

Pamela Forner Henning Müller  
Roberto Paredes Paolo Rosso  
Benno Stein (Eds.)

LNCS 8138

# Information Access Evaluation

Multilinguality, Multimodality,  
and Visualization

4th International Conference  
of the CLEF Initiative, CLEF 2013  
Valencia, Spain, September 2013, Proceedings



*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*

Alfred Kobsa

*University of California, Irvine, CA, 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

*TU Dortmund University, Germany*

Madhu Sudan

*Microsoft Research, Cambridge, MA, USA*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max Planck Institute for Informatics, Saarbruecken, Germany*

Pamela Forner Henning Müller  
Roberto Paredes Paolo Rosso  
Benno Stein (Eds.)

# Information Access Evaluation

Multilinguality, Multimodality,  
and Visualization

4th International Conference  
of the CLEF Initiative, CLEF 2013  
Valencia, Spain, September 23-26, 2013  
Proceedings

## Volume Editors

Pamela Forner

Center for the Evaluation of Language and Communication Technologies (CELCT)

Povo, Italy

E-mail: forner@celct.it

Henning Müller

University of Applied Sciences Western Switzerland, HES-SO Valais

Sierre, Switzerland

E-mail: henning.mueller@hevs.ch

Roberto Paredes

Paolo Rosso

Universitat Politècnica de València, Dept. de Sistemas Informáticos y Computación

València, Spain

E-mail: {rparedes; proso}@dsic.upv.es

Benno Stein

Bauhaus-Universität Weimar, Germany

E-mail: benno.stein@uni-weimar.de

ISSN 0302-9743

e-ISSN 1611-3349

ISBN 978-3-642-40801-4

e-ISBN 978-3-642-40802-1

DOI 10.1007/978-3-642-40802-1

Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013946902

CR Subject Classification (1998): I.7, I.2.7, H.3.1, H.3.3, H.3.7, H.4.1, H.5.3, H.2.8, I.1.3

LNCS Sublibrary: SL 3 – Information Systems and Application,  
incl. Internet/Web and HCI

© Springer-Verlag Berlin Heidelberg 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

*Typesetting:* Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

Since 2000 the Cross-Language Evaluation Forum (CLEF) has played a leading role in stimulating research and innovation in a wide range of key areas in the domain of multimodal and multilingual information access. Through the years, CLEF has promoted the study and implementation of evaluation methodologies for diverse tasks, resulting in the creation of a broad, strong and multidisciplinary research community.

Until 2010, the outcomes of the experiments carried out under the CLEF umbrella were presented and discussed at annual workshops in conjunction with the European Conference for Digital Libraries. CLEF 2010 represented a radical change from this “classic” CLEF format. While preserving CLEF’s traditional core goals, namely, benchmarking activities carried out in various tracks, we complemented these activities with a peer-reviewed conference component that aimed at advancing research in the evaluation of complex information access systems in different languages and modalities.

The theme of the CLEF conference this year is “Information Access Evaluation Meets Multilinguality, Multimodality, and Visualization”. Thus, the papers accepted for the conference included research on information access that meets these three key aspects: multilinguality, multimodality, and visualization. Two keynote speakers highlight important issues related to our field. Rada Mihalcea presents the growing need in Internet for effective solutions for multilingual natural language processing. In her talk, Rada explores the hypothesis that a multilingual representation can enrich the feature space for natural language processing tasks, and hence entail significant improvements over traditional solutions that rely exclusively on a monolingual representation. Evangelos Kanoulas presents the increasing interest in creating test collections that better model the variability encountered in real-life search scenarios in order to assess the information retrieval effectiveness. This includes experimenting over a variety of queries, corpora and even users and their interactions with the search results.

This year the overviews of the different evaluation campaigns are included in the publication in hand, while the experiments carried out by systems during the evaluation campaigns are described in a separate publication, namely, the *Working Notes*.

The success of CLEF 2013 would not have been possible without the invaluable contributions of the members of the Program Committee, Organizing Committee, students, and volunteers who supported the conference in its various

stages. We would also like to express our gratitude to the sponsoring organizations for their significant and timely support. These proceedings were prepared with the assistance of the Center for the Evaluation of Language and Communication Technologies (CELCT), Trento, Italy.

July 2013

Pamela Forner  
Henning Müller  
Roberto Paredes  
Paolo Rosso  
Benno Stein

# Organization

CLEF 2013 was organized by the Universitat Politècnica de València, Spain.

## General Chairs

Paolo Rosso	Universitat Politècnica de València, Spain
Benno Stein	Bauhaus-Universität Weimar, Germany

## Program Chairs

Henning Müller	University of Applied Sciences Western Switzerland, Switzerland
Roberto Paredes	Universitat Politècnica de València, Spain

## Evaluation Labs Chairs

Roberto Navigli	Sapienza University of Rome, Italy
Dan Tufis	RACAI, Romania

## Resource Chair

Khalid Choukri	Evaluations and Language resources Distribution Agency (ELDA), France
----------------	--

## Organization Chair

Pamela Forner	CELCT, Italy
---------------	--------------

## Organizing Committee

Centro de Formación Permanente, Universitat Politècnica de València, Spain:

Ma Francisca Collado López  
Ester Srougi Ramon

Universitat Politècnica de València, Spain

Enrique Flores  
Parth Gupta  
Mauricio Villegas

Center for the Evaluation of Language and Communication Technologies (CELCT), Italy:

Pamela Forner  
Giovanni Moretti

## Program Committee

Shlomo Argamon	Illinois Institute of Technology, USA
Alexandra Balahur	Joint Research Centre - JRC - European Commission, Italy
Barbara Caputo	IDIAP Research Institute, Switzerland
Khalid Choukri	ELDA, France
Walter Daelemans	University of Antwerp, Belgium
Nicola Ferro	University of Padua, Italy
Norbert Fuhr	University of Duisburg-Essen, Germany
Allan Hanbury	Vienna University of Technology, Austria
Donna Harman	National Institute of Standard and Technology, USA
Eduard Hovy	Carnegie Mellon University, USA
Antoine Isaac	Europeana & VU University Amsterdam, The Netherlands
Alexis Joly	Inria, France
Evangelos Kanoulas	Google, Switzerland
Birger Larsen	Royal School of Library and Information Science, Denmark
Mihai Lupu	Vienna University of Technology, Austria
Walid Magdy	Qatar Computing Research Institute, Qatar
Thomas Mandl	University of Hildesheim, Germany
Paul McNamee	Johns Hopkins University, USA
Manuel Montes-Y-Gómez	INAOE, Mexico
Jian-Yun Nie	University of Montreal, Canada
Anselmo Peñas	National Distance Learning University, Spain
Carol Peters	ISTI-CNR Pisa, Italy
Vivien Petras	Humboldt University, Germany
Florina Piroi	Vienna University of Technology, Austria
Martin Potthast	Bauhaus-Universität Weimar, Germany
Alvaro Rodrigo	National Distance Learning University, Spain
Paolo Rosso	Universitat Politècnica de València, Spain
Giuseppe Santucci	Sapienza University of Rome, Italy
Tobias Schreck	University of Konstanz, Germany
Efstathios Stamatatos	University of the Aegean, Greece



Benno Stein	Bauhaus-Universität Weimar, Germany
Bart Thomee	Yahoo! Research, Spain
Elaine Toms	University of Sheffield, UK
Theodora Tsirikla	University of Applied Sciences Western Switzerland, Switzerland
Jose-Luis Vicedo	University of Alicante, Spain
Christa Womser-Hacker	University of Hildesheim, Germany
David Zellhoefer	Brandenburg Technical University Cottbus, Germany

## Sponsoring Institutions

CLEF 2013 benefited from the support of the following organizations:

PROMISE Network of Excellence



European Science Foundation (ELIAS project)



Universitat Politècnica de València



Escola Tècnica Superior d'Enginyeria Informàtica



Departamento de Sistemas Informáticos y Computación



WIQ-EI - Web Information Quality Evaluation Initiative



Corex, Building Knowledge Solutions



ELIAS



# Keynotes

# Multilingual Natural Language Processing

Rada Mihalcea

University of North Texas, Computer Science and Engineering  
PO Box 311366, Denton, 76203 United States  
[rada@cs.unt.edu](mailto:rada@cs.unt.edu)

**Abstract.** With rapidly growing online resources, such as Wikipedia, Twitter, or Facebook, there is an increasing number of languages that have a Web presence, and correspondingly there is a growing need for effective solutions for multilingual natural language processing. In this talk, I will explore the hypothesis that a multilingual representation can enrich the feature space for natural language processing tasks, and lead to significant improvements over traditional solutions that rely exclusively on a monolingual representation. Specifically, I will describe experiments performed on three different tasks: word sense disambiguation, subjectivity analysis, and text semantic similarity, and show how the use of a multilingual representation can leverage additional information from the languages in the multilingual space, and thus improve over the use of only one language at a time.

# Comparative Evaluation Redux, or: How to Stop Worrying and Learn to Love the Variance

Evangelos Kanoulas

Google Zurich  
Brandschenkestrasse 110  
Zurich, CH-8002 Switzerland  
ekanoulas@gmail.com

**Abstract.** Information retrieval effectiveness evaluation typically takes one of three forms: batch experiments based on static test collections, lab studies measuring actual users interacting with a system, or online experiments tracking user's interactions with a live system. Test collection experiments are sometimes viewed as introducing too many simplifying assumptions to accurately predict the usefulness of a system to its users. As a result, there is great interest in creating test collections that better model the variability encountered in real-life search scenarios. This includes experimenting over a variety of queries, corpora or even users and their interactions with the search results. In this talk I will discuss how to control different aspects of batch experimentation, how to model the variance control variables introduce to measurements of effectiveness, and how to extend our statistical significance test arsenal to allow comparing retrieval algorithms.

# Table of Contents

## Evaluation and Visualization

The Scholarly Impact of CLEF (2000–2009) . . . . .	1
<i>Theodora Tsirikla, Birger Larsen, Henning Müller, Stefan Endrullis, and Erhard Rahm</i>	
A Quantitative Look at the CLEF Working Notes . . . . .	13
<i>Thomas Wilhelm-Stein and Maximilian Eibl</i>	
Building a Common Framework for IIR Evaluation . . . . .	17
<i>Mark Michael Hall and Elaine Toms</i>	
Improving Ranking Evaluation Employing Visual Analytics . . . . .	29
<i>Marco Angelini, Nicola Ferro, Giuseppe Santucci, and Gianmaria Silvello</i>	
A Proposal for New Evaluation Metrics and Result Visualization Technique for Sentiment Analysis Tasks . . . . .	41
<i>Francisco José Valverde-Albacete, Jorge Carrillo-de-Albornoz, and Carmen Peláez-Moreno</i>	
A New Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection . . . . .	53
<i>Imene Bensalem, Paolo Rosso, and Salim Chikhi</i>	
Selecting Success Criteria: Experiences with an Academic Library Catalogue . . . . .	59
<i>Paul Clough and Paula Goodale</i>	
A Dependency-Inspired Semantic Evaluation of Machine Translation Systems . . . . .	71
<i>Mohammad Reza Mirsarraf and Nazanin Dehghani</i>	
A Turing Test to Evaluate a Complex Summarization Task . . . . .	75
<i>Alejandro Molina, Eric SanJuan, and Juan-Manuel Torres-Moreno</i>	
A Formative Evaluation of a Comprehensive Search System for Medical Professionals . . . . .	81
<i>Veronika Stefanov, Alexander Sachs, Marlene Kritz, Matthias Samwald, Manfred Gschwandtner, and Allan Hanbury</i>	

## Multilinguality and Less-Resourced Languages

Exploiting Multiple Translation Resources for English-Persian Cross Language Information Retrieval . . . . .	93
<i>Hosein Azarbonyad, Azadeh Shakery, and Hesham Faily</i>	
ALQASIM: Arabic Language Question Answer Selection in Machines . . .	100
<i>Ahmed Magdy Ezzeldin, Mohamed Hamed Kholief, and Yasser El-Sonbaty</i>	
A Web-Based CLIR System with Cross-Lingual Topical Pseudo Relevance Feedback . . . . .	104
<i>Xuwen Wang, Xiaojie Wang, and Qiang Zhang</i>	
A Case Study in Decomposing for Bengali Information Retrieval . . . .	108
<i>Debasis Ganguly, Johannes Leveling, and Gareth J.F. Jones</i>	
Context-Dependent Semantic Annotation in Cross-Lingual Biomedical Resources . . . . .	120
<i>Rafael Berlanga, Antonio Jimeno-Yepes, María Pérez-Catalán, and Dietrich Rebholz-Schuhmann</i>	
A Comparative Evaluation of Cross-Lingual Text Annotation Techniques . . . . .	124
<i>Lei Zhang, Achim Rettinger, Michael Färber, and Marko Tadić</i>	

## Applications

Mining Query Logs of USPTO Patent Examiners . . . . .	136
<i>Wolfgang Tannebaum and Andreas Rauber</i>	
Relevant Clouds: Leveraging Relevance Feedback to Build Tag Clouds for Image Search . . . . .	143
<i>Luis A. Leiva, Mauricio Villegas, and Roberto Paredes</i>	
Counting Co-occurrences in Citations to Identify Plagiarised Text Fragments . . . . .	150
<i>Solange de L. Pertile, Paolo Rosso, and Viviane P. Moreira</i>	
The Impact of Belief Values on the Identification of Patient Cohorts . . .	155
<i>Travis Goodwin and Sanda M. Harabagiu</i>	
Semantic Discovery of Resources in Cloud-Based PACS/RIS Systems . . .	167
<i>Rafael Berlanga, María Pérez-Catalán, Lledó Museros, and Rafael Forcada</i>	
Subtopic Mining Based on Head-Modifier Relation and Co-occurrence of Intents Using Web Documents . . . . .	179
<i>Se-Jong Kim and Jong-Hyeok Lee</i>	

## Lab Overviews

Cultural Heritage in CLEF (CHiC) 2013 . . . . .	192
<i>Vivien Petras, Toine Bogers, Elaine Toms, Mark Michael Hall, Jacques Savoy, Piotr Malak, Adam Pawłowski, Nicola Ferro, and Ivano Masiero</i>	
Overview of the ShARe/CLEF eHealth Evaluation Lab 2013 . . . . .	212
<i>Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J.F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeriot, David Martinez, and Guido Zuccon</i>	
Overview of CLEF-IP 2013 Lab: Information Retrieval in the Patent Domain . . . . .	232
<i>Florina Piroi, Mihai Lupu, and Allan Hanbury</i>	
ImageCLEF 2013: The Vision, the Data and the Open Challenges . . . . .	250
<i>Barbara Caputo, Henning Müller, Bart Thomee, Mauricio Villegas, Roberto Paredes, David Zellhöfer, Hervé Goëau, Alexis Joly, Pierre Bonnet, Jesus Martínez-Gómez, Ismael García Varea, and Miguel Cazorla</i>	
Overview of INEX 2013 . . . . .	269
<i>Patrice Bellot, Antoine Doucet, Shlomo Geva, Sairam Gurajada, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Arunav Mishra, Véronique Moriceau, Josiane Mothe, Michael Preminger, Eric SanJuan, Ralf Schenkel, Xavier Tannier, Martin Theobald, Matthew Trappett, and Qiuyue Wang</i>	
Recent Trends in Digital Text Forensics and Its Evaluation: Plagiarism Detection, Author Identification, and Author Profiling . . . . .	282
<i>Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efsthios Stamatatos, and Benno Stein</i>	
QA4MRE 2011-2013: Overview of Question Answering for Machine Reading Evaluation . . . . .	303
<i>Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, and Roser Morante</i>	
Multilingual Question Answering over Linked Data (QALD-3): Lab Overview . . . . .	321
<i>Philipp Cimiano, Vanessa Lopez, Christina Unger, Elena Cabrio, Azel-Cyrille Ngonga Ngomo, and Sebastian Walter</i>	

Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems .....	333
<i>Enrique Amigó, Jorge Carrillo-de-Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, and Damiano Spina</i>	
Entity Recognition in Parallel Multi-lingual Biomedical Corpora: The CLEF-ER Laboratory Overview .....	353
<i>Dietrich Rebholz-Schuhmann, Simon Clemenide, Fabio Rinaldi, Senay Kafkas, Erik M. van Mulligen, Chinh Bui, Johannes Hellrich, Ian Lewin, David Milward, Michael Poprat, Antonio Jimeno-Yepes, Udo Hahn, and Jan A. Kors</i>	
<b>Author Index</b> .....	369



# The Scholarly Impact of CLEF (2000–2009)

Theodora Tsirikla<sup>1</sup>, Birger Larsen<sup>1</sup>, Henning Müller<sup>2</sup>,  
Stefan Endrullis<sup>3</sup>, and Erhard Rahm<sup>3</sup>

<sup>1</sup> Royal School of Library & Information Science, University of Copenhagen,  
Denmark

<sup>2</sup> University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland

<sup>3</sup> University of Leipzig, Germany

**Abstract.** This paper assesses the scholarly impact of the CLEF evaluation campaign by performing a bibliometric analysis of the citations of the CLEF 2000–2009 proceedings publications collected through Scopus and Google Scholar. Our analysis indicates a significant impact of CLEF, particularly for its well-established Adhoc, ImageCLEF, and QA labs, and for the lab/task overview publications that attract considerable interest. Moreover, initial analysis indicates that the scholarly impact of ImageCLEF is comparable to that of TRECvid.

