

**Springer-Verlag
Berlin Heidelberg GmbH**

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

Advances in Cross-Language Information Retrieval

Third Workshop of the
Cross-Language Evaluation Forum, CLEF 2002
Rome, Italy, September 19-20, 2002
Revised 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

Martin Braschler

Eurospider Information Technology AG
Schaffhauserstr. 18, 8006 Zürich, Switzerland
E-mail: martin.braschler@eurospider.com

Julio Gonzalo

Universidad Nacional de Educación a Distancia
Lenguajes y Sistemas Informáticos
Ciudad Universitaria, 28040 Madrid, Spain
E-mail: julio@lsi.uned.es

Michael Kluck

Informationszentrum Sozialwissenschaften der
Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (IZ)
Lennéstr. 30, 53113 Bonn, Germany
E-mail: kluck@bonn.iz-soz.de

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie;
detailed bibliographic data is available in the Internet at <<http://dnb.ddb.de>>.

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

ISSN 0302-9743

ISBN 978-3-540-40830-7

ISBN 978-3-540-45237-9 (eBook)

DOI 10.1007/978-3-540-45237-9

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-Verlag. Violations are liable for prosecution under the German Copyright Law.

© Springer-Verlag Berlin Heidelberg 2003

Originally published by Springer-Verlag Berlin Heidelberg New York in 2003

Typesetting: Camera-ready by author, data conversion by DA-TeX Gerd Blumenstein
Printed on acid-free paper SPIN: 10931776 06/3142 5 4 3 2 1 0

Preface

The third campaign of the Cross-Language Evaluation Forum (CLEF) for European languages was held from January to September 2002. Participation in this campaign showed a slight rise in the number of participants with, 37 groups from both academia and industry and a steep rise in the number of experiments they submitted for one or more of the five official tracks. The campaign culminated in a two-day workshop held in Rome, Italy, 19–20 September, immediately following the Sixth European Conference on Digital Libraries (ECDL 2002), attended by nearly 70 researchers and system developers. The objective of the workshop was to bring together the groups that had participated in CLEF 2002 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 a complete run-down and detailed analysis of the results, and it thus provides an exhaustive record of the CLEF 2002 campaign.

CLEF 2002 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 2003

Carol Peters
Martin Braschler
Julio Gonzalo
Michael Kluck

CLEF 2002 Workshop Steering Committee

Martin Braschler	Eurospider Information Technology AG, Switzerland
Khalid Choukri	Evaluations and Language Resources Distribution Agency, France
Julio Gonzalo Arroyo	Universidad Nacional de Educación a Distancia, Spain
Donna Harman	National Institute of Standards and Technology, USA
Noriko Kando	National Institute of Informatics, Japan
Michael Kluck	IZ Sozialwissenschaften, Bonn, Germany
Patrick Kremer	Institute de Information Scientifique et Technique, Vandoevre, France
Carol Peters	Italian National Research Council, Pisa, Italy
Peter Schäuble	Eurospider Information Technology AG, Switzerland
Laurent Schmitt	Institute de Information Scientifique et Technique, Vandoevre, France
Ellen Voorhees	National Institute of Standards and Technology, USA

Table of Contents

Introduction	
<i>C. Peters</i>	1

I System Evaluation Experiments at CLEF 2002

CLEF 2002 – Overview of Results	
<i>M. Braschler</i>	9

Cross-Language and More

Cross-Language Retrieval Experiments at CLEF 2002	
<i>A. Chen</i>	28

ITC-irst at CLEF 2002:

Using *N*-Best Query Translations for CLIR

<i>N. Bertoldi and M. Federico</i>	49
--	----

Océ at CLEF 2002

<i>R. Brand and M. Brünner</i>	59
--------------------------------------	----

Report on CLEF 2002 Experiments:

Combining Multiple Sources of Evidence

<i>J. Savoy</i>	66
-----------------------	----

UTACLIR @ CLEF 2002 –

Bilingual and Multilingual Runs with a Unified Process

<i>E. Airio, H. Keskustalo, T. Hedlund, and A. Pirkola</i>	91
--	----

A Multilingual Approach to Multilingual Information Retrieval

<i>J.-Y. Nie and F. Jin</i>	101
-----------------------------------	-----

Combining Evidence for Cross-Language Information Retrieval

<i>J. Kamps, C. Monz, and M. de Rijke</i>	111
---	-----

Exeter at CLEF 2002:

Experiments with Machine Translation for Monolingual and Bilingual Retrieval

<i>A.M. Lam-Adesina and G.J.F. Jones</i>	127
--	-----

Portuguese-English Experiments Using Latent Semantic Indexing

<i>V.M. Orenco and C. Huyck</i>	147
---------------------------------------	-----

