain

E-mail: julio@lsi.uned.es

Martin Braschler

Eurospider Information Technology AG

Schaffhauser Str. 18, 8006 Zürich, Switzerland

E-mail: martin.braschler@eurospider.com

Michael Kluck

Informationszentrum Sozialwissenschaften der

Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (IZ)

Lennéstr. 30, 53113 Bonn, Germany

E-mail: kluck@bonn.iz-soz.de

Library of Congress Control Number: 2004115726

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

ISSN 0302-9743

ISBN 3-540-24017-9 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2004

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper SPIN: 11342052 06/3142 5 4 3 2 1 0

# Preface

The fourth campaign of the Cross-language Evaluation Forum (CLEF) for European languages was held from January to August 2003. Participation in this campaign showed a slight rise in the number of participants from the previous year, with 42 groups submitting results for one or more of the different tracks (compared with 37 in 2002), but a steep rise in the number of experiments attempted. A distinctive feature of CLEF 2003 was the number of new tracks and tasks that were offered as pilot experiments. The aim was to try out new ideas and to encourage the development of 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.

The campaign culminated in a two-day workshop held in Trondheim, Norway, 21–22 August, immediately following the 7th European Conference on Digital Libraries (ECDL 2003), and attended by more than 70 researchers and system developers. The objective of the workshop was to bring together the groups that had participated in the CLEF 2003 campaign so that they could report on the results of their experiments. Attendance at the workshop was thus limited to participants in the campaign plus several invited guests with recognized expertise in the multilingual information access field. This volume contains thoroughly revised and expanded versions of the preliminary papers presented at the workshop accompanied by detailed analyses of the results in the various track overview papers and thus provides an exhaustive record of the CLEF 2003 campaign.

CLEF 2003 was conducted within the framework of a project of the Information Society Technologies programme of the European Commission (IST-2000-31002). The campaign was organized in collaboration with the US National Institute of Standards and Technology (NIST) and with the support of the DELOS Network of Excellence for Digital Libraries. The support of NIST and DELOS in the running of the evaluation campaign is gratefully acknowledged. We would also like to thank the other members of the Workshop Steering Committee for their assistance in the coordination of this event.

June 2004

Carol Peters  
Julio Gonzalo  
Martin Braschler  
Michael Kluck

## **CLEF 2003 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, Departamento de 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  
Carol Peters, ISTI-CNR, Pisa, Italy  
Mark Sanderson, University of Sheffield, UK  
Peter Schäuble, Eurospider Information Technology, Switzerland  
Ellen Voorhees, National Institute of Standards and Technology, USA

# Table of Contents

Introduction	
<i>Julio Gonzalo, Carol Peters</i> .....	1
CLEF 2003 Methodology and Metrics	
<i>Martin Braschler, Carol Peters</i> .....	7
Analysis of the Reliability of the Multilingual Topic Set for the Cross Language Evaluation Forum	
<i>Thomas Mandl, Christa Womser-Hacker</i> .....	21
Evaluation of Information Access Technologies at the NTCIR Workshop	
<i>Noriko Kando</i> .....	29
<b>Part I. Ad-hoc Text Retrieval Tracks</b>	
CLEF 2003 – Overview of Results	
<i>Martin Braschler</i> .....	44
<b>Mainly Cross-Language Experiments</b>	
Report on CLEF-2003 Multilingual Tracks	
<i>Jacques Savoy</i> .....	64
The Impact of Word Normalization Methods and Merging Strategies on Multilingual IR	
<i>Eija Airio, Heikki Keskustalo, Turid Hedlund, Ari Pirkola</i> .....	74
JHU/APL Experiments in Tokenization and Non-word Translation	
<i>Paul McNamee, James Mayfield</i> .....	85
Report on CLEF-2003 Experiments: Two Ways of Extracting Multilingual Resources from Corpora	
<i>Nicola Cancedda, Hervé Déjean, Éric Gaussier, Jean-Michel Renders, Alexei Vinokourov</i> .....	98
Combining Query Translation and Document Translation in Cross-Language Retrieval	
<i>Aitao Chen, Fredric C. Gey</i> .....	108
Cross-Language Experiments with the IR-n System	
<i>Fernando Llopis, Rafael Muñoz</i> .....	122