## 1 Introduction

The scholarly impact of *research activities* is commonly measured by their associated publications (i.e., the publications generated as a result of such activities) and the citations they receive. Existing work in bibliometrics and scientometrics has mainly focussed on assessing the scholarly impact of specific publication venues [5] (e.g., journals and conference proceedings) or of the research activities of individual authors [1], institutions, countries, or particular domains [2].

In the field of information retrieval, evaluation campaigns at the international level (e.g., TREC<sup>1</sup>, CLEF<sup>2</sup>, INEX<sup>3</sup>, NTCIR<sup>4</sup>, and FIRE<sup>5</sup>) constitute a research activity that has been widely credited with contributing tremendously to the advancement of the field. Measuring the impact of such benchmarking activities is crucial for assessing which of their aspects have been successful, and thus obtain guidance for the development of improved evaluation methodologies and information retrieval systems. Given that their contribution to the field is mainly indicated by the research that would otherwise not have been possible, it is reasonable to consider that their success can be measured, to some extent, by the scholarly impact of the research they foster. Recent investigations have reported on the scholarly impact of TRECvid<sup>6</sup> [7] and ImageCLEF<sup>7</sup> [8]. Building on this

---

<sup>1</sup> Text REtrieval Conference (<http://trec.nist.gov/>)

<sup>2</sup> Cross-Language Evaluation Forum (<http://www.clef-initiative.eu/>)

<sup>3</sup> INitiative for the Evaluation of XML retrieval (<http://www.inex.otago.ac.nz/>)

<sup>4</sup> NTCIR Evaluation of Information Access Technologies (<http://ntcir.nii.ac.jp/>)

<sup>5</sup> Forum for Information Retrieval Evaluation (<http://www.isical.ac.in/~clia/>)

<sup>6</sup> TREC Video Retrieval Evaluation (<http://trecvid.nist.gov/>)

<sup>7</sup> CLEF Image Retrieval Evaluation (<http://www.imageclef.org/>)

work, this paper presents a preliminary study on assessing the scholarly impact of the first ten years of CLEF activities. To this end, it performs a citation analysis on a dataset of publications obtained from the CLEF proceedings.

## 2 The CLEF Evaluation Campaign

Evaluation campaigns enable the reproducible and comparative evaluation of new approaches, algorithms, theories, and models, through the use of standardised resources and common evaluation methodologies within regular and systematic evaluation cycles. Motivated by the need to support users from a global community accessing the ever growing body of multilingual and multimodal information, the CLEF annual evaluation campaign, launched in 1997 as part of TREC, became an independent event in 2000 with the goal to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information. To this end, it provides an infrastructure for: (i) the comparative evaluation of multilingual and multimodal information access systems, (ii) the creation of reusable resources for such benchmarking purposes, (iii) the exploration of new evaluation methodologies and innovative ways of using experimental data, and (iv) the exchange of ideas.

CLEF is organised as a series of evaluation *labs* (referred to as *tracks* before 2010), each with a focus on a particular research area, ranging from the core cross-lingual adhoc retrieval (*Adhoc*) to multilingual question answering (*QA@CLEF*), cross-language image retrieval (*ImageCLEF*), and interactive retrieval (*iCLEF*). Some labs are in turn structured into tasks, each with even more focussed research objectives. In 2010, CLEF changed its format by accompanying its labs with a peer-reviewed conference. This paper focusses on the first ten years of CLEF and does not consider the changes that took place thereafter.

CLEF's annual evaluation cycle culminates in a workshop where participants of all labs present and discuss their findings with other researchers. This event is accompanied by the **CLEF working notes**, where research groups publish, separately for each lab and task, *participant* notebook papers that describe their techniques and results. In addition, the organisers of each lab (and/or each task) publish *overview* papers that present the evaluation resources used, summarise the approaches employed by the participating groups, and provide an analysis of the main evaluation results. Moreover, *evaluation* papers reflecting on evaluation issues, presenting other evaluation initiatives, or describing and analysing evaluation resources and experimental data may also be included. These (non-refereed) CLEF working notes papers are available online on the CLEF website.

From 2000 to 2009, participants were invited to publish after each workshop more detailed descriptions of their approaches and more in-depth analyses of the results of their participation, together with further experimentation, if possible, to the **CLEF proceedings**. These papers went through a reviewing process and the accepted ones, together with updated versions of the overview papers, were published in a volume of the Springer Lecture Notes in Computer Science series in the year following the workshop and the CLEF evaluation campaign.

Moreover, CLEF participants and organisers may extend their work and publish in journals, conferences, and workshops. The same applies for research groups from academia and industry that, while not official participants of the CLEF activities, may decide at a later stage to use CLEF resources to evaluate their approaches. These **CLEF-derived** publications are a good indication of the impact of CLEF beyond the environment of the evaluation campaign.

### 3 Bibliometric Analysis Method

Bibliometric studies provide a quantitative and qualitative indication of the scholarly impact of research by examining the number of publications derived from it and the number of citations these publications receive. The most comprehensive citation data sources are: (i) *ISI Web of Science*, (ii) *Scopus*, and (iii) *Google Scholar*. ISI and Scopus also provide citation analysis tools to calculate various metrics of scholarly impact, such as the h-index [3]. Google Scholar, on the other hand, does not offer such capabilities for arbitrary publication sets; citation analysis using its data can though be performed by systems such as the *Online Citation Service* (OCS – <http://dbs.uni-leipzig.de/ocs/>) and *Publish or Perish* (PoP – <http://www.harzing.com/pop.htm>).

Each of these sources follows a different data collection policy that affects both the publications covered and the number of citations found. Differences in their coverage can enormously affect the assessment of scholarly impact metrics; the degree to which this happens varies among disciplines [1,2]. For computer science, where publications in peer-reviewed conference proceedings are highly valued and cited in their own right, ISI greatly underestimates the number of citations found [5,1], given that its coverage of conference proceedings is very partial. Scopus offers broader coverage, which may though be hindered by its lack of coverage before 1996; this does not affect this study. Google Scholar offers an even wider coverage and thus further benefits citation analyses performed for the computer science field [5,2]. As a result, this study employs both Scopus and Google Scholar (in particular its OCS and PoP wrappers) for assessing the scholarly impact of CLEF. This allows us to also explore a further goal: to compare and contrast these data sources in the context of such an analysis.

Similarly to [8], the focus is on the CLEF proceedings; analysis of the CLEF working notes and CLEF-derived publications is left as future work. The CLEF 2000–2009 proceedings contain 873 publications. These were obtained through DBLP and were semi-automatically annotated with their *type* (i.e., evaluation, participant, or overview) and the *lab(s)* and/or *tasks(s)* they refer to.

Their citations were obtained as follows in an 24-hour period in April 2013. In Scopus, the query “SRCTITLE(lecture notes in computer science) AND VOLUME(*proceedings\_volume*)” was entered in the Advanced Search separately for each year and the results were manually cross-checked against the publication list. In OCS, the list of publications was directly uploaded to the system which matched them to one or more Google Scholar entries. The result list consisting of tuples of the form  $\langle \textit{input\_publication}, \textit{Google\_Scholar\_match}, \textit{number\_of\_citations} \rangle$  was manually refined so as to remove false positive matches.

Furthermore, the citations (if any) of publications for which OCS did not find a match were manually added to the list. In PoP, the proceedings title was used in the Publication field and the proceedings publication year in the Year field. The results were also manually refined by removing false positive matches, merging entries deemed equivalent, and adding the citations of unmatched publications.

It should be noted that the reliability of Google Scholar as a data source for bibliometric studies is being received with mixed feelings [1], and some outright scepticism [4], due to its widely reported shortcomings [5,4,1]. In particular, Google Scholar frequently has several entries for the same publication, e.g., due to misspellings or incorrectly identified years, and therefore may deflate citation counts [5,4]. OCS rectifies this through multiple matching and PoP through support for manual merging. Inversely, Google Scholar may also inflate citation counts by grouping together citations of different papers, e.g., the journal and conference version of a paper with the same or similar titles [5,4]. Furthermore, Google Scholar is not always able to correctly identify the publication year of an item [4]. These deficiencies have been taken into account and addressed with manual data cleaning when possible, but we should acknowledge that examining the validity of citations in Google Scholar is beyond the scope of this study.

## 4 Results of the Bibliometric Analysis

The results of the bibliometric analysis of the citation data found by the three sources for the 873 CLEF proceedings publications are presented in Table 1. Over the years, there is a steady increase in the number of publications, in line with the continuous increase in the number of offered labs (with the exception of 2007). The coverage of publications varies significantly between Scopus and Google Scholar, with the former indexing a subset that does not include the entire 2000 and 2001 CLEF proceedings and another four individual publications, and thus contains 92% of all publications, while the latter does not index 22 (0.02%) of all publications. Table 2 indicates that Spain is the country that has produced the most CLEF proceedings publications, with five of its institutions and four of its authors being among the top 10 most prolific. Although the statistics in Table 2 are obtained from Scopus, and therefore cover only the years 2002–2009, they can still be considered representative of the whole dataset since they describe over 90% of all publications; OCS and PoP do not readily support such analysis.

The number of citations varies greatly between Scopus and Google Scholar, with the latter finding around ten times more citations than Scopus. Overall, the total number of citations over the 873 CLEF proceedings publications are 9,137 and 8,878 as found by OCS and PoP, respectively, resulting in 10.47 and 10.17 average cites per paper, respectively, while Scopus only finds 905 citations.

The differences between these data sources are investigated further by examining the correlations of the citations they find. Scopus' low coverage does not allow for meaningful comparisons to the other two sources and therefore our investigation focusses on the differences between OCS and PoP. Since both rely on Google Scholar, their differences are not substantial. Figure 1(a) shows a strong

**Table 1.** The citations, average number of citations per publication, and h-index of the CLEF proceedings publications as found by the three sources

	# labs	# publ.	OCS			PoP			Scopus		
			# cit.	avg.	h-index	# cit.	avg.	h-index	# cit.	avg.	h-index
2000	3	27	501	18.56	15	507	18.78	15	-	-	-
2001	2	37	904	24.43	17	901	24.35	17	-	-	-
2002	4	44	636	14.45	14	634	14.41	14	74	1.68	4
2003	6	65	787	12.11	15	776	11.94	15	87	1.34	5
2004	6	81	989	12.21	17	942	11.63	16	137	1.69	5
2005	8	112	1231	10.99	18	1207	10.78	17	133	1.19	5
2006	8	127	1278	10.06	18	1250	9.84	18	133	1.05	5
2007	7	116	1028	8.86	16	902	7.78	15	119	1.03	5
2008	10	131	1002	7.65	16	989	7.55	16	78	0.60	3
2009	10	133	781	5.87	12	770	5.79	12	144	1.08	5
Total	14	873	9,137	10.47	41	8,878	10.17	41	905	1.04	10

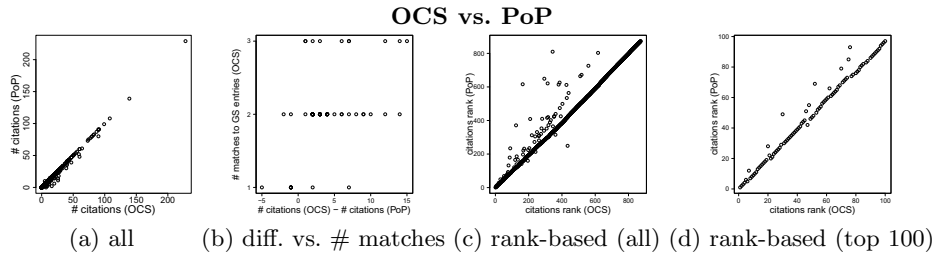
**Table 2.** Top 10 countries, affiliations, and authors of the CLEF 2002–2009 proceedings publications as found by Scopus

Country		Affiliation		Author	
Spain	178	Universidad de Alicante	44	Jones G.J.F.	29
Germany	105	UNED	33	Mandl T.	25
United States	93	Dublin City University	30	Llopis F.	24
France	67	University of Amsterdam	29	de Rijke M.	24
United Kingdom	61	Universidad de Jaen	27	Garcia-Cumbreras M.A.	20
Italy	55	Universität Hildesheim	25	Urena-Lopez L.A.	20
Netherlands	54	Universidad Carlos III de Madrid	24	Clough P.	19
Switzerland	52	UC Berkeley	23	Penas A.	18
Ireland	41	Universidad Politecnica de Madrid	22	Rosso P.	18
Canada	25	University of Sheffield	21	Leveling J.	17

correlation between the number of citations OCS and PoP find for each publication, particularly for publications with high citation counts. This is further confirmed by Figures 1(c)–(d) that show the correlations between the rankings based on the citation counts over all publications and over the 100 most cited publications, respectively. Here, ties in the rankings are resolved using the titles, but similar results are obtained when using the authors' names. The overlap in publications ranked by both in the top  $k = \{100, 200, 300, 400, 500\}$  is over 96%.

Overall, OCS finds 259 (3%) more citations than PoP. The difference for a single publication ranges from 1 to 15 citations, as illustrated in Figure 1(b). Small differences could be attributed to changes in the Google Scholar index that may have taken place during the time period that intervened between obtaining the citation data from each source. Larger differences could be attributed to the different policies adopted by OCS and PoP for matching each input publication to a Google Scholar entry. Figure 1(b) plots the differences in citation counts against the number of Google Scholar matches found by OCS; the higher the difference, the more likely that OCS found more matches. This indicates that OCS achieves a slightly higher recall, and therefore OCS data will be used for the analysis performed in the following sections, unless stated otherwise.

Finally, when examining the distributions over the years, OCS and PoP reach their peak in terms of number of citations and h-index values in 2006. The average number of citations per publication peaks much earlier though, indicating



**Fig. 1.** Correlations between the citations found by the different sources

that the publications of the early CLEF years have on average much more impact than the more recent ones. This could be attributed to the longer time period afforded to these earlier publications for accumulating citations. Given though the current lack of access to the citing papers through the OCS and PoP systems, only a future analysis that will monitor changes in regular intervals (e.g., yearly) could provide further insights (see also Section 4.4).

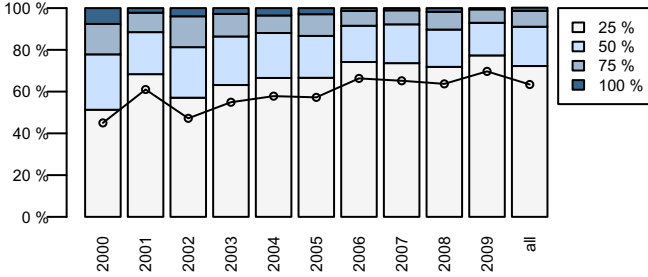
#### 4.1 Citation Distribution

Metrics such as the total number of citations and the average number of citations per publication do not allow us to gauge the impact of individual publications, given that scientific publications are typically cited to a variable extent and citation distributions across such publications are found to be highly skewed [6]. To determine the degree of citation skew and thus gain insights into the variability of the impact of particular publications, the distribution of citations into publication quartiles are examined for each year and overall.

Figure 2 indicates the relative cumulative citation count for each quartile of publications. The 25% of top cited publications account for 50 to 75% of all citations (72% on average), while the bottom 25% of publications merely attract 0.5–7.5% of all citations (1.5% on average). This citation skewness appears to be increasing over the years. For the first three years, the top 25% of publications account for less than 60% of all citations, for the next three years, for around 65% of all citations, while for the last four years, for close to 75% of all citations.

