

Lecture Notes in Computer Science

2406

Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer

Berlin

Heidelberg

New York

Barcelona

Hong Kong

London

Milan

Paris

Tokyo

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

Evaluation of Cross-Language Information Retrieval Systems

Second Workshop of the
Cross-Language Evaluation Forum, CLEF 2001
Darmstadt, Germany, September 3-4, 2001
Revised Papers



Springer

Volume Editors

Carol Peters

Istituto di Scienza e Tecnologie

Consiglio Nazionale delle Ricerche

Via G. Moruzzi 1, 56124 Pisa, Italy

E-mail: carol@iei.pi.cnr.it

Martin Braschler

Eurospider Information Technology AG

Schaffhauserstrasse 18, 8006 Zürich, Switzerland

e-mail: martin.braschler@eurospider.com

Julio Gonzalo

Universidad Nacional de Educación a Distancia

E.T.S.I. Industriales, Ciudad Universitaria s/n

28040 Madrid, Spain

e-mail: julio@lsi.uned.es

Michael Kluck

InformationsZentrum Sozialwissenschaften

Lennestr. 30, 53113 Bonn, Germany

e-mail: Kluck@bonn.iz-soz.de

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Evaluation of cross language information retrieval systems : revised papers

/ Second workshop of the Cross Language Evaluation Forum, CLEF 2001,

Darmstadt, Germany, September 3 - 4, 2001. Carol Peters ... (ed.). -

Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ;

Paris ; Tokyo : Springer, 2002

(Lecture notes in computer science ; Vol. 2406)

ISBN 3-540-44042-9

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

ISSN 0302-9743

ISBN 3-540-44042-9 Springer-Verlag 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-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York,

a member of BertelsmannSpringer Science+Business Media GmbH

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2002

Printed in Germany

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Stefan Sossna e. K.

Printed on acid-free paper SPIN: 10873667 06/3142 5 4 3 2 1 0

Preface

The second evaluation campaign of the Cross-Language Evaluation Forum (CLEF) for European languages was held from January to September 2001. This campaign proved a great success, and showed an increase in participation of around 70% compared with CLEF 2000. It culminated in a two-day workshop in Darmstadt, Germany, 3–4 September, in conjunction with the 5th European Conference on Digital Libraries (ECDL 2001). On the first day of the workshop, the results of the CLEF 2001 evaluation campaign were reported and discussed in paper and poster sessions. The second day focused on the current needs of cross-language systems and how evaluation campaigns in the future can best be designed to stimulate progress. The workshop was attended by nearly 50 researchers and system developers from both academia and industry. It provided an important opportunity for researchers working in the same area to get together and exchange ideas and experiences. Copies of all the presentations are available on the CLEF web site at <http://www.clef-campaign.org>. This volume contains thoroughly revised and expanded versions of the papers presented at the workshop and provides an exhaustive record of the CLEF 2001 campaign.

CLEF 2001 was conducted as an activity of the DELOS Network of Excellence for Digital Libraries, funded by the EC Information Society Technologies program to further research in digital library technologies. The activity was 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. We would also like to thank the other members of the Workshop Steering Committee for their assistance in the organization of this event.

April 2002

Carol Peters
Martin Braschler
Julio Gonzalo
Michael Kluck

CLEF 2001 Workshop Steering Committee

Martin Braschler, Eurospider Information Technology AG, Switzerland
Julio Gonzalo Arroyo, UNED, Madrid, Spain
Donna Harman, National Institute of Standards and Technology (NIST), USA
Djoerd Hiemstra, University of Twente, The Netherlands
Noriko Kando, National Institute of Informatics (NII), Japan
Michael Kluck, Informationszentrum Sozialwissenschaften (IZ), Bonn, Germany
Carol Peters, IEI-CNR, Pisa, Italy
Peter Schäuble, Eurospider Information Technology AG, Switzerland
Ellen Voorhees, National Institute of Standards and Technology (NIST), USA
Christa Womser-Hacker, University of Hildesheim, Germany

Table of Contents

Introduction	1
<i>Carol Peters</i>	
Part I. System Evaluation Experiments at CLEF 2001	
CLEF 2001 - Overview of Results	9
<i>Martin Braschler</i>	
Mainly Cross-Language	
Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach	27
<i>Jacques Savoy</i>	
Multilingual Information Retrieval Using English and Chinese Queries	44
<i>Aitao Chen</i>	
Exeter at CLEF 2001: Experiments with Machine Translation for Bilingual Retrieval	59
<i>Gareth J.F. Jones, Adenike M. Lam-Adesina</i>	
TNO at CLEF-2001: Comparing Translation Resources	78
<i>Wessel Kraaij</i>	
ITC-irst at CLEF 2001: Monolingual and Bilingual Tracks	94
<i>Nicola Bertoldi, Marcello Federico</i>	
Experiments with the Eurospider Retrieval System for CLEF 2001	102
<i>Martin Braschler, Bärbel Ripplinger, Peter Schäuble</i>	
Using Co-occurrence, Augmented Restrictions, and C-E WordNet for Chinese-English Cross-Language Information Retrieval at CLEF 2001	111
<i>Wen-Cheng Lin, Hsin-Hsi Chen</i>	
Utaclir @ CLEF 2001 - Effects of Compound Splitting and N-Gram Techniques	118
<i>Turid Hedlund, Heikki Keskustalo, Ari Pirkola, Eija Airio, Kalervo Järvelin</i>	
Using Statistical Translation Models for Bilingual IR	137
<i>Jian-Yun Nie, Michel Simard</i>	
Cross-Lingual Pseudo-Relevance Feedback Using a Comparable Corpus . . .	151
<i>Monica Rogati, Yiming Yang</i>	