Multilingual Information Retrieval Using Open, Transparent Resources in CLEF 2003 <i>Monica Rogati, Yiming Yang</i> .....	133
ITC-irst at CLEF 2003: Monolingual, Bilingual and Multilingual Information Retrieval <i>Nicola Bertoldi, Marcello Federico</i> .....	140
Language-Dependent and Language-Independent Approaches to Cross-Lingual Text Retrieval <i>Jaap Kamps, Christof Monz, Maarten de Rijke, Börkur Sigurbjörnsson</i> .....	152
Multilingual Retrieval Experiments with MIMOR at the University of Hildesheim <i>René Hackl, Ralph Kölle, Thomas Mandl, Alexandra Ploedt, Jan-Hendrik Scheufen, Christa Womser-Hacker</i> .....	166
Concept-Based Searching and Merging for Multilingual Information Retrieval: First Experiments at CLEF 2003 <i>Romaric Besançon, Gaël de Chalendar, Olivier Ferret, Christian Fluhr, Olivier Mesnard, Hubert Naets</i> .....	174
CLEF 2003 Experiments at UB: Automatically Generated Phrases and Relevance Feedback for Improving CLIR <i>Miguel E. Ruiz</i> .....	185
SINAI at CLEF 2003: Decompounding and Merging <i>Fernando Martínez-Santiago, Arturo Montejo-Ráez, Luis Alfonso Ureña-López, M. Carlos Díaz-Galiano</i> .....	192
Merging Results by Predicted Retrieval Effectiveness <i>Wen-Cheng Lin, Hsin-Hsi Chen</i> .....	202
MIRACLE Approaches to Multilingual Information Retrieval: A Baseline for Future Research <i>José L. Martínez, Julio Villena, Jorge Fombella, Ana G. Serrano, Paloma Martínez, José M. Goñi, José C. González</i> .....	210
Experiments to Evaluate Probabilistic Models for Automatic Stemmer Generation and Query Word Translation <i>Giorgio M. Di Nunzio, Nicola Ferro, Massimo Melucci, Nicola Orio</i> .....	220

Clairvoyance CLEF-2003 Experiments <i>Yan Qu, Gregory Grefenstette, David A. Evans</i> .....	236
Simple Translations of Monolingual Queries Expanded through an Association Thesaurus. X-IOTA IR system for CLIPS Bilingual Experiments <i>Gilles Sérasset, Jean-Pierre Chevallet</i> .....	242
Two-Stage Refinement of Query Translation in a Pivot Language Approach to Cross-Lingual Information Retrieval: An Experiment at CLEF 2003 <i>Kazuaki Kishida, Noriko Kando</i> .....	253
Regular Sound Changes for Cross-Language Information Retrieval <i>Michael P. Oakes, Souvik Banerjee</i> .....	263
Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual, Bilingual and Multilingual Retrieval <i>Adenike M. Lam-Adesina, Gareth J. F. Jones</i> .....	271
<b>Monolingual Experiments</b>	
Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer <sup>TM</sup> at CLEF 2003 <i>Stephen Tomlinson</i> .....	286
Océ at CLEF 2003 <i>Roel Brand, Marvin Brünner, Samuel Driessen, Pascha Iljin, Jakob Klok</i> .....	301
Comparing Weighting Models for Monolingual Information Retrieval <i>Gianni Amati, Claudio Carpineto, Giovanni Romano</i> .....	310
Pruning Texts with NLP and Expanding Queries with an Ontology: TagSearch <i>Gil Francopoulo</i> .....	319
Report on CLEF-2003 Monolingual Tracks: Fusion of Probabilistic Models for Effective Monolingual Retrieval <i>Jacques Savoy</i> .....	322
Selective Compound Splitting of Swedish Queries for Boolean Combinations of Truncated Terms <i>Rickard Cöster, Magnus Sahlgren, Jussi Karlgren</i> .....	337



COLE Experiments at CLEF 2003 in the Spanish Monolingual Track  
*Jesús Vilares, Miguel A. Alonso, Francisco J. Ribadas*..... 345

Experiments with Self Organizing Maps in CLEF 2003  
*Javier Fernández, Ricardo Mones, Irene Díaz, José Ranilla,  
 Elías F. Combarro* ..... 358

Ricoh at CLEF 2003  
*Yuichi Kojima, Hideo Itoh, Hiroko Mano, Yasushi Ogawa* ..... 367

MediaLab at CLEF 2003: Using Keyword Disambiguation  
*Peter van der Weerd* ..... 373

**Part II. Domain-Specific Document Retrieval**

The GIRT Data in the Evaluation of CLIR Systems - from 1997  
 Until 2003  
*Michael Kluck*..... 376

A Data-Compression Approach to the Monolingual GIRT Task:  
 An Agnostic Point of View  
*Daniela Alderuccio, Luciana Bordonni, Vittorio Loreto* ..... 391

UC Berkeley at CLEF-2003 – Russian Language Experiments and  
 Domain-Specific Retrieval  
*Vivien Petras, Natalia Perelman, Fredric Gey*..... 401