These results are corroborated by also measuring the skewness of the citation distribution using the *Gini coefficient*, a measure of statistical dispersion that reflects the inequality among values of a frequency distribution. The Gini coefficient corresponds to a nonnegative real number, with higher values indicating more diverse distributions; 0 indicates complete equality, and 1 total inequality. Its overall value of 0.63 in CLEF indicates the high degree of variability in the citations of individual publications, and this diversity is continuously increasing as indicated by the values of the Gini coefficient being below 0.5, around 0.55, and over 0.65 for the first three, next three, and final four years, respectively.

The exception to the above observations is the year 2001, which is more skewed compared to the other early CLEF years; its Gini coefficient is 0.61, while its



**Fig. 2.** The distributions of citations found by OCS (split by quarters) over the years and overall, and the Gini coefficient of these distributions plotted as a line

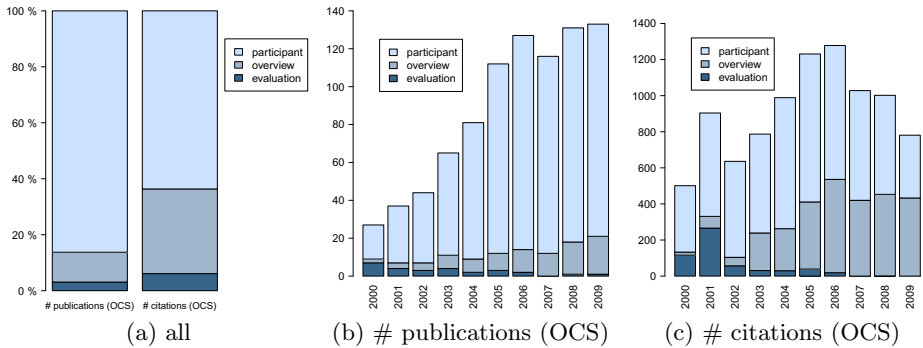
**Table 3.** Top 10 cited publications as found by OCS: their rank and number of citations by the three sources, and their author(s), title, year, and type (E = *evaluation*, O = *overview*, P = *participant*). Terms in *italics* denote abbreviations of original title terms.

OCS / PoP / Scopus rank	# citations	Author(s)	Title	Year	Type				
1	1	-	228 229	-	-				
2	2	139	139	17	Voorhees	The Philosophy of Information Retrieval Evaluation.	2001	E	
3	3	5	108	108	12	Müller et al.	Overview of the ImageCLEFmed 2006 Medical Retrieval [...]	2006	O
4	4	1	99	99	17	Clough et al.	The CLEF 2005 Cross-Language Image Retrieval Track.	2005	O
5	6	290	91	91	4	Clough et al.	The CLEF 2004 Cross-Language Image Retrieval Track.	2004	O
6	5	6	90	91	11	Vallin et al.	Overview of the CLEF 2005 Multilingual QA Track.	2005	O
7	12	29	90	80	5	Chen	Cross-Language Retrieval Experiments at CLEF 2002.	2002	P
8	7	-	90	90	-	Grubinger et al.	Overview of the ImageCLEFphoto 2007 [...] Task.	2007	O
9	8	14	87	87	7	Monz & de Rijke	Shallow Morphological Analysis in Monolingual IR [...]	2001	P
10	9	4	83	83	13	Müller et al.	Overview of the CLEF 2009 Medical Image Retrieval Track.	2009	O
						Magnini et al.	Overview of the CLEF 2004 Multilingual QA Track.	2004	O

top 25% publications account for almost 70% of all citations. This high degree of variability is due to the inclusion of two of the top 10 cited publications over all years, listed in Table 3, and in particular due to the domination of the most cited publication, a paper by Ellen Voorhees [9], which achieves around 65% more citations than the second most cited publication. The remaining top cited publications in Table 3 are more or less evenly spread across the years.

## 4.2 Citation Analysis of CLEF Publications Types

Figure 3(a) compares the relative number of publications of the three types (*evaluation*, *overview*, and *participant*) with their relative citation frequency. As also listed in the last column of Table 4, the participants' publications account for a substantial share of all publications, namely 86%, but only receive 64% of all citations. On the other hand, overview and evaluation publications receive three times or twice the percentage of citations compared to their publications' percentage. This indicates the significant impact of these two types; the significant impact of overview publications is further illustrated in Table 3 where 7 out of the 10 most cited publications are overviews, while the impact of evaluation publications can be attributed to a single publication, the Voorhees paper [9].



**Fig. 3.** Relative impact of different types of CLEF proceedings publications

**Table 4.** Relative percentages of different types of CLEF proceedings publications and their citations over the years

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2000–2009
	<b>% publications</b>										
evaluation	25.93	10.81	6.82	6.15	2.47	2.68	1.57	0.00	0.76	0.75	0.03
overview	7.41	8.11	9.09	10.77	8.64	8.04	9.45	10.34	12.98	15.04	0.11
participant	66.67	81.08	84.09	83.08	88.89	89.29	88.98	89.66	86.26	84.21	0.86
	<b>% citations</b>										
evaluation	23.15	29.42	8.96	3.94	3.03	3.17	1.49	0.00	0.10	0.00	0.06
overview	3.39	7.19	7.39	26.43	23.56	30.22	40.45	40.86	45.11	55.44	0.30
participant	73.45	63.38	83.65	69.63	73.41	66.61	58.06	59.14	54.79	44.56	0.64

Figures 3(b)–(c) and Table 4 drill down from the summary data into the time dimension. During the early years, CLEF proceedings included several evaluation publications, many of them invited, which attracted a considerable number of citations, with the Voorhees [9] paper in 2001 being the most prominent example. More recently, such publications and consequently their citations have all but disappeared. The number of participants’ publications has mostly followed a steady increase both in absolute and in relative terms, reaching almost 90% of all publications for some years. However, such publications manage to attract only between 44% and 74% of all citations, with the exception of 2002, where participants’ publications received almost 84% of all citations. This is mostly due to a single participant’s publication included among the 10 most cited publications (see Table 3). Finally, the impact of overview publications has significantly increased during the more recent years, where overviews constitute only 10 to 15% of all publications, but account for 40 to 55% of all citations.

### 4.3 Citation Analysis of CLEF Labs and Tasks

Table 5 presents the results of the citation analysis for the publications of the 14 labs and their tasks organised by CLEF during its first 10 years. Two more “pseudo-labs”, *CLEF* and *Other* are also listed; these are used for classifying the



**Table 5.** CLEF labs and tasks in alphabetical order, the number of years they have run, their publications, citations, average number of citations per publication, and the type of the most cited publication (E = *evaluation*, O = *overview*, P = *participant*). The number of publications and citations over all tasks for a lab may not sum up to the total listed for *all tasks* for that lab, since a publication may refer to more than one task. Similarly for the number of publications and citations over all labs.

Lab	Task	#years	# publications	# citations	average	most cited	
Adhoc	( <i>all tasks</i> )	10	237	2540	10.72	P	
	Cross/Mono-lingual	8	188	2285	12.15	P	
	Persian	2	11	97	8.82	O	
	Robust	4	30	192	6.40	O	
	TEL	2	19	150	7.89	O	
CL-SR		6	29	208	7.17	O	
CLEF		10	23	203	8.83	E	
CLEF-IP		1	15	85	5.67	O	
Domain-Specific		9	47	555	11.81	P	
GeoCLEF		4	58	561	9.67	O	
GRID@CLEF		1	3	8	2.67	O	
iCLEF		9	41	378	9.22	O	
ImageCLEF	( <i>all tasks</i> )	7	179	2018	11.27	O	
	Interactive	1	2	4	2.00	P	
	Medical Annotation	5	37	586	15.84	O	
	Medical Retrieval	6	62	1002	16.16	O	
	Photo Annotation	4	21	245	11.67	O	
	Photo Retrieval	7	86	1002	11.65	O	
	Robot Vision	1	6	23	3.83	O	
	Wikipedia Retrieval	2	11	74	6.73	O	
	INFILE		2	8	5	0.62	O
	LogCLEF		1	6	25	4.17	O
MorphoChallenge		3	20	247	12.35	P	
Other		5	8	277	34.62	E	
QA@CLEF	( <i>all tasks</i> )	7	173	2023	11.69	O	
	AVE	3	25	274	10.96	O	
	GikiCLEF	1	7	32	4.57	O	
	QA	6	114	1489	13.06	O	
	QAST	3	11	89	8.09	O	
	ResPubliQA	1	10	95	9.50	O	
	WiQA	1	7	52	7.43	O	
VideoCLEF		2	14	79	5.64	O	
WebCLEF		4	28	180	6.43	P	
All		10	873	9,137	10.47	E	

evaluation type publications not assigned to specific labs, but rather pertaining to evaluation issues related to CLEF or other evaluation campaigns, respectively.

Three labs, *Adhoc*, *ImageCLEF*, and *QA@CLEF*, clearly dominate in terms of publication and citation numbers; they account for 67% of all publications and for 72% of all citations. They also account for 9 of the 10 most cited publications in Table 3. The highest number of citations per publication is observed for the Other evaluation publications, which are highly skewed due to the presence of the Voorhees [9] paper. Excluding these from further consideration, the aforementioned three labs are among the top ranked ones, together with the *Domain-Specific* and *MorphoChallenge*. Overall, the *Medical Retrieval* and *Medical Annotation* ImageCLEF tasks have had the greatest impact among all labs and tasks, closely followed by the main *QA* task and the main *Cross/Mono-lingual* Adhoc task. This also indicates a bias towards older, most established labs and tasks. Finally, the most cited publication in each lab or task is in most cases its overview, further indicating the high impact of such publications.

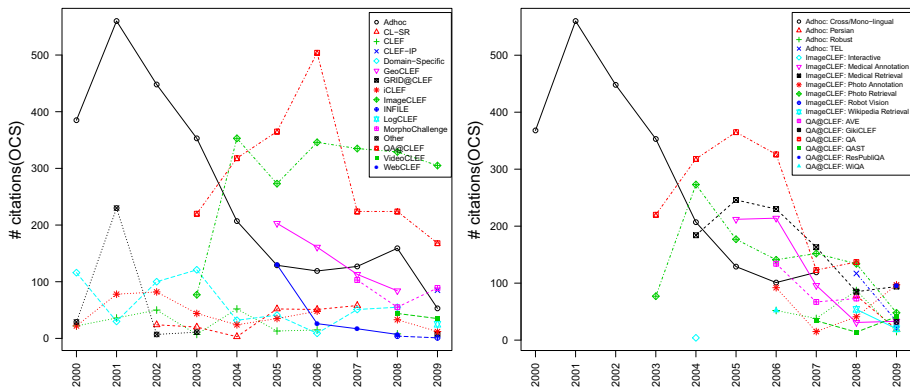


Fig. 4. The impact of CLEF labs (left) and tasks (right) over the years

Figure 4 depicts the number of citations for the CLEF labs and tasks over the years. Although it is difficult to identify trends over all labs and tasks, in many cases there appears to be a peak in their second or third year of operation, followed by a decline. Exceptions include the *Photo Annotation* ImageCLEF task, which attracted significant interest in its fourth year when it employed a new collection and adopted new evaluation methodologies, and also the *Cross-Language Speech Retrieval* (CL-SR) lab that increased its impact in 2005 following a move from broadcast news to conversational speech. Such novel aspects result in renewed interest in labs and tasks, and also appear to strengthen their impact.

#### 4.4 Assessing the Impact of ImageCLEF in 2011 and in 2013

A previous study [8] assessed the scholarly impact of ImageCLEF by performing a bibliometric analysis of citation data collected in April 2011 through Scopus and PoP. Table 6 compares and contrasts the results of this earlier study with the results of this work using the same data sources two years later. The earlier study also took into account iCLEF publications that relied on ImageCLEF datasets or were otherwise closely related to ImageCLEF. However, the impact of these additional publications is negligible, since their citations account for less than 0.04% of all citations; these two results sets can be viewed as being comparable.

There is a considerable increase in the number of citations over these two years: 364 (+23%) more citations are found by PoP and 91 (+50%) by Scopus. For PoP, most citations are added to the 2004 and 2006 publications, while for Scopus to the 2007–2009 ones. Overall, the impact of ImageCLEF tasks appears to increase several years after they took place, however further analysis is needed to determine whether these citations originate from papers published over these two years, or from papers simply added to the sources' indexes during this time.

**Table 6.** Bibliometric analyses of the ImageCLEF publications published in the CLEF 2003–2009 proceedings performed in 2011 and in 2013 using Scopus and PoP

		#publications		# citations		average		h-index	
		2011	2013	2011	2013	2011	2013	2011	2013
Scopus	2003	5	5	13	14	2.60	2.80	2	3
	2004	20	20	50	64	2.50	3.20	4	5
	2005	25	22	24	30	0.96	1.36	3	3
	2006	27	23	25	38	0.93	1.65	2	3
	2007	29	29	18	34	0.62	1.17	3	3
	2008	45	40	14	34	0.31	0.85	2	3
	2009	44	40	38	59	0.86	1.48	4	5
	<b>Total</b>	195	179	182	273	0.93	1.53	6	7
PoP	2003	5	5	65	74	13.00	14.80	3	4
	2004	20	20	210	340	10.50	17.00	8	10
	2005	25	22	247	265	9.88	12.05	7	8
	2006	27	23	259	344	9.59	14.96	7	8
	2007	29	29	249	291	8.59	10.03	7	9
	2008	45	40	284	318	6.31	7.95	7	8
	2009	44	40	259	305	5.89	7.63	7	7
	<b>Total</b>	195	179	1,573	1,937	8.06	10.82	18	22

**Table 7.** Bibliometric analyses of all TRECVID (*TVa*) [7], TRECVID working notes (*TV*), CLEF proceedings (*C*), and ImageCLEF (*I*) publications using PoP

	#publications				# citations				average				h-index			
	TVa	TV	C	I	TVa	TV	C	I	TVa	TV	C	I	TVa	TV	C	I
2003	64	27	65	5	1,066	561	787	74	16.66	20.78	12.11	14.80	18	10	15	4
2004	158	29	81	20	2,124	423	989	340	13.44	14.59	12.21	17.00	24	11	17	10
2005	225	26	112	22	2,537	433	1231	265	11.28	16.65	10.99	12.05	28	8	18	8
2006	361	35	127	23	4,068	437	1278	344	11.27	12.49	10.06	14.96	30	11	18	8
2007	382	34	116	29	3,562	244	1028	291	8.97	7.18	8.86	10.03	28	6	16	9
2008	509	40	131	40	1,691	175	1002	318	3.32	4.37	7.65	7.95	16	10	16	8
2009	374	13	133	40	780	12	781	305	2.09	0.92	5.87	7.63	12	2	12	7
<b>Total</b>	2,073	205	765	179	15,828	2,285	7,096	1,937	7.63	11.21	9.28	10.82	52	25	38	22

#### 4.5 Comparing to the Impact of other Evaluation Campaigns

Assessments of the scholarly impact of other evaluation campaigns have only been performed for TRECVID (2003–2009) [7], where a list containing both the *TRECVID working notes* and the *TRECVID-derived* publications was analysed. For comparability to the CLEF proceedings, we obtained the data used in [7] (<http://www.cdvp.dcu.ie/scholarly-impact/>) and manually identified the subset of the TRECVID working notes publications. Table 7 analyses these three sets (all TRECVID, TRECVID working notes, CLEF publications), and also the ImageCLEF publications, since this lab and TRECVID focus on similar domains.

Overall, there are about three times more TRECVID publications than CLEF proceedings ones, but receive on average less citations. It is difficult though to draw conclusions given the multidisciplinary nature of CLEF coupled with the different citation practices in different domains. The number of TRECVID working notes publications is close to that of ImageCLEF, with the former attracting a slightly higher number of citations, but not significantly so; both perform better than the larger sets. It appears that ImageCLEF is on par with TRECVID, taking also into account the fact that ImageCLEF was first established in 2003,

while TRECVID was part of TREC already from 2001 and became an independent event in 2003. On the other hand, the TRECVID working notes publications list is rather incomplete (cf. [7]). Also, the data in [7] were collected earlier and thus it is likely that the TRECVID publications have attracted more citations over time. Further investigation is needed for reaching more reliable conclusions.