Investigation on Disambiguation in CLIR: Aligned Corpus and Bi-directional Translation-Based Strategies	158
<i>M. Boughanem, C. Christment, N. Nassr</i>	
Vector-Based Semantic Analysis Using Random Indexing for Cross-Lingual Query Expansion	169
<i>Magnus Sahlgren, Jussi Karlgren</i>	
Query Expansion Techniques for the CLEF Bilingual Track	177
<i>Fatiha Sadat, Akira Maeda, Masatoshi Yoshikawa, Shunsuke Uemura</i>	
Intelligent Information Access Systems (SINAI) at CLEF 2001: Calculating Translation Probabilities with SemCor	185
<i>Fernando Martínez-Santiago, L. Alfonso Ureña-López, M. Carlos Díaz-Galiano, Maite Martín-Valdivia, Manuel García-Vega</i>	
JHU/APL Experiments at CLEF: Translation Resources and Score Normalization	193
<i>Paul McNamee, James Mayfield</i>	
Dictionary-Based Thai CLIR: An Experimental Survey of Thai CLIR.....	209
<i>Jaruskulchai Chuleerat</i>	
English-Dutch CLIR Using Query Translation Techniques	219
<i>Mirna Adriani</i>	
Thomson Legal and Regulatory at CLEF 2001: Monolingual and Bilingual Experiments	226
<i>Hugo Molina-Salgado, Isabelle Moulinier, Mark Knudson, Elizabeth Lund, Kirat Sekhon</i>	
Working with Russian Queries for the GIRT, Bilingual, and Multilingual CLEF Tasks	235
<i>Fredric C. Gey, Hailing Jiang, Natalia Perelman</i>	
IR-n: A Passage Retrieval System at CLEF-2001	244
<i>Fernando Llopis, José L. Vicedo</i>	
Monolingual Experiments	
Spanish Monolingual Track: The Impact of Stemming on Retrieval	253
<i>Carlos G. Figuerola, Raquel Gómez, Angel F. Zazo Rodríguez, José Luis Alonso Berrocal</i>	
Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German, and Italian	262
<i>Christof Monz, Maarten de Rijke</i>	

Stemming Evaluated in 6 Languages by Hummingbird SearchServer™ at CLEF 2001	278
<i>Stephen Tomlinson</i>	
Minimalistic Test Runs of the Eidetica Indexer	288
<i>Teresita Frizzarin, Annius Groenink</i>	
Across the Bridge: CLEF 2001 - Non-english Monolingual Retrieval. The French Task	291
<i>Eugenia Matoyo, Tony Valsamidis</i>	
Mpro-IR in CLEF 2001	300
<i>Bärbel Ripplinger</i>	
Some Terms Are More Interchangeable than Others	303
<i>Jakob Klok, Samuel Driessen, Marvin Brunner</i>	
Interactive Track	
The CLEF 2001 Interactive Track	308
<i>Douglas W. Oard, Julio Gonzalo</i>	
Noun Phrase Translations for Cross-Language Document Selection	320
<i>Fernando López-Ostenero, Julio Gonzalo, Anselmo Peñas, Felisa Verdejo</i>	
iCLEF at Sheffield	332
<i>Mark Sanderson, Zoë Bathie</i>	
iCLEF 2001 at Maryland: Comparing Term-for-Term Gloss and MT	336
<i>Jianqiang Wang, Douglas W. Oard</i>	
Part II. Evaluation Issues and Results	
The Philosophy of Information Retrieval Evaluation	355
<i>Ellen M. Voorhees</i>	
CLIR System Evaluation at the Second NTCIR Workshop	371
<i>Noriko Kando</i>	
Multilingual Topic Generation within the CLEF 2001 Experiments	389
<i>Christa Womser-Hacker</i>	
CLEF Methodology and Metrics	394
<i>Martin Braschler, Carol Peters</i>	
Appendix	
List of Run Characteristics	405
Multilingual Runs	409

Bilingual to English Runs 435

Bilingual to Dutch Runs 496

Monolingual Dutch Runs 499

Monolingual French Runs 518

Monolingual German Runs 536

Monolingual Italian Runs 561

Monolingual Spanish Runs 575

GIRT Domain-Specific Runs..... 597

Author Index 601