University of Hagen at CLEF 2003: Natural Language Access to the  
 GIRT4 Data  
*Johannes Leveling*..... 412

**Part III. Interactive Cross-Language Retrieval**

The CLEF 2003 Interactive Track  
*Douglas W. Oard, Julio Gonzalo*..... 425

iCLEF 2003 at Maryland: Translation Selection and Document Selection  
*Bonnie Dorr, Daqing He, Jun Luo, Douglas W. Oard,  
 Richard Schwartz, Jianqiang Wang, David Zajic* ..... 435

UNED at iCLEF 2003: Searching Cross-Language Summaries  
*Fernando López-Ostenero, Julio Gonzalo, Felisa Verdejo*..... 450

Comparing Syntactic-Semantic Patterns and Passages in Interactive Cross Language Information Access <i>Borja Navarro, Fernando Llopis, Miguel Ángel Varó</i> .....	462
Continued Experiments on Cross-Language Relevance Assessment <i>Jussi Karlgren, Preben Hansen</i> .....	468
<b>Part IV. Cross-Language Question Answering</b>	
The Multiple Language Question Answering Track at CLEF 2003 <i>Bernardo Magnini, Simone Romagnoli, Alessandro Vallin, Jesús Herrera, Anselmo Peñas, Víctor Peinado, Felisa Verdejo, Maarten de Rijke</i> .....	471
Creating the DISEQuA Corpus: A Test Set for Multilingual Question Answering <i>Bernardo Magnini, Simone Romagnoli, Alessandro Vallin, Jesús Herrera, Anselmo Peñas, Víctor Peinado, Felisa Verdejo, Maarten de Rijke</i> .....	487
Bridging Languages for Question Answering: DIOGENE at CLEF 2003 <i>Matteo Negri, Hristo Tanev, Bernardo Magnini</i> .....	501
Cross-Language Question Answering at the USC Information Sciences Institute <i>Abdessamad Echihabi, Douglas W. Oard, Daniel Marcu, Ulf Hermjakob</i> .....	514
How Frogs Built the Berlin Wall. A Detailed Error Analysis of a Question Answering System for Dutch <i>Valentin Jijkoun, Gilad Mishne, Maarten de Rijke</i> .....	523
Cross-Lingual QA: A Modular Baseline in CLEF 2003 <i>Lucian Vlad Lita, Monica Rogati, Jaime Carbonell</i> .....	535
Question Answering in Spanish <i>José L. Vicedo, Ruben Izquierdo, Fernando Llopis, Rafael Muñoz</i> ....	541
Quantum, a French/English Cross-Language Question Answering System <i>Luc Plamondon, George Foster</i> .....	549
A Cross-Language Question/Answering-System for German and English <i>Günter Neumann, Bogdan Sacaleanu</i> .....	559

Cross-Language French-English Question Answering Using the DLT System at CLEF 2003  
*Richard F. E. Sutcliffe, Igal Gabbay, Aoife O’Gorman* ..... 572

**Part V. Cross-Language Image Retrieval**

The CLEF 2003 Cross Language Image Retrieval Track  
*Paul Clough, Mark Sanderson* ..... 581

Assessing Translation Quality for Cross Language Image Retrieval  
*Paul Clough, Mark Sanderson* ..... 594

Foreign Name Backward Transliteration in Chinese-English Cross-Language Image Retrieval  
*Wen-Cheng Lin, Changhua Yang, Hsin-Hsi Chen* ..... 611

Image Retrieval: the MIRACLE Approach  
*Julio Villena, José L. Martínez, Jorge Fombella, Ana G. Serrano, Alberto Ruiz, Paloma Martínez, José M. Goñi, José C. González* ..... 621

Scene of Crime Information System: Playing at St. Andrews  
*Bogdan Vrusias, Mariam Tariq, Lee Gillam* ..... 631

**Part VI. Cross-Language Spoken Document Retrieval**

The CLEF 2003 Cross-Language Spoken Document Retrieval Track  
*Marcello Federico, Gareth J. F. Jones* ..... 646

Exeter at CLEF 2003: Cross-Language Spoken Document Retrieval Experiments  
*Gareth J. F. Jones, Adenike Lam-Adesina* ..... 653

N-grams for Translation and Retrieval in CL-SDR  
*Paul McNamee, James Mayfield* ..... 658

Spoken Document Retrieval Experiments with IR-n System  
*Fernando Llopis, Patricio Martínez-Barco* ..... 664

ITC-irst at CLEF 2003: Cross-Language Spoken Document Retrieval  
*Nicola Bertoldi, Marcello Federico* ..... 672

**Appendix** ..... 677

**Author Index** ..... 701