## 5 Conclusions

Measuring the impact of evaluation campaigns may prove useful for supporting research policy decisions by determining which aspects have been successful, and thus obtaining guidance for the development of improved evaluation methodologies and systems. This bibliometric analysis of the CLEF 2000–2009 proceedings has shown the considerable impact of CLEF during its first ten years in several diverse multi-disciplinary research fields. The high impact of the overview publications further indicates the significant interest in the created resources and the developed evaluation methodologies, typically described in such papers. It is necessary though to extend this analysis and include the working notes and all derived work. Finally, our analysis has highlighted the differences between the available citation analysis tools: Google Scholar provides a much wider coverage than Scopus, while OCS and PoP are in essence comparable, each with different querying facilities that might prove advantageous in different situations.

**Acknowledgements.** This work was partially supported by the ELIAS ESF Research Networking Programme, and by the Promise (258191) and Khresmoi (257528) FP7 projects.

## References

1. Bar-Ilan, J.: Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics* 74(2), 257–271 (2008)
2. Harzing, A.-W.: Citation analysis across disciplines: The impact of different data sources and citation metrics (2010), [http://www.harzing.com/data\\_metrics\\_comparison.htm](http://www.harzing.com/data_metrics_comparison.htm) (retrieved)
3. Hirsch, J.E.: An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences (PNAS)* 102(46), 16569–16572 (2005)
4. Jacsó, P.: Deflated, inflated and phantom citation counts. *Online Information Review* 30(3), 297–309 (2006)
5. Rahm, E., Thor, A.: Citation analysis of database publications. *SIGMOD Record* 34, 48–53 (2005)
6. Seglen, P.O.: The skewness of science. *JASIS* 43(9), 628–638 (1992)
7. Thornley, C.V., Johnson, A.C., Smeaton, A.F., Lee, H.: The scholarly impact of TRECVID (2003–2009). *JASIST* 62(4), 613–627 (2011)
8. Tsikrika, T., Seco de Herrera, A.G., Müller, H.: Assessing the scholarly impact of ImageCLEF. In: Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., de Rijke, M. (eds.) CLEF 2011. LNCS, vol. 6941, pp. 95–106. Springer, Heidelberg (2011)
9. Voorhees, E.M.: The philosophy of information retrieval evaluation. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 355–370. Springer, Heidelberg (2002)

# A Quantitative Look at the CLEF Working Notes

Thomas Wilhelm-Stein and Maximilian Eibl

Technische Universität Chemnitz, 09107 Chemnitz, Germany  
{wilt,eibl}@hrz.tu-chemnitz.de

**Abstract.** After seven years of participation in CLEF we take a look back at the developments and trends in different domains like evaluation measures and retrieval models. For that purpose a new collection containing all CLEF working notes including their metadata was created and analysed.

**Keywords:** data mining, evaluation, retrospection, retrieval models, evaluation measures.

## 1 Motivation

We have participated in CLEF since 2006 [1]. Sometimes we involved students in our experiments and evaluations. Some had never worked on an evaluation of an information retrieval system before and struggled with the various decisions during the setup of a retrieval experiment and with the interpretation of the evaluation measures. In contrast, long time participants are able to make these decisions based on their experience. To support our students making their decision and to provide them with an overview of the CLEF labs, we created a new collection of the CLEF working notes and a method to analyse it.

## 2 Experiment Setup

The collection was built using all available working notes from the CLEF web site<sup>1</sup>. There were a total of 1413 referenced working notes. Three documents could not be retrieved hence only their metadata is in the collection.

The created index contained the following fields: title, authors, publishing year, track, task, type, and the content of the document, which was extracted from the PDF file. The field type denotes if the document is a working note or an overview paper.

For each concept a query was created manually and tested against the collection by examination of samples from the results. The queries consisted of synonyms and variations of the wording of the concept. A total of 24 queries

---

<sup>1</sup> Retrieved April 19, 2013, from

<http://www.clef-initiative.eu/publication/working-notes>

were created and divided into four groups. In order to analyse the usage of the concepts over the course of the years, the queries were expanded to include a filter for the publishing year. The result was a table, which indicates for each concept and each year in how many working notes a concept was used.

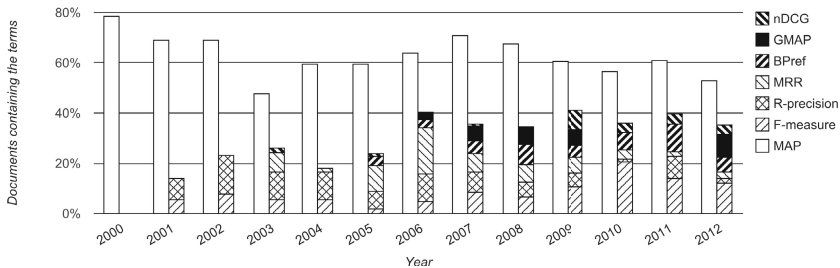
The following concepts were compiled:

- Evaluation measures (see [2], chapter 7): Mean Average Precision (MAP), F-measure, R-precision, Mean Reciprocal Rank (MRR), Binary Preference (BPref), Geometric Mean Average Precision (GMAP), and Normalized Discounted Cumulative Gain (nDCG)
- Retrieval engines: Apache Lucene <sup>2</sup>, SMART (System for the Mechanical Analysis and Retrieval of Text, see [3]), Lemur and Indri <sup>3</sup>, Terrier <sup>4</sup>, and Xapian<sup>5</sup>
- Retrieval models: Okapi/BM25, TF-IDF, Vector Space Model, Language Model, Latent Semantic Indexing, Divergence from Randomness, and Latent Dirichlet Allocation.
- General techniques: Stemming, Stop Words, Query Expansion, Clustering, and Bag of Words

### 3 Results

In this section the results are discussed using charts. For each year the relative number of documents containing the terms are shown. One hundred percent corresponds to all documents of the respective year.

Figure 1 shows the results for the retrieval measures. In the first year Mean Average Precision (MAP) was found in 78 percent of the documents. In the following years this proportion decreased in favor of other measures. But it remained the most widely used measure. The other measures vary considerably in their use.



**Fig. 1.** Comparison of the usage of the concepts for retrieval measures

<sup>2</sup> see <http://lucene.apache.org/>

<sup>3</sup> see <http://www.lemurproject.org/>

<sup>4</sup> see <http://terrier.org/>

<sup>5</sup> see <http://xapian.org/>

As shown in figure 2 Apache Lucene has gained in its importance since its broad public availability in 2005. In 2006 there were 36 working notes mentioning Lucene, which indicates that it was used in nearly 3 out of 10 experiments.

Even if the SMART retrieval system is apparently still used today, most hits are attributable to references on papers. Some hits are english words like, for example, smart phones[4], some relate to components of SMART like stop word lists[5] and some actually used the retrieval system[6].

Though Xapian is still in active development, it was only used for a small number of experiments in the years 2003 to 2007.

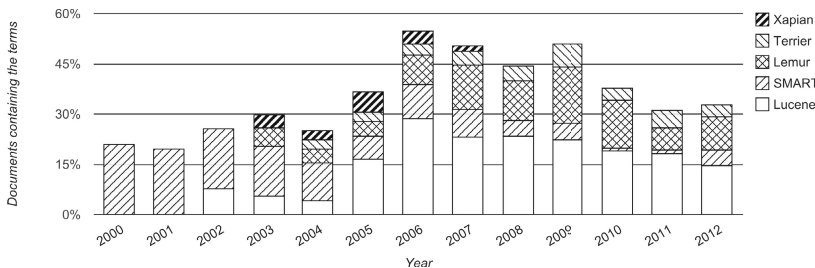


Fig. 2. Comparison of the usage of the concepts for common retrieval engines

Figure 3 shows the use of different retrieval model terms. TF-IDF [7] and the Vector Space Model [8] are used almost constantly over the examined time period. Other models like Okapi/BM25 and Divergence from Randomness had periods where they were used more often.

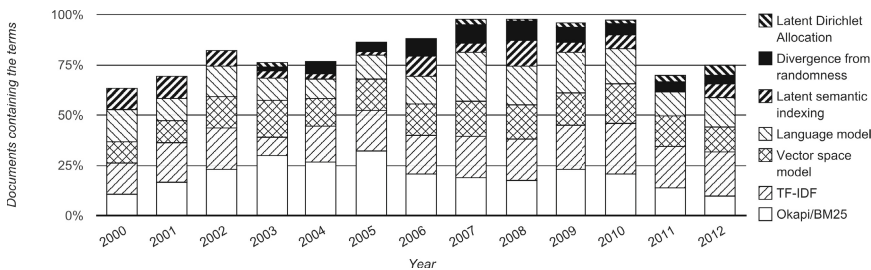


Fig. 3. Comparison of the usage of the concepts for retrieval models

All the general techniques that can be seen in figure 4 were present since the first years of CLEF. Bag of Words and Clustering are shown to have increased in their use, while Query Expansion, Stop Words, and Stemming decreased slightly. Even if this can not be linked directly to their performance, one can see that all of these techniques have held up over the years as an integral component of the retrieval experiments.

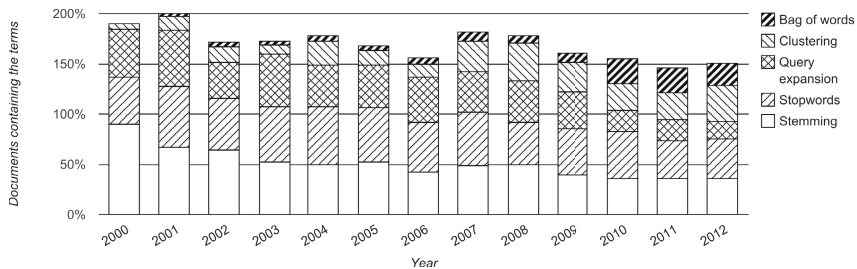


Fig. 4. Comparison of the usage of the concepts for general techniques

## 4 Conclusions and Future Work

These results show only a small part of the information contained in the newly created collection. Despite the small size of the new collection, we provided first insight into the last 13 years of CLEF. Further analyses can improve this statistical overview on the relations between evaluation tasks and the methodology in corresponding experiments. More research could be done in evaluating the measures, depending on the track of the working notes.

Other visualizations could further improve the evaluation. We also want to visualise the results with our *Compeval* tool, which we already used in [9].

## References

- [1] Eibl, M., Kürsten, J.: The importance of being grid: Chemnitz university of technology at grid@clef. Working Notes for the CLEF (2009)
- [2] van Rijsbergen, C.J.: Information Retrieval. Butterworth (1979)
- [3] Salton, G., Lesk, M.E.: The smart automatic document retrieval systems - an illustration. *Commun. ACM* 8(6), 391–398 (1965)
- [4] Suominen, H., Basilakis, J., Johnson, M., Dawson, L., Hanlen, L., Kelly, B., Yeo, A., Sanchez, P.: Clinical speech to text evaluation setting. In: [10]
- [5] Bogers, T., Larsen, B.: Rslis at inex 2012: Social book search track. In: [10]
- [6] Crouch, C.J., Crouch, D.B., Chittilla, S., Nagalla, S., Kulkarni, S., Nawale, S.: The 2012 inex snippet and tweet contextualization tasks. In: [10]
- [7] Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1), 11–21 (1972)
- [8] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. *Communications of the ACM* 18(11), 613–620 (1975)
- [9] Wilhelm, T., Kürsten, J., Eibl, M.: A tool for comparative ir evaluation on component level. In: Ma, W.Y., Nie, J.Y., Baeza-Yates, R.A., Chua, T.S., Croft, W.B. (eds.) *SIGIR*, pp. 1291–1292. ACM (2011)
- [10] Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20 (2012)*



# Building a Common Framework for IIR Evaluation

Mark Michael Hall and Elaine Toms

Information School  
University of Sheffield  
Sheffield S1 4DP, UK  
{m.mhall,e.toms}@sheffield.ac.uk

**Abstract.** Cranfield-style evaluations standardised Information Retrieval (IR) evaluation practices, enabling the creation of programmes such as TREC, CLEF, and INEX, and long-term comparability of IR systems. However, the methodology does not translate well into the Interactive IR (IIR) domain, where the inclusion of the user into the search process and the repeated interaction between user and system creates more variability than the Cranfield-style evaluations can support. As a result, IIR evaluations of various systems have tended to be non-comparable, not because the systems vary, but because the methodologies used are non-comparable. In this paper we describe a standardised IIR evaluation framework, that ensures that IIR evaluations can share a standardised baseline methodology in much the same way that TREC, CLEF, and INEX imposed a process on IR evaluation. The framework provides a common baseline, derived by integrating existing, validated evaluation measures, that enables inter-study comparison, but is also flexible enough to support most kinds of IIR studies. This is achieved through the use of a “pluggable” system, into which any web-based IIR interface can be embedded. The framework has been implemented and the software will be made available to reduce the resource commitment required for IIR studies.

**Keywords:** evaluation, methodology, interactive information retrieval.

## 1 Introduction

Cranfield-style evaluations standardised Information Retrieval (IR) evaluation practices, and served as the foundation for a host of evaluation programmes including TREC, CLEF, and INEX. These set the pace for evaluating the output from information retrieval systems with a view to improving system performance. Many accomplishments over the past three decades in search systems effectiveness can be linked to these programmes. In parallel, the interactive IR (IIR) research community focused somewhat similar research on the user as a core ingredient in the research. While there is overlap, IIR has additional goals: a) assess search systems and components of search systems using user-centred evaluation methods typically found in human experimentation and human computer

interaction (e.g., [12]), and b) examine user actions and activities – both cognitive and behavioural – to understand how people search for information and which aspect of context (e.g., characteristics of the user, the work environment, situation, etc.) influences the process (e.g. [4,10]).

While the TREC and CLEF programmes have enjoyed standardised protocols and measures to assess performance and output, and to experimentally compare among systems, the IIR evaluation field has not had that advantage. The TREC and CLEF evaluation programmes specified standard test collections, test topics and sets of expert-assessed relevant items (including training sets) as the minimum ingredients, and a standard way of presenting and comparing the results – the ubiquitous reverse-ranked list of relevant items per topic and additionally aggregated by system and collection. On the other hand, IIR research was and still is researcher driven with non-standard “collections”, user-imposed search tasks, and diverse sets of measures to support multiple research objectives. In the midst of all of this is usually a set of participants, a sample of convenience. Thus, it is difficult to compare across studies.

The challenge is two-fold: developing a standard methodological protocol that may service multiple types of IIR evaluations and research, and developing a standard set of meaningful measures that are more than descriptive of the process. In this work, we address the first: we designed, developed, implemented and tested a common research infrastructure and protocol that can be used by the IIR research community to systematically conduct IIR studies. Over time, the accumulated studies will also provide a comprehensive data set that includes both context and process data that may be used by the IR community to test and develop algorithms seated in human cognition and behaviour, and additionally to provide a sufficiently robust, detailed, reliable data set that may be used to test existing measures and develop new ones. This paper describes the rationale and the design of the infrastructure, and its subsequent implementation.

## 2 Interactive IR Research – Past and Present

Typically IIR research was conducted using a single system in a laboratory setting in which a researcher observed and interacted with a participant [21]. This was a time-consuming, resource exhaustive and labour intensive process [23,26]. As a result, IIR research used a small number of participants doing a few tasks, which challenged the validity and reliability of the research [11]. In their recent systematic review of 127 IIR studies, Kelly and Sugimoto [13], found extreme variability in IIR studies: from 4 to 283 participants with a mean of 37, and between six and ten task instantiations was typical, although the maximum observed was 56 in a single study.

Similarly what was measured varied significantly; 1533 measures were identified [13]. Clearly the situation has not changed since Yuan and Meadow examined the measures used in 1999 [27], and Tague-Sutcliff in 1992 [21]. The challenge has been that the same: concepts are not always measured using the same “yardstick” and there is no standard set. For example, in the outcome from

the TREC Interactive Track, lab participants used a similar protocol, but the variables tested differed and measurement was not consistent [6]. All of this variability in IIR studies has not allowed for comparison across a series of studies, or the aggregation of data from multiple studies to test hypotheses in large data sets.