Thomson Legal and Regulatory Experiments for CLEF 2002 <i>I. Moulinier and H. Molina-Salgado</i>	155
Eurospider at CLEF 2002 <i>M. Braschler, A. Göhring, and P. Schäuble</i>	164
Merging Mechanisms in Multilingual Information Retrieval <i>W.-C. Lin and H.-H. Chen</i>	175
SINAI at CLEF 2002: Experiments with Merging Strategies <i>F. Martínez, L.A. Ureña, and M.T. Martín</i>	187
Cross-Language Retrieval at the University of Twente and TNO <i>D. Reidsma, D. Hiemstra, F. de Jong, and W. Kraaij</i>	197
Scalable Multilingual Information Access <i>P. McNamee and J. Mayfield</i>	207
Some Experiments with the Dutch Collection <i>A.P. de Vries and A. Diekema</i>	219
Resolving Translation Ambiguity Using Monolingual Corpora <i>Y. Qu, G. Grefenstette, and D.A. Evans</i>	223

Monolingual Experiments

Experiments in 8 European Languages with Hummingbird SearchServer™ at CLEF 2002 <i>S. Tomlinson</i>	242
Italian Monolingual Information Retrieval with PROSIT <i>G. Amati, C. Carpineto, and G. Romano</i>	257
COLE Experiments in the CLEF 2002 Spanish Monolingual Track <i>J. Vilares, M.A. Alonso, F.J. Ribadas, and M. Vilares</i>	265
Improving the Automatic Retrieval of Text Documents <i>M. Agosti, M. Bacchin, N. Ferro, and M. Melucci</i>	279
IR-n System at CLEF-2002 <i>F. Llopis, J.L. Vicedo, and A. Ferrández</i>	291
Experiments in Term Expansion Using Thesauri in Spanish <i>Á.F. Zazo, C.G. Figuerola, J.L.A. Berrocal, E. Rodríguez, and R. Gómez</i>	301
SICS at CLEF 2002: Automatic Query Expansion Using Random Indexing <i>M. Sahlgren, J. Karlgren, R. Cöster, and T. Järvinen</i>	311

Pliers and Snowball at CLEF 2002	
<i>A. MacFarlane</i>	321
Experiments with a Chunker and Lucene	
<i>G. Francopoulo</i>	336
Information Retrieval with Language Knowledge	
<i>E. Dura and M. Drejak</i>	338

Mainly Domain-Specific Information Retrieval

Domain Specific Retrieval Experiments with MIMOR at the University of Hildesheim	
<i>R. Hackl, R. Kölle, T. Mandl, and C. Womser-Hacker</i>	343
Using Thesauri in Cross-Language Retrieval of German and French Indexed Collections	
<i>V. Petras, N. Perelman, and F. Gey</i>	349
Assessing Automatically Extracted Bilingual Lexicons for CLIR in Vertical Domains: XRCE Participation in the GIRT Track of CLEF 2002	
<i>J.-M. Renders, H. Déjean, and É. Gaussier</i>	363

Interactive Track

The CLEF 2002 Interactive Track	
<i>J. Gonzalo and D.W. Oard</i>	372
SICS at iCLEF 2002:	
Cross-Language Relevance Assessment and Task Context	
<i>J. Karlgren and P. Hansen</i>	383
Universities of Alicante and Jaen at iCLEF	
<i>F. Llopis, J.L. Vicedo, A. Ferrández, M.C. Díaz, and F. Martínez</i>	392
Comparing User-Assisted and Automatic Query Translation	
<i>D. He, J. Wang, D.W. Oard, and M. Nossal</i>	400
Interactive Cross-Language Searching:	
Phrases Are Better than Terms for Query Formulation and Refinement	
<i>F. López-Ostenero, J. Gonzalo, A. Peñas, and F. Verdejo</i>	416
Exploring the Effect of Query Translation when Searching Cross-Language	
<i>D. Petrelli, G. Demetriou, P. Herring, M. Beaulieu, and M. Sanderson</i> ...	430

Cross-Language Spoken Document Retrieval

CLEF 2002 Cross-Language Spoken Document Retrieval Pilot Track Report <i>G.J.F. Jones and M. Federico</i>	446
Exeter at CLEF 2002: Cross-Language Spoken Document Retrieval Experiments <i>G.J.F. Jones and A.M. Lam-Adesina</i>	458
Cross-Language Spoken Document Retrieval on the TREC SDR Collection <i>N. Bertoldi and M. Federico</i>	476

II Cross-Language Systems Evaluation Initiatives, Issues and Results

CLIR at NTCIR Workshop 3: Cross-Language and Cross-Genre Retrieval <i>N. Kando</i>	485
Linguistic and Statistical Analysis of the CLEF Topics <i>T. Mandl and C. Womser-Hacker</i>	505
CLEF 2002 Methodology and Metrics <i>M. Braschler and C. Peters</i>	512

III Appendix

List of Run Characteristics	529
Overview Graphs	533
Multilingual Runs	544
Bilingual to German Runs	580
Bilingual to English Runs	593
Bilingual to Spanish Runs	609
Bilingual to Finnish Runs	625
Bilingual to French Runs	627
Bilingual to Italian Runs	641
Bilingual to Dutch Runs	655
Bilingual to Swedish Runs	664

Monolingual German Runs	665
Monolingual Spanish Runs	666
Monolingual Finnish Runs	714
Monolingual French Runs	725
Monolingual Italian Runs	741
Monolingual Dutch Runs	766
Monolingual Swedish Runs	785
AMARYLLIS Domain-Specific Runs	794
GIRT Domain-Specific Runs	809
Author Index	827