The main challenge lies in creating a framework that is sufficiently standardised to enable comparability of evaluation results, while at the same time being flexible enough to be applied to a wide range of experiments and variables in order to ensure its uptake. The matter has been richly discussed by Tague-Sutcliffe [21] who outlined ten key decisions in the research design. Later, first Ingwersen and Jarvelin 2005 [7] and later Kelly's synthesis of IIR [11], synthesized and elaborated on this process. However, the closest we have come to a standard protocol is the set of instruments used by TREC Interactive Track, and a practice of pre- and post-task data capture that has been used more or less consistently.

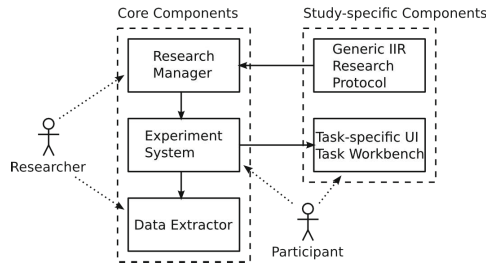
While the traditional method for IIR experiments has been in-the-lab studies, the web introduced alternatives that reduced cost, enabled 24-7 experimentation, provided for a high degree of external validity, and to an extent automated parts of the experimental setup [17,18]. One of the first disciplines to adapt research to the Web was psychology. Its Psychological Research on the Web (<http://psych.hanover.edu/Research/exponnet.html>) continues to provide links to hundreds of web-based surveys and experiments, but this remains simply a list of links. The Web Experiment List (<http://www.wexlist.net>) is a similar but parallel service that provides links to and descriptions of current and past web experiments.

In 2004, Toms, Freund and Li designed and implemented the WiIRE (Web-based Interactive Information Retrieval) system [24], which devised an experimental workflow process that took the participant from information page through a variety of questionnaires and the search interface. Used in TREC 11 Interactive Track, it was built using typical Microsoft Office desktop technologies, which severely limited its capabilities. Data collection relied on server logs limiting the amount of client-side data that could be collected. The concept was later implanted in a new version using PHP, JavaScript, and MySQL used in INEX2007 [25]. This version still provided the basics in implementation of a web-based experiment, but lacked flexibility in setup and data extraction. More recently, SCAMP (Search ConfigurAtor for experiMenting with PuppyIR) was developed by Renaud and Azzopardi [19] which is used to assess IR systems, but does not include the range of IIR research designs that are typically done. Another development is the experiment system described in [2], but to our knowledge it is not publicly released. Thus in IIR, there is a significant amount of interest and need to develop standard protocols and systematic approaches to data collection. Given the diversity in past studies and inconsistencies in what is collected and how much, there is a significant need to develop an approach.

### 3 IIR Evaluation Framework

To overcome these limitations the proposed evaluation framework was designed around five core objectives:

1. Provide a systematic way of setting up an experiment or user study that may be intuitively used by students and researchers;
2. Provide a standard set of evaluation measures to improve comparability;
3. Ensure that standard and consistent data formats are used to simplify the comparison and aggregation of studies;
4. Extract a standard procedure for the conduct of IIR studies from past research, so that studies can share a common protocol even if the system, the tasks, and the participant samples are different;
5. Reduce resource commitment in the conduct of such studies.



**Fig. 1.** Design of the proposed evaluation framework, with the three core and the two study-specific components. In a non-IIR study different study-specific components would be used. In the framework, the researcher interacts only with the *Research Manager* and *Data Extractor*, while the participant only ever sees the *Experiment System* and *Task-specific UI*.

The difficulty in designing a framework that implements these objectives is balancing the standardisation and simplification efforts with the ability to support the wide range of evaluation experiments conducted within IIR. To achieve this we have developed a flexible framework, inspired by the WiIRE system [24,25] and work in the POODLE project [2], that provides the core functionality required by all experiments and into which the experiment-specific functionality can easily be plugged-in (fig. 1). The three core components of the framework are:

- The **Research Manager** is the primary point of interaction for the researcher setting up an experiment. It is used to specify the workflow of the experiment, the tasks and interfaces to use, and all other measures to acquire. To simplify and standardise both the experiment process and results, the **Research Manager** is primed with a *generic research protocol*, such as

the *Generic IIR Research Protocol* provided in this paper, that specifies the basic experiment workflow and into which the researcher only has to add the experiment-specific aspects;

- the **Experiment System** takes the experiment defined by the *Research Manager* and generates the UI screens that the participants interact with. It also ensures that the tasks and interfaces are correctly distributed and rotated between the participants, in accordance with the settings specified in the **Research Manager**. Finally it loads the **Task-specific UI** and records the participants' responses and ensures that they conform to the requirements specified by the researcher. To ensure the flexibility of the system, any web-based system can be used as the **Task-specific UI**;
- the **Results Extractor** takes the participant data gathered by the **Experiment system** and provides them in a format that can be used by analysis packages such as SPSS or R. The data includes not only the participants' responses, but also data on tasks / interfaces used by the participants used and the order in which they appeared.

To simplify the setup and further standardise IIR studies, the following two IIR-specific components have been developed. In a non-IIR context, these would be replaced with components developed for that context.

- the **Generic IIR Research Protocol** aims to define a standardised and re-usable workflow and set of evaluation measures for IIR evaluation studies;
- the **Task Workbench** provides an extensible and pluggable set of UI components for IIR interfaces, with the aim of simplifying the set-up of IIR evaluation experiments.

### 3.1 Research Manager

The *Research Manager* addresses requirements #1 and #5, in that it provides a structured process for setting up experiments and through this reduces the resource commitment required. The *Research Manager* achieves this through the use of *generic research protocols* that specify a structure for the type of experiment the researcher wishes to conduct. The researcher then adapts this *generic research protocol* to their specific requirements. This provides the desired level of standardisation, while at the same time being flexible enough to support a wide range of experiments. The details will be discussed in the context of IIR evaluation, using the *Generic IIR Research Protocol* in section 3.4, but are equally applicable to any other study that can be conducted via the web.

When setting up an experiment, the researcher first selects the *generic research protocol* that they wish to use, although if there is no applicable *generic research protocol*, then the experiment can also be built from scratch. Assuming the researcher selects the *Generic IIR Research Protocol* to setup an IIR study, they are first asked to provide basic information including title, purpose, key researcher names, and contact information, which are used to generate the initial and final information pages. Next, the researcher selects which of the optional

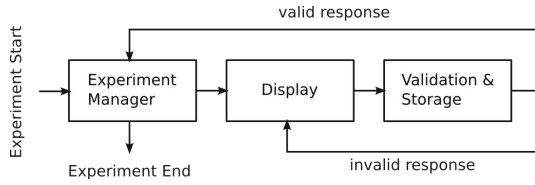
steps in the *Generic IIR Research Protocol* to include in their study. Naturally this choice can be changed at any time, if testing reveals that optional steps are superfluous or should be included. This specifies the basic structure of the experiment and the next step is to define the core tasks to test or control for, in IIR generally including:

- **Task Type:** categorisation of task based on attributes of a task which may be Fact-finding, Know-item, Topical, Transactional and so on. Unfortunately there is no well-defined taxonomy of task type [22], although multiple types have been created. In this case, Task Type will be defined by the participant, although we hope that current research may provide some parameters around these for greater consistency. Each Task Type, e.g., Topical, is represented by multiple instantiations of that type that specify the exact task that a participant will do using the particular interface and collection. For example, find out who should not get a flu shot. The actual number of task instantiations will vary with the amount of effort that is required of the participant, and this is a decision of the researcher.
- **System:** this may be different IR systems, different interfaces to the same IR system; or a single UI with interface objects.
- **Participant Group:** different groups of participants may be recruited based on selected characteristics. For example, novices may be compared to experts, or youth to seniors, or sometimes by scores on a particular human characteristics such as scores on a cognitive style test.

The researcher first identifies which of these elements will be tested, and whether the design will be between- or within-subjects for *Task Type* and/or *System*, and between-subjects for *Participant Group*. Mixed approaches are also possible to handle scenarios where a pure between- or within-subjects approach is not desired or not feasible. Based on these settings the *Research Manager* creates the final experiment that is then passed onto the *Experiment System*, which then uses the settings to ensure that participants are assigned to *Task type / System / Participant Group* combinations and that participants are evenly distributed between the combinations.

### 3.2 Experiment System

The *Experiment System* addresses requirements #3 and #5 by providing a full integrated system that handles the whole workflow of the experiment as it is used by the participants. It takes the experiment designed using the *Research Manager* and guides the participants through the experiment using the three-step workflow shown in figure 2. When a new participant starts the experiment the *Experiment Manager* selects the initial step to show the participant and displays it to the participant. For example, in the *Generic IIR Research Protocol* this is the information and consent form. The participant reads the instructions on the page and answers any questions. They then submit their answers back to the system, which validates the answers against the answer schema defined



**Fig. 2.** The main loop implementing the *Experiment System*. Before showing the first step and then after each step the *Experiment Manager* determines the next step, based on the experiment workflow defined in the *Research Manager*, the steps seen so far, and the participant’s answers.

in the *Research Manager*. If the results do not match the schema, for example if a required question was not answered, or if the answer is invalid, then the applicable error messages are generated and the page show to the participant again, with their existing answers pre-filled. If the results are acceptable, then the answers are stored and the *Experiment Manager* uses the workflow defined in the *Research Manager* to determine which step to show next. This decision can take into account which steps the participant has completed, which *Task type / System* combination they were assigned to, and also what answers the participant has provided so far.

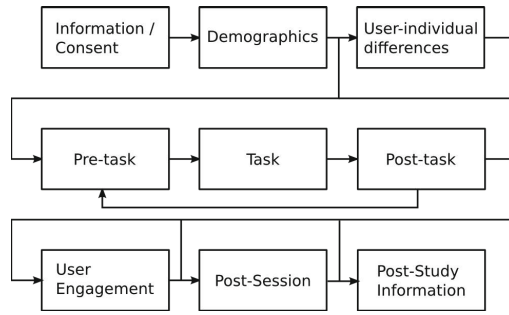
To ensure that the *Experiment System* can be used in a wide range of experiments, it does not itself include the task interface. At the *Task* steps in the experiment workflow, it simply loads the applicable task UI, as defined in the *Research Manager*, into the interface. A number of different techniques for the embedding are available, including an inline-frame-based, a simple re-direction-based, and a API-callback-based approach. This ensures that the framework can be deployed with most types of web-based UIs and can thus be widely used.

### 3.3 Data Extractor

The *Data Extractor* addresses requirements #3 in that it outputs the results from the experiment in a standardised format for further processing in analysis packages such as SPSS or R. In addition to the data acquired from the participants, the output also includes data on the *Task type / System* combinations the participants were shown. Simple post-processing steps, such as filtering columns or participant answers, can be applied to the data to reduce the amount of pre-processing required before loading the data into the analysis package.

### 3.4 Generic IIR Research Protocol

The *Generic IIR Research Protocol* supports requirements #2, #3, and #5 for the IIR evaluation context. By providing a standardised set of steps, ordering of those steps, and measures within those steps, it ensures that results from different studies become comparable. Because the standardised measures are pre-defined, it also reduces the resource commitment required to set up the experiment.



**Fig. 3.** The main work-flow through the *Generic IIR research Protocol*, showing the optional *user-individual differences* and *post-session* steps and also that the *pre-task – task – post-task* structure can be repeated multiple times within an experiment if the aim is to evaluate multiple tasks.

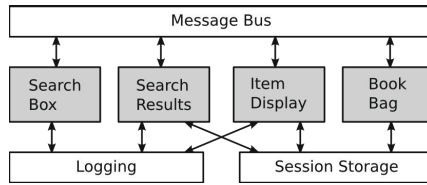
To be able to support the varied IIR evaluation landscape, it makes no constraint on the IIR UI that is under test, and it also allows the researcher to augment the process with the specific research questions they are interested in (in the *post-task* or *post-session* steps). The protocol has adapted and augmented the protocols used by early TREC Interactive Tracks and INEX Interactive Tracks, all of which are based on many earlier IIR studies. Some aspects have been extracted from more recent work. The main work-flow through the protocol is shown in figure 3 and consists of nine steps:

1. **Study information and Consent:** this is the typical introduction to a study together with a consent form that enables informed consent to be made (which is now expected and required for human-based experiments) and advises participants of their rights in participating. Most of the actual textual content is provided by the researcher when setting up the experiment in the *research manager*. However, because the basic protocol has received Research Ethics approval by Sheffield University, some of the content cannot be modified.
2. **Demographics Questionnaire:** a standard set of questions asked of all participants is used to create a profile of the set of participants in a study. A minimum set of standard variables is required (gender, age, education, cultural background, and employment) to ensure comparability across studies, and in some instances may help explain results (e.g., inexperienced, mostly of one gender, mostly undergraduates and so on). But additional experiment-specific variables can be added to the default set in the *research manager*;
3. **User-individual Differences:** depending on the study objectives, there is a large variety of user characteristics that one might observe, control or test, such as Cognitive Style [20], Need for Cognition [3], Curiosity [9], and Openness to Experience [14]. The basic research protocol does not include any of these as a default; we need more research to emphatically determine that any of these are core predictors of search actions and outcomes. The



*Generic IIR Research Protocol* defines a standard template to insert these into the experiment, but they will in the short term be study-specific. This customisation is available through the *Research Manager* which may be used to add scales or questions that are not currently specified by the protocol;

4. **Pre-Task Questions:** prior to assigning a participant to a task, the knowledge, experience and interest in the task topic is collected. For this, a set of standard questions derived from TREC and INEX interactive track protocols as well as other IIR studies was used [1]. These will be required, enabling the future comparison across studies. Unlike the implementation in TREC and INEX, the questions have been converted to standard Likert scales requesting agreement with statements;
5. **Task:** at this point in the procedure, the participants are shown the task UI. The UI may be created using our *Task Workbench* or the UI to any web-based system may be inserted. The system used is not discussed further in this paper, as search interfaces is a different topic. The system also handles the insertion of tutorial, and practice in the case of novel interfaces for which a participant may require training and some exposure;
6. **Post-Task Questions:** as with the *pre-task questions* a set of post-task questions also derived from past TREC and INEX interactive tracks, and reproduced in other studies, are integrated into the research protocol as a required step. These questions address the user-perception of completing the assigned task.
7. **User Engagement:** after completing all tasks, a set of post-session questions assesses the participants' engagement with the whole study. By default the *generic research protocol* provides the User Engagement Scale [15]. This scale measures six components of user experience, namely Focused Attention, Perceived Usability, Aesthetics, Endurability, Novelty, and Felt Involvement. At present, there is no competitor for this measure. While we recommend that it be included so that the scale can be further generalised and potentially improved, it is not a required feature.
8. **Post Study:** an additional but not required feature is the option of assessing the interface to the system used and/or the content. However researchers may substitute specific questions aimed at evaluating the whole session. For example in studies testing a novel IIR interface or component, questions evaluating the participants' interactions with the novel interface or component would be asked at this point.
9. **Post Study Information:** minimally this will contain acknowledgement and contact information. Optionally, the participants will also be able to sign up for future studies, with the goal of building up a pool of potential participants for future IIR evaluations. In this case, the system will collect contact information and a brief profile so that targeted recruitment may be conducted.



**Fig. 4.** The pluggable task work-bench provides three shared modules (*Message Bus*, *Logging*, and *Session Storage*) into which the actual evaluation UI components (sample shown in grey) are plugged.

### 3.5 Task Workbench

To further reduce the resource commitments (requirement #5) required to set up an IIR evaluation experiment, an extensible, pluggable task work-bench is provided (fig. 4). The task work-bench provides three standard modules (*Message Bus*, *Logging*, *Session Storage*) into which the experiment / task-specific components are plugged. Each component defines a set of messages it can send and listen for. The researcher then specifies which components should listen to which messages from which other components and the *message bus* ensures that the messages are correctly delivered. This means that new components can easily be integrated with existing components, simply by linking them via their messages.

```
{
  "participant": 322, "timestamp": "2013-02-13T14:34:23",
  "action": "query", "parameters": {"q": "Railwy"},
  "components": {
    "search_box": {"spelling": "Railway", "q": "Railwy"},
    "search_results": {"numFound": 4, "docs": [{...}, {...]}}
  }
}
```

**Fig. 5.** Example entry for the log-file generated by the *Task Workbench*. The entry shows that participant 322 sent a query “Railwy”, together with a list of those components that reacted to the query and what data they showed the participant.

The *Task Workbench* provides standard *logging* and *session storage* modules to simplify the creation of new components. A set of standard components (search box, search results, item display, task display, book-bag for collecting items) that can be re-used or extended. It also generates a very rich log file (fig. 5). In addition to the standard fields it also includes detailed information on which UI components were updated based on the request, and all the data that the updated UI components displayed to the participant. This makes it possible to fully re-play the participant’s interaction with the system.

## 4 Conclusion

In this paper we present a novel, standardised design and system for Interactive Information Retrieval (IIR) experiments, building on past implementations

[21,23,18,2]. The framework defines a standardised set of questions that enables the comparability of IIR evaluation results, while still being flexible enough to allow for the investigation of experiment-specific research questions. To reduce the resource requirements of setting up IIR evaluations the framework is supported through a number of extensible software components, that can easily be integrated with existing IIR systems. The goal of the framework is to achieve a level of standardisation in IIR that extends the comparability that Cranfield-style evaluation brought to IR in general to the IIR evaluation domain.

The system has successfully been deployed for the data-collection in the 2013 CLEF CHiC Interactive task [16] and also in the 2013 TREC Session Track [8]. It has also been used in non-IIR studies [5].

**Acknowledgements.** The research leading to these results was supported by the Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191. The authors gratefully acknowledge the advice and contributions of V. Petras, B. Larsen, and P. Hansen to the design.

## References

1. Trec 2002 interactive track guidelines. Technical report (2002)
2. Bierig, R., Gwizdka, J., Cole, M.: A user-centered experiment and logging framework for interactive information retrieval. In: Proceedings of the SIGIR 2009 Workshop on Understanding the User: Logging and Interpreting User Interactions in Information Search and Retrieval, pp. 8–11 (2009)
3. Cacioppo, J.T., Petty, R.E., Kao, C.F.: The efficient assessment of need for cognition. *Journal of Personality Assessment* 48(3), 306–307 (1984)
4. Gwizdka, J.: Distribution of cognitive load in web search. *Journal of the American Society for Information Science and Technology* 61(11), 2167–2187 (2010)
5. Hall, M., Clough, P., Stevenson, M.: Evaluating the use of clustering for automatically organising digital library collections. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) TPDFL 2012. LNCS, vol. 7489, pp. 323–334. Springer, Heidelberg (2012)
6. Hersh, W.: Trec 2002 interactive track report. In: Proc. TREC (2002)
7. Ingwersen, P., Järvelin, K.: The turn: Integration of information seeking and retrieval in context, vol. 18. Springer (2005)
8. Kanoulas, E., Hall, M., Clough, P., Carterette, B.: Overview of the trec 2013 session track. In: Proceedings of the Twentieth Text REtrieval Conference (TREC 2013) (2013)
9. Kashdan, T.B., Gallagher, M.W., Silvia, P.J., Winterstein, B.P., Breen, W.E., Terhar, D., Steger, M.F.: The curiosity and exploration inventory-ii: Development, factor structure, and psychometrics. *Journal of Research in Personality* 43(6), 987–998 (2009)
10. Kelly, D.: Measuring online information seeking context, part 1: background and method. *Journal of the American Society for Information Science and Technology* (14), 1862–1874 (2006)
11. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3(1), 1–224 (2009)

12. Kelly, D., Gyllstrom, K., Bailey, E.W.: A comparison of query and term suggestion features for interactive searching. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 371–378. ACM (2009)
13. Kelly, D., Sugimoto, C.: A systematic review of interactive information retrieval evaluation studies, 1967–2006. *JASIST* 64(4), 745–770 (2013)
14. Lee, K., Ashton, M.: The hexaco personality inventory: A new measure of the major dimensions of personality. *Multivariate Behavioral Research* 39, 329–358 (2004)
15. O’Brien, H.L., Toms, E.G.: The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology* 61(1), 50–69 (2009)
16. Petras, V., Hall, M., Savoy, J., Bogers, T., Malak, P., Toms, E., Pawlowski, A.: Cultural heritage in clef (chic) (2013)
17. Reips, U.-D.: Standards for internet-based experimenting. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie)* 49(4), 243–256 (2002)
18. Reips, U.-D., Lengler, R.: Theweb experiment list: A web service for the recruitment of participants and archiving of internet-based experiments. *Behavior Research Methods* 37(2), 287–292 (2005)
19. Renaud, G., Azzopardi, L.: Scamp: a tool for conducting interactive information retrieval experiments. In: Proceedings of the 4th Information Interaction in Context Symposium, pp. 286–289. ACM (2012)
20. Riding, R.J., Rayner, S.: Cognitive styles and learning strategies: Understanding style differences in learning and behaviour. D. Fulton Publishers (1998)
21. Tague-Sutcliffe, J.: The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management* 28(4), 467–490 (1992)
22. Toms, E.: Task-based information searching and retrieval, pp. 43–59. Facet Publishing (2011)
23. Toms, E.G., Freund, L., Li, C.: Wiire: the web interactive information retrieval experimentation system prototype. *Information Processing & Management* 40(4), 655–675 (2004)
24. Toms, E.G., Freund, L., Li, C.: Wiire: the web interactive information retrieval experimentation system prototype. *Information Processing & Management* 40(4), 655–675 (2004)
25. Toms, E.G., O’Brien, H., Mackenzie, T., Jordan, C., Freund, L., Toze, S., Dawe, E., MacNutt, A.: Task effects on interactive search: The query factor. In: Fuhr, N., Kamps, J., Lalmas, M., Trotman, A. (eds.) *INEX 2007*. LNCS, vol. 4862, pp. 359–372. Springer, Heidelberg (2008)
26. Toms, E.G., Villa, R., McCay-Peet, L.: How is a search system used in work task completion? *Journal of Information Science* 39(1), 15–25 (2013)
27. Yuan, W., Meadow, C.T.: A study of the use of variables in information retrieval user studies. *Journal of the American Society for Information Science* 50(2), 140–150 (1999)

# Improving Ranking Evaluation Employing Visual Analytics

Marco Angelini<sup>2</sup>, Nicola Ferro<sup>1</sup>, Giuseppe Santucci<sup>2</sup>, and Gianmaria Silvello<sup>1</sup>

<sup>1</sup> University of Padua, Italy

{ferro,silvello}@dei.unipd.it

<sup>2</sup> “La Sapienza” University of Rome, Italy

{angelini,santucci}@dis.uniroma1.it

**Abstract.** In order to satisfy diverse user needs and support challenging tasks, it is fundamental to provide automated tools to examine system behavior, both visually and analytically. This paper provides an analytical model for examining rankings produced by IR systems, based on the discounted cumulative gain family of metrics, and visualization for performing failure and “what-if” analyses.

## 1 Introduction

Information Retrieval (IR) systems, ranging from World Wide Web search engines to enterprise search or expertise retrieval systems and passing through information access components in wider systems such as digital libraries, are key technologies to get access to relevant information items in a context where information overload is day-to-day experience of every user.

In order to present this considerable amount of information to the user, IR systems rely on sophisticated ranking models where many different parameters affect the obtained results. Furthermore, they are comprised of several components interacting together in complex ways to produce a list of relevant documents in response to a user query. Ranking is a central and ubiquitous issue in this context since it is necessary to return the results retrieved in response to a user query according to the estimation of their relevance to that query. The interactions among the components of an IR system are often hard to trace down, to explain in the light of the obtained results, and to interpret in the perspective of possible modifications to be made to improve the ranking of the results, thus making this activity extremely difficult. This activity is usually called, in the IR field, *failure analysis* and it is deemed a fundamental activity in experimental evaluation even if it is too often overlooked due to its difficulty [1].

To give the reader an idea of how much demanding failure analysis can be, please consider the case of the the Reliable Information Access (RIA) workshop [4], which was aimed at investigating in a systematic way the behaviour of just one component in a IR system, namely the relevance feedback module. [4] reported that, for analysing 8 systems, 28 people from 12 organizations worked for 6 weeks requiring from 11 to 40 person-hours per topic for 150 overall topics.

Such a big effort was just aimed at understanding why a system behaved in a certain way. Nevertheless, in a real setting, after such inspection, you have to come back to design and development and implement the modifications and new features that the previous analysis suggested as possible solutions to the identified problems and, then, you have to start a new experimentation cycle to verify whether the newly added features actually give the expected contribution. Therefore, the overall process of improving an IR system is much more time and resource demanding than failure analysis alone.

The contribution of the paper is the design, implementation, and initial test of a Visual Analytics (VA) system, called Visual Analytics Tool for Experimental Evaluation (VATE<sup>2</sup>), which supports all the phases of the evaluation of an IR system, namely performance and failure analysis, greatly reducing the effort needed to carry them out by providing effective interaction with the experimental data. Moreover, VATE<sup>2</sup> introduces a completely new phase in the experimental evaluation process, called *what-if analysis*, which is aimed at getting an estimate of what could be the effects of a modification to the IR system under examination before needing to actually implement it.

The paper is organized as follows: Section 2 discusses related work. Section 3 describes how the analytical models for interaction we adopt to conduct failure analysis and what-if analysis. Section 4 explains how the visualization and interaction part works and gives an overview of VATE<sup>2</sup> and Section 5 presents an initial evaluation of the system conducted with experts of the field. Finally, Section 6 concludes the paper, pointing out ongoing research activities.

## 2 Related Work

The graded-relevance metrics considered in this paper are based on cumulative gain [5]; the Discounted Cumulated Gain (DCG) measures are based on the idea that documents are divided in multiple ordered categories, e.g. highly relevant, relevant, fairly relevant, not relevant. DCG measures assign a gain to each relevance grade and for each position in the rank a discount is computed. Then, for each rank, DCG is computed by using the cumulative sum of the discounted gains up to that rank. This gives rise to a whole family of measures, depending on the choice of the gain assigned to each relevance grade and the used discounting function.

A work that exploits DCG to support analysis is [8] where the authors propose the potential for personalization curve. The potential for personalization is the gap between the optimal ranking for an individual and the optimal ranking for a group. The curves plots the average nDCG's (normalized DCG) for the best individual, group and web ranking against different group size. These curves were adopted to investigate the potential of personalization of implicit content-based and behavior features. Our work shares the idea of using a curve that plots DCG against rank position, as in [5], but using the gap between curves to support analysis as in [8]. Moreover, the models proposed in this paper provide the basis for the development of VA environment that can provide us with: (i) a quick and

intuitive idea of what happened in a ranking list; (ii) an understanding of what are the main reasons of its perceived performances; and, (iii) the possibility of exploring the consequences of modifying the system characteristics through an interactive what-if scenario. The work presented here builds on a precedent work by the authors [1] refining the what-if model and introducing a validation with expert users.

### 3 The Models Behind VATE<sup>2</sup>

#### 3.1 Clustering via Supervised Learning

IR systems are seen as black boxes in experimental evaluation, because, in most cases, we can analyze the ranking lists produced by a system, but we cannot analyze the system which produced them. This means that we cannot modify a systems, run new and diversified tests to understand how the system behaves and how it can be improved. To this end we have to rely only on the outputted ranking lists and from these we need to infer how the system behave under specific conditions.

In this context machine learning based on supervised learning techniques can help because they are effective tools to automatically tune parameters and combine multiple evidences [6] and they can be employed starting from the rankings outputted by test systems. Supervised learning methods are feature-based and a widely-used list of features usually adopted by these techniques is described in [3].

The purpose of learning to rank techniques is to improve the original ranking model in order to obtain better performances or to grip on machine learning to build new and more effective ranking models. In VATE<sup>2</sup> we leverage on these techniques with a slightly different purpose; indeed, we use the produced ranking lists, the experimental collection and a machine learning algorithm (i.e. a classification algorithm based on regression trees) to learn a ranking model of a given IR system in order to thoroughly study it without actually having it available.

Most of the state-of-the-art learning to rank algorithms are “feature-based”, which means that they learn the optimal way of combining features extracted from topic-document pairs. So, the topic-document pairs under investigation are represented as vectors of features, representing the relevance of documents w.r.t. a given topic. We can divide the typical features used in learning to rank into three main categories: document-based, topic-based, and model-based. Document-based features are extracted from the given document; topic-based features are the same as the document-based but calculated on the text of the topic, and model-based features are the output of ranking models. In VATE<sup>2</sup> we adopt document-based and topic-based features and we do not consider the model-based ones. This choice derives from the fact that our goal is to learn the ranking model of a system in the most reliable way and not to improve their performances. The most used and reliable list of features used in learning to

rank framework is provided by the LEarning TO Rank (LETOR)<sup>1</sup> initiative run by Microsoft Research and proposed by Liu et al. in [7].

In this work we exploit this framework to learn the ranking model of the IR system under investigation in order to simulate the way in which it ranks the documents. Our aim is to support a “what if” investigation on the ranking list outputted by the system taken into account; the basic idea is to show how the ranking list and the DCG change when we move upward or downward a document in the list. To this purpose, the “cluster hypothesis” saying that “closely associated documents tend to be relevant to the same requests” [9] has to be taken into account; indeed, there can be a correlation in the ranking list between a document and its “closed associated documents”. We lever on the hypothesis that if we change the rank of a document also the cluster of documents associated with it will accordingly change their rank.

There are several algorithms for clustering as described in [2]. In this work we focus on the ranking of the considered documents and on how the ranking model can be improved. To this purpose we form the cluster for a target document by grouping together the documents which are similar from the considered ranking model point-of-view. Let us take into account a full result vector  $FV_j$  retrieved for a given query  $q_j$ , for each document  $FV_j[i]$  we create a cluster of documents  $C_i$  by: (i) employing a test IR system and submitting  $FV_j[i]$  as a query, thus retrieving a result vector  $FV_i$  of documents; (ii) determining  $C_i = FV_j \cap FV_i$ ; and, (iii) ranking the documents in  $C_i$  by employing the learned ranking model.

Therefore, we retrieve a result vector  $FV_i$  of relevant documents w.r.t.  $FV_j[i]$ , then we pick out only those documents which are in the original result vector (say  $FV_j$ ), and lastly we use the learned ranking model to order these documents accordingly to their “ranking” similarity to  $FV_j[i]$ . In this way, the higher a document is into the cluster  $C_i$ , the more similar it is to the target document  $FV_j[i]$ . We can see that the similarity measure is based on how the documents are seen by the learned ranking model.

In the end of this process, for each document  $FV_j[i]$  obtained by an IR system for a query  $q_j$ , we define a cluster of documents  $C_i$  ordered by their relevance with respect to  $FV_j[i]$ .

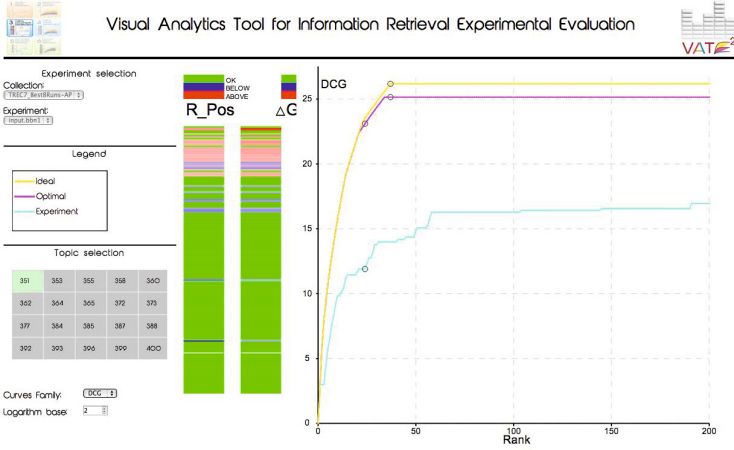
### 3.2 Rank Gain/Loss Model

According to [5] we model the retrieval results as a ranked vector of  $n$  documents  $V$ , i.e.  $V[1]$  contains the identifier of the document predicted by the system to be most relevant,  $V[n]$  the least relevant one. The ground truth  $GT$  function assigns to each document  $V[i]$  a value in the relevance interval  $\{0..k\}$ , where  $k$  represents the highest relevance score. Thus, the higher the index of a relevant document the less useful it is for the user; this is modeled through a discounting function  $DF$  that progressively reduces the relevance of a document,  $GT(V[i])$  as  $i$  increases. We do not stick with a particular proposal of  $DF$  and we develop

---

<sup>1</sup> <http://research.microsoft.com/en-us/um/beijing/projects/letor/>





**Fig. 1.** A Screen-shot of the failure analysis interface of VATE<sup>2</sup>

a model that is parametric with respect to this choice. However, to fix the ideas, we recall the original  $DF$  proposed in [5]:

$$DF(V[i]) = \begin{cases} GT(V[i]), & \text{if } i \leq x \\ GT(V[i]) / \log_x(i), & \text{if } i > x \end{cases} \quad (3.1)$$

that reduces, in a logarithmic way, the relevance of a document whose index is greater than the logarithm base.

The DCG function allows for comparing the performances of different IR systems, e.g. plotting the  $DCG(i)$  values of each IR system and comparing the curve behavior. However, if the user's task is to improve the ranking performance of a single IR system, looking at the misplaced documents (i.e. ranked too high or too low with respect to the other documents) the DCG function does not help, because the same value  $DCG(i)$  could be generated by different permutations of  $V$  and because it does not point out the loss in cumulative gain caused by misplaced elements. To this end, we introduce the following definitions and novel metrics.

Using the above definitions we can define the relative position  $R\_Pos(V[i])$  function for each document in  $V$  as follows:

$$R\_Pos(V[i]) = \begin{cases} 0, & \text{if } \min\_index(V, GT(V[i])) \leq i \leq \max\_index(V, GT(V[i])) \\ \min\_index(V, GT(V[i])) - i, & \text{if } i < \min\_index(V, GT(V[i])) \\ \max\_index(V, GT(V[i])) - i, & \text{if } i > \max\_index(V, GT(V[i])) \end{cases} \quad (3.2)$$

$R\_Pos(V[i])$  allows for pointing out misplaced elements and understanding how much they are misplaced: 0 values denote documents that are within the optimal interval, negative values denote elements that are below the optimal interval (pessimistic ranking), and positive values denote elements that are above the optimal (optimistic ranking). The absolute value of  $R\_Pos(V[i])$  gives the minimum distance of a misplaced element from its optimal interval.

According to the actual relevance and rank position, the same value of  $R\_Pos(V[i])$  can produce different variations of the DCG function. We measure the contributions of misplaced elements with the function  $\Delta\_Gain(V, i)$  which quantifies the effect of a misplacement in the overall computation of DCG. The  $\Delta\_Gain(V, i)$  function can assume both positive and negative values, where negative values correspond to elements that are presented too early (with respect to, their relevance) to the user and positive values to elements that are presented too late.

### 3.3 What-if Analysis Model

The clusters of documents defined above play a central role in the document movement estimation of VATE<sup>2</sup>. Indeed, once a user spots a misplaced document, say  $d_4$ , and s/he decides to move it upward or downward, also the ten documents in the  $C_4$  cluster are moved accordingly. The current implementation of VATE<sup>2</sup> employs the simple linear movement strategy where the movement of the document and the related document cluster happens according to a straightforward algorithm that tries to move the documents in the cluster of the same amount of positions as the document dragged and dropped by the user. However, this is not always possible since, for example, a document in the cluster might be ranked higher than the document selected by the user and may not exist enough space on the top of the ranking to place it; in this and similar cases, the movement algorithm “compresses” the movement of the documents in the cluster, approximating at its best the user intent.

The retrieval results are modeled as a ranked vector  $V$  containing the first 200 documents of the full result vector  $FV$ . The clustering algorithm we described, associates to each document  $V[i]$  a cluster  $C_i$  of similar documents (we consider only the documents whose relevance with  $V[i]$  is greater than a suitable threshold). Moreover, for the sake of notation we define the index cluster set  $IC_i$ , i.e., the set of indexes of  $FV$  corresponding to elements in  $C_i$ :  $IC_i = \{j | FV[j] \in C_i\}$ . As a consequence, according to the “cluster hypothesis”, moving up or down the document  $V[i]$  will affect in the same way all the documents in  $C_i$  and that might result in rescuing some documents below the 200 threshold pushing down some documents that were above such threshold.

We model the what-if interaction with the system with the operator  $Move(i, j)$  whose goal is to move the element in position  $i$  in position  $j$ . In order to understand the effect on  $V$  of such an operation, we have to consider all the  $C_i$  elements and the relative position of their indexes, that ranges between  $min(IC_i)$  and  $max(IC_i)$ . Different cases may occur and we analyze them assuming, without loss of generality, that  $i < j$ , i.e., that the analyst goal is to move up the element  $V[i]$  of  $j - i$  positions. For the clustering hypothesis that implies that all the  $C_i$  elements will move up of  $j - i$  positions as well. There are, however, situations in which that is not possible: the maximum upshift is  $max(min(IC_i) - 1, j - i)$  and if  $j - i > min(IC_i) - 1$  the best we can do is to move up all the  $C_i$  elements of just  $IC_i - 1$  positions. That corresponds to the situation in which the analyst wants to move up the element in position  $i$  of  $k$  positions, but there exists a

document in  $C_i$  whose index is  $\leq k$  and, obviously, it is not possible to move it up of  $k$  positions. In such a case, the system moves up all the documents in the cluster of  $\min(IC_i) - 1$  positions, approximating the user intent.

## 4 Overview of VATE<sup>2</sup>

VATE<sup>2</sup> allows the analyst to perform three main activities: performance analysis, failure analysis and what-if analysis by employing the models described above. These three main activities can be carried out at the “topic level” or at the “experiment level”.

At the topic level VATE<sup>2</sup> takes as input the ranked document list for the topic  $t$  and the ideal ranked list, obtained choosing the most relevant documents in the collection  $D$  for the topic  $t$  and ordering them in the best way. At the experiment level VATE<sup>2</sup> evaluates the overall quality of the ranking for all the topics of the experiment, focusing on the variability of the results. Basically, at the experiment level VATE<sup>2</sup> shows an aggregate representation based on the boxplot statistical tool showing the variability of the DCG family of metrics calculated on all the topics considered by an experiment. In this way the analyst will have a clearer insight on what to expect from her/his ranking algorithm both in a static way and in a dynamic one (which involves an interactive reordering of the ranked list of documents).

While visually inspecting the ranked list (i.e. failure analysis), it is possible to simulate the effect of interactively reordering the list, moving a target document  $d$  and observing the effect on the ranking while this shift is propagated to all the documents of the cluster containing the documents similar to  $d$  (i.e. what-if analysis). This cluster of documents simulates the “domino effect” within the given topic  $t$ . When the analyst is satisfied with the results, i.e. when he has produced a new ranking of the documents that corresponds to the effect that is expected by modifications that are planned for the system, he can feed the Clustering via Supervised Learning model with the newly produced ranked list, obtain a new model which takes into account the just introduced modifications, and inspecting the effects of this new model for other topics. This re-learning phase simulates the “domino effect” on the other topics different from  $t$  caused by a possible modification in the system.

### 4.1 How to Perform the Failure Analysis

Figure 1 shows the DCG Graph for the topic level analysis. On the left side we can see two vertical bars representing the visualization of the ranking list. The first one represents the  $R\_Pos$  vector. The visualization system computes the optimal ranking list of the documents and assigns to each document a color based on its rank. A green color is assigned to a document at the correct rank w.r.t. the calculated optimal rank; whereas a blue color is assigned to a document ranked below the optimal and a red color is assigned to a document ranked above the optimal. The color intensity gives the user an indication of how far the document

is from its optimal rank: a weak intensity means that the document is close to the optimal, a strong intensity means it is far to the optimal. The second vertical bar represents the  $\Delta\_Gain$  function values for each document. We adopted the same color code as in the previous vector, but in this case the red color represents a loss and a blue color represents a gain in terms of  $\Delta\_Gain$ .

On the right side of Figure 1 we can see a graph showing three curves:

**Experiment Ranking** refers to the top  $n$  ranked results provided by the system under investigation;

**Optimal Ranking** refers to an optimal re-ranking of the experiment;

**Ideal Ranking** refers to the ideal ranking of the top  $n$  documents in the pool.

The visualization system is built in such a way that if a user selects a document in the  $R\_Pos$  vector, also the DCG loss/gain in the  $\Delta\_Gain$  vector and all its contributions to the different curves (i.e. Experiment, Optimal and Ideal) will be highlighted.

The visualization described so far is well-suited to cope with a static analysis of the ranked result: the user can understand if there is the need to re-rank the documents or to perform a re-querying to retrieve a different set of documents with the aim of obtaining a better value of the DCG metric.

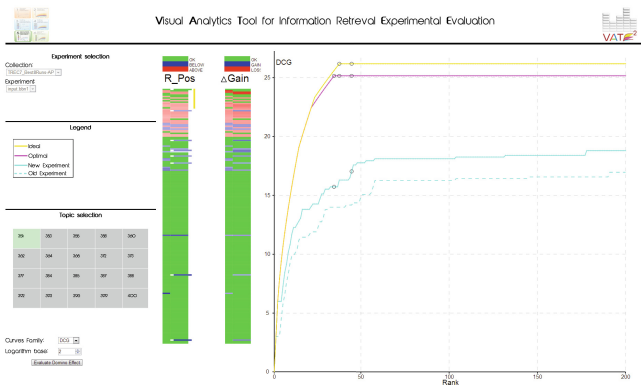
## 4.2 How to Perform the What-if Analysis

The what-if functionality allows the users to interact with the ranked vector of  $R\_Pos$ . The system allows the user to shift a target document  $t$  from its actual position to a new one in a “drag&drop” fashion, with the goal of investigating the effect of this movement in the ranking algorithm by inspecting the DCG of the modified ranking list. Clearly, a change in the ranking algorithm will affect not only the target document  $t$ , but also all the documents in its cluster.

In Figure 2 it is possible to see the animated phase of interactive re-ranking of the documents at the topic level: after highlighting and moving the target document  $t$  from the starting position to a new one, the user will be presented with an animated re-ranking of the documents connected to the target one. Once the new position of the target document has been selected, the system moves it to the new position and the documents in its associated cluster are moved together into their new positions. This leads to the redrawing of the  $R\_Pos$ ,  $\Delta\_Gain$  and DCG graphs according to the new values assigned to each document involved in the ranking process.

It is possible to see that when a user select a document in the leftest bar, all the documents in its cluster are highlighted in yellow helping the user to understand which documents are involved in a potential movement.

Figure 2 shows also the result of the what-if process: the image presents two new curves, representing the new values assigned for both the experiment curve (purple one) and the optimal curve (orange one). To evaluate the changes in the DCG function, the image shows, in a dash-stroke fashion, the old curve trends. Thanks to this visualization, the user can appreciate the gain or the



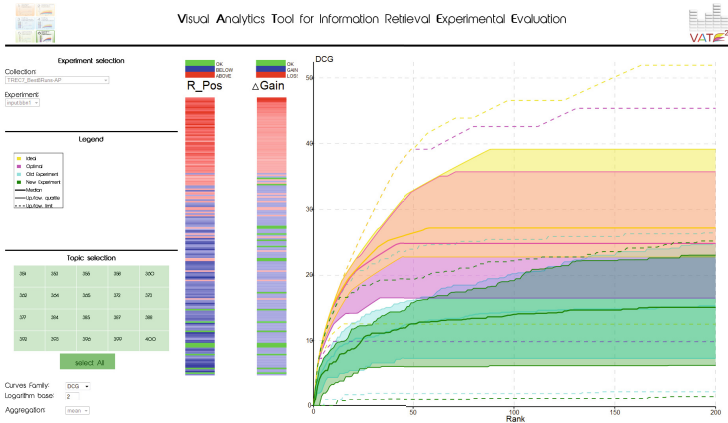
**Fig. 2.** A Screen-shot of the topic level what-if analysis interface of VATE<sup>2</sup>

loss obtained from this particular re-rank. In the case shown in Figure 2 the movements performed by the user improved the performances at the topic level; indeed, the dashed line – i.e. the old experiment curve – is lower than the solid one – i.e. the new experiment curve. This means that we are simulating a change in the system that does improve it. On top of that, at the experiment level, the change in the ordering of a particular ranking list will result in changing also the other ranking lists within the same experiment: these changes can be intercepted by this graph in terms of variability of the curves and on the raising/declining of the “box” region of the boxplots (showed as filled area in the graph).

To maintain the graph as clear as possible, the choice of not representing the single boxplots, but simply the continuous lines joining the similar points has been taken. So, in the graph area there are five different curves which are: upper limit, upper quartile, median, lower quartile, and lower limit. All these curves are determined for the ideal, the optimal and the experiment cases. For each case, the area between lower and upper quartile is color filled in order to highlight the central area (the box of the boxplot) of the analysis.

In figure 3 we can appreciate that, in this particular case, the optimal and experiment areas do not overlap very much, and the median curve of the experiments is quite far from the one of the optimal. This can be asserted from an aggregate point of view, and not by a specific topic analysis like the one we proposed with the DCG graph. Different considerations can also be made on variability: in this case, while experiment and optimal box areas are quite broad, demonstrating a heterogeneity in values, and also the ideals box area is big meaning a high variability of the data among the different topics.

The domino effect due to the what-if analysis is highlighted by the experiment areas: the old one (before the what-if analysis) is shaded in blue, whereas the new one (after the what-if analysis) is shaded in green. We can see that a change in one topic at the topic level worsens the global performances; indeed, the blue area is better than the green one. This means that the change the user did at the



**Fig. 3.** A Screen-shot of the experiment level what-if analysis interface of VATE<sup>2</sup>

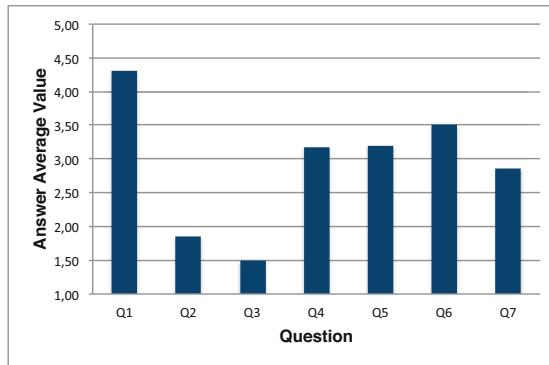
topic level (which improved the local performances) reflects at the experiment (global) level worsening the overall performances of the system.

## 5 Initial Validation with Experts

VATE<sup>2</sup> has been tested in a laboratory setting involving 13 experts (i.e. academics, post-docs and PhD students) in IR. The functioning of VATE<sup>2</sup> was described by means of an oral presentation where its peculiar functions were explained. This introduction was necessary to get the experts to know the system and to let them understand how to use it. The performance analysis part as well as the failure analysis one are more straightforward and close to the day to day experience of the experts; whereas, the what if analysis evaluation represents a totally new paradigm which requires some time to be properly understood.

The study was conducted by allowing the experts to freely use VATE<sup>2</sup> for an hour and, at the end, by asking them to compile a questionnaire. The questionnaire was divided into seven parts, one for each interface and one for an overall evaluation of VATE<sup>2</sup> as a whole. Every part repeated the following seven questions referring to the specific functionality under evaluation:

- Q1.** Is the addressed problem relevant for involved stakeholders (researchers and developers)?
- Q2.** Are the currently available tools and techniques adequate for dealing with the addressed problem?
- Q3.** Do currently available tools and techniques for dealing with the addressed problem offer interactive visualizations?
- Q4.** Is the proposed visual tool understandable?



**Fig. 4.** The histogram reporting the average answers of the experts evaluating VATE<sup>2</sup> as a whole

- Q5.** Is the proposed visual tool suitable and effective for dealing with the addressed problem?
- Q6.** To what extent the proposed visual tool is innovative with respect to the currently available tools and techniques?
- Q7.** To what extent the proposed visual tool will enhance the productivity of involved stakeholders (researchers and developers)?

The first three questions regard the scientific relevance of VATE<sup>2</sup> and they are aimed to understand if the experts think the problem addressed is relevant and if there exist other tools with the same purpose. The last four questions are aimed to understand if the experts think VATE<sup>2</sup> is useful for experimental evaluation and if it is well-suited for its purposes. Every answer was graded from 1 to 5, where 1 stand for “not at all” and 5 for “quite a lot”. In Figure 4 we report the average results of the questionnaire regarding the overall part which allows us to understand what the experts think about VATE<sup>2</sup> as a whole.

We can see that the problem addressed is of high relevance for the involved stakeholder (question 1) and that there not exist any other tool doing the work of VATE<sup>2</sup>. Indeed, answers to questions 2 and 3 are both below 2 as an average value which means that VATE<sup>2</sup> proposes something totally new in the field. Questions 4 to 6 report that the tool is understandable, suitable and effective for dealing with the addressed problem, and innovative. The last question is about productivity; on average the experts think VATE<sup>2</sup> can improve productivity but the answer is not clear like for the other questions. We think this is due to the time necessary to learn how to effectively use the system. By analyzing the results of every single part we see that experts think that VATE<sup>2</sup> improves productivity for performance analysis and failure analysis, but it is less clear if it is useful for what-if analysis which as explained above is a brand new topic in IR evaluation and probably it requires more time to become useful to the experts.

## 6 Conclusion and Future Work

This paper presented a fully-fledged analytical and visualization model to support interactive exploration of IR experimental results with a two-fold aim: (i) to ease and support deep failure analysis in order to better understand system behavior; (ii) to conduct a what-if analysis to have an estimate of the impact that possible modifications to the system, identified in the previous step and aimed at improving the performances, can have before needing to actually re-implement the system.

Future work will concern two main issues: (i) while the informal results about the system usage are quite encouraging we plan to run a more structured user study, involving people that have not participated in the system design; and (ii) we want to improve the way in which the clusters produced by the The Clustering via Supervised Learning methods are used to compute the new ranking and the associated DCG functions.

**Acknowledgements.** The work reported in this paper has been supported by the PROMISE network of excellence (contract n. 258191) project as a part of the 7th Framework Program of the European commission (FP7/2007-2013).

## References

1. Angelini, M., Ferro, N., Santucci, G., Silvello, G.: Visual Interactive Failure Analysis: Supporting Users in Information Retrieval Evaluation. In: Proc. of the 4th Information Interaction in Context Symposium, IIX 2012, pp. 194–203. ACM, New York (2012)
2. Berkhin, P.: A Survey of Clustering Data Mining Techniques. In: Kogan, J., Nicholas, C., Teboulle, M. (eds.) *Grouping Multidimensional Data*, pp. 25–71. Springer, Heidelberg (2006)
3. Geng, X., Liu, T.-Y., Qin, T., Li, H.: Feature Selection for Ranking. In: Kraaij, W., de Vries, A.P., Clarke, C.L.A., Fuhr, N., Kando, N. (eds.) *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, pp. 407–414. ACM Press, New York (2007)
4. Harman, D., Buckley, C.: Overview of the Reliable Information Access Workshop. *Information Retrieval* 12(6), 615–641 (2009)
5. Järvelin, K., Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information System (TOIS)* 20(4), 422–446 (2002)
6. Liu, T.-Y.: Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval* 3(3), 225–331 (2009)
7. Liu, T.-Y.Y., Xu, J., Qin, T., Xiong, W., Li, H.: LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval. In: Joachims, T., Li, H., Liu, T.-Y., Zhai, C. (eds.) *SIGIR 2007 Workshop on Learning to Rank for Information Retrieval* (2007)
8. Teevan, J., Dumais, S.T., Horvitz, E.: Potential for Personalization. *ACM Transactions on Computer-Human Interaction (TOCHI)* 17(1), 1–31 (2010)
9. van Rijsbergen, C.J.: *Information Retrieval*, 2nd edn. Butterworths, London (1979)



# A Proposal for New Evaluation Metrics and Result Visualization Technique for Sentiment Analysis Tasks\*

Francisco José Valverde-Albacete<sup>1</sup>,  
Jorge Carrillo-de-Albornoz<sup>1</sup>, and Carmen Peláez-Moreno<sup>2</sup>

<sup>1</sup> Departamento de Lenguajes y Sistemas Informáticos  
Univ. Nacional de Educación a Distancia, c/ Juan del Rosal, 16. 28040 Madrid, Spain  
{fva,jcalbornoz}@lsi.uned.es

<sup>2</sup> Departamento de Teoría de la Señal y de las Comunicaciones  
Universidad Carlos III de Madrid, 28911 Leganés, Spain  
carmen@tsc.uc3m.es

**Abstract.** In this paper we propound the use of a number of entropy-based metrics and a visualization tool for the intrinsic evaluation of Sentiment and Reputation Analysis tasks. We provide a theoretical justification for their use and discuss how they complement other accuracy-based metrics. We apply the proposed techniques to the analysis of TASS-SEPLN and RepLab 2012 results and show how the metric is effective for system comparison purposes, for system development and postmortem evaluation.

## 1 Introduction

The appropriate evaluation of multi-class classification is a founding stone of Machine Learning. For Sentiment and Reputation Analysis (SA and RA), where different polarities—for instance *positive*, *neutral*, *negative*—and several degrees of such polarities may be of interest, it is a crucial tool.

However, accuracy-based methods in predictive analytics suffer from the well-known accuracy paradox, viz. a high level of accuracy is not a necessarily an indicator of high classifier performance [1, 2, 3]. In other words, a high accuracy figure does not necessarily imply that the classifier has been able to model the underlying phenomena.

Since accuracy-improving methods try to improve the *heuristic rule* of minimizing the number of errors, we have to question whether rather than a shortcoming of accuracy, this paradox might be a *shortcoming of the heuristic*.

An alternative heuristic is to maximize the information transferred from input to output through the classification process, as described by the contingency matrix. In [4] an information-theoretic visualization scheme was proposed,

---

\* FJVA and JCdA are supported by EU FP7 project LiMoSINe (contract 288024). CPM has been partially supported by the Spanish Government-Comisión Interministerial de Ciencia y Tecnología project TEC2011-26807 for this paper.

the *entropy triangle*, where the *mutual information (MI)* of the contingency matrix is related to the distance of the input and output distributions from uniformity and to the *variation of information* [5], another distance measuring how much information from input was not learnt and how much information at the output is not predicted by the classifier.

Unfortunately, MI is expressed in bits, not in efficiency, and this detracts from its intended reading as a metric. Furthermore, it is actually one aspect of a tripolar manifestation [4], hence not adequate as a *binary* indicator of goodness. Also, it measures how well has the classifier learnt the input distribution, but not what its expected accuracy is.

On the other hand, the Normalized Information Transfer (NIT) factor [6] is a measure that relates to MI in the same way that the reduction in perplexity of a language model relates to the entropy of a source: it quantifies how well the classifier has done its job of reducing the uncertainty in the input distribution. This reading allows us to justify an Entropy-Modulated Accuracy that can be used as a complement to more standard, error-based metrics, like precision, recall or F-score.

In the following we introduce more formally these two tools (Section 2) and apply them to the systems that took part in the last TASS-SEPLN and RepLab 2012 campaigns (Section 3). We conclude with some suggestions for their use.

## 2 The Entropy Triangle and the Normalized Information Transfer

### 2.1 The Entropy Triangle: A Visualization Tool

The entropy triangle is a contingency matrix visualization tool based on an often overlooked decomposition of the joint entropy of two random variables[4]. Figure 1 shows such a decomposition showing the three crucial regions:

- The *mutual information*,

$$MI_{P_{XY}} = H_{P_X \cdot P_Y} - H_{P_{XY}}$$

- The *variation of information*, the addition of the conditional perplexities on input and output [5],

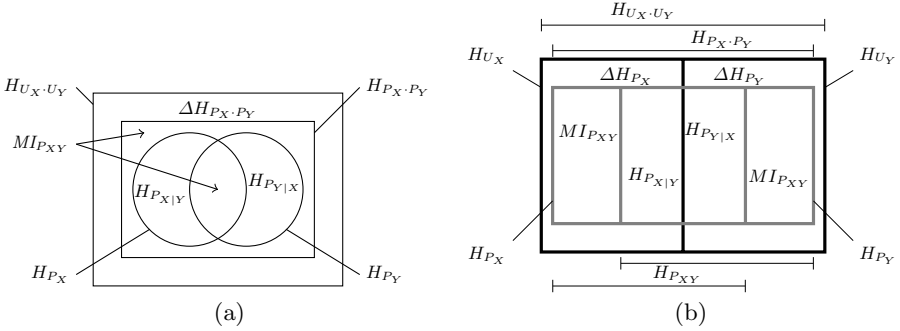
$$VI_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} \quad (1)$$

- And the *entropy decrement* between the uniform distributions with the same cardinality of events as  $P_X$  and  $P_Y$  and the entropy of the joint distribution where both are independent,

$$\Delta H_{P_X \cdot P_Y} = H_{U_X \cdot U_Y} - H_{P_X \cdot P_Y} \quad (2)$$

Note that all of these quantities are positive. In fact from the previous decomposition the following *balance equation* is evident,

$$\begin{aligned} H_{U_X \cdot U_Y} &= \Delta H_{P_X \cdot P_Y} + 2 * MI_{P_{XY}} + VI_{P_{XY}} \\ 0 &\leq \Delta H_{P_X \cdot P_Y}, MI_{P_{XY}}, VI_{P_{XY}} \leq H_{U_X \cdot U_Y} \end{aligned} \quad (3)$$



**Fig. 1. Extended entropy diagrams related to a bivariate distribution, from [4].** The bounding rectangle is the joint entropy of two uniform (hence independent) distributions  $U_X$  and  $U_Y$  of the same cardinality as input probability distribution  $P_X$  and output  $P_Y$ , resp. The expected mutual information  $MI_{P_{XY}}$  appears *twice* in (a) and this makes the diagram split for each variable symmetrically in (b).

where the bounds are easily obtained from distributional considerations. If we normalize (3) by the overall entropy  $H_{U_X \cdot U_Y}$  we obtain the equation of the 2-simplex in entropic space,

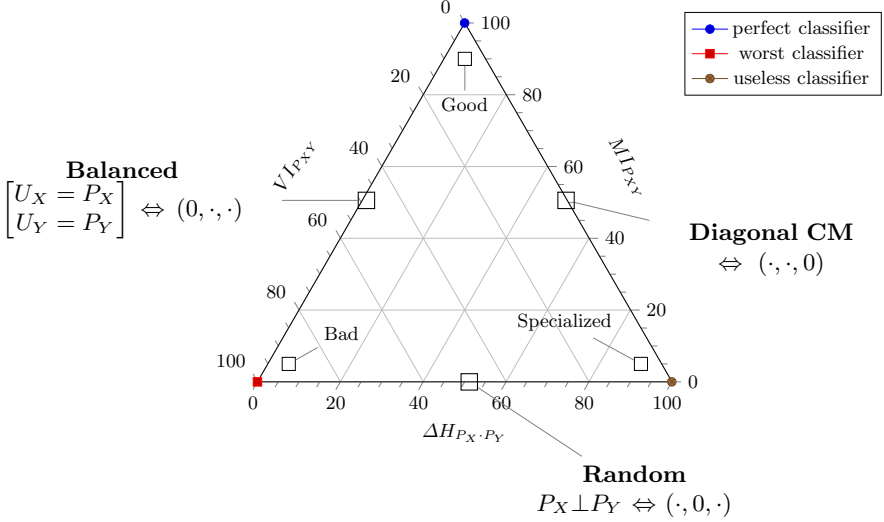
$$\begin{aligned}
 1 &= \Delta' H_{P_X \cdot P_Y} + 2 * MI'_{P_{XY}} + VI'_{P_{XY}} \\
 0 &\leq \Delta' H_{P_X \cdot P_Y}, MI'_{P_{XY}}, VI'_{P_{XY}} \leq 1
 \end{aligned}
 \tag{4}$$

representable by a De Finetti or ternary entropy diagram or simply *entropy triangle (ET)*.

The evaluation of classifiers is fairly simple using the schematic in Fig. 2.

1. Classifiers on the bottom side of the triangle *transmit no mutual information* from input to output: they have not profited by being exposed to the data.
2. Classifiers on the right hand side have diagonal confusion matrices, hence *perfect (standard) accuracy*.
3. Classifiers on the left hand side operate on perfectly balanced data distributions, hence they are *solving the most difficult multiclass problem* (from the point of view of an uninformed decision).

Of course, combinations of these conditions provide specific kinds of classifiers. Those at the apex or close to it are obtaining the highest accuracy possible on very balanced datasets and transmitting a lot of mutual information hence they are the *best classifiers* possible. Those at or close to the left vertex are essentially not doing any job on very difficult data: they are *the worst classifiers*. Those at or close to the right vertex are not doing any job on very easy data for which they claim to have very high accuracy: they are *specialized (majority) classifiers* and our intuition is that they are the kind of classifiers that generate the accuracy paradox [1].



**Fig. 2. Schematic Entropy Triangle showing interpretable zones and extreme cases of classifiers.** The annotations on the center of each side are meant to hold for that whole side.

In just this guise, the ET has already been successfully used in the evaluation of Speech Recognition systems [4, 7]. But a simple extension of the ET is to endow it with a graduated axis or colormap that also allows us to visualize the correlation of such information-theoretic measures with other measures like accuracy, greatly enhancing its usefulness. Examples of its use can be seen in Figs. 3 and 4, and this is the main tool we propose to complement other Sentiment Analysis metrics.

## 2.2 The Normalized Information Transfer (NIT) Factor and the Entropy-Modified Accuracy (EMA)

The problem with the ET is that in spite of being helpful as a visualization and exploration tool, it does not allow for system ranking at the heart of modern competitive, task-based evaluation. For such purposes we use a corrected version of the accuracy and a measure derived from mutual information.

A measure of the *effectiveness of the learning process* is the *information transfer factor*  $\mu_{XY} = 2^{MI_{P_{XY}}}$  but we prefer to report it as a fraction of the number of classes, the *Normalized Information Transfer factor (NIT)*,

$$q(P_{XY}) = \frac{\mu_{XY}}{k} = 2^{MI_{P_{XY}} - H_{U_X}} \quad (5)$$

The NIT is explained in the context of the perplexity of the classifier [6]. The quantity  $\mu_X = 2^{MI_{XY}}$  is interpreted there as the reduction in the number of classes afforded by a classifier on average, as seen from the point of view of an

uninformed decision: the higher this reduction, the better. In the worst case—random decision—, this reduction is  $MI_{P_{XY}} = 0$ ,  $2^{MI_{P_{XY}}} = 1$  whence the NIT is  $1/k$ . In the best possible case (perfect classifier, balanced class distribution) this reduction is  $MI_{P_{XY}} = \log_2 k$ ,  $2^{MI_{P_{XY}}} = k$ , whence the normalized rate is 1 so that the range of the NIT factor is  $1/k \leq q((P_{XY})) \leq 1$  matching well the intuition that a random decision on a balanced data set can only guess right  $1/k$  of the times on average but the best informed decision guesses right always.

Considering the two paragraphs above,  $k_{X|Y} = 2^{H_{P_{X|Y}}}$  can be interpreted as the *remnant number of equiprobable classes* seen by the classifier (after learning the task). But  $k_{X|Y}$  is precisely the number of equiprobable classes the classifier sees after subtracting the NIT, whence the *entropy-modulated accuracy (EMA)* of the classifier would be

$$a'(P_{XY}) = 1/k_{X|Y} = 2^{-H_{P_{X|Y}}}$$

We can see that the EMA is corrected by the input distribution and the learning process, i.e. the more efficient the learning process, the higher the NIT and the higher the EMA but, the more imbalanced the input class distribution, the lower  $k_X$  and the higher the EMA.

Note that this last commentary makes the EMA a suspicious metric: classifiers should only be compared when the effective perplexities of the tasks they are applied to are comparable, that is, with similar  $k_X$ . For classifiers across tasks, then, the NIT is a better measure of success, although when measuring performance *on the same task*, modified accuracy is a good metric. In the following, we will report both.

### 3 Experiments and Evaluation

#### 3.1 Sentiment Analysis in TASS-SEPLN

The aim of the TASS-SEPLN competition was to classify tweets into different degrees of *Sentiment* polarity. The data consists of tweets, written in Spanish by nearly 200 well-known personalities and celebrities of the world [8]. Each tweet is tagged with its global polarity, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. Five levels have been defined: *strong positive* (P+), *positive* (P), *neutral* (NEU), *negative* (N), *strong negative* (N+) and one additional *no sentiment* tag (NONE). Table 1 shows the distribution of these classes in the training and test sets, and their effective perplexities: the training sets are much more balanced.

In TASS-SEPLN, polarity classification is evaluated as two different tasks. The goal of TASS5 is to automatically classify each of the tweets into one of the 5 polarity levels mentioned above. However, prior to this classification, the task requires to filter out those tweets not expressing any sentiment (i.e., those tagged as NONE), so the number of classes is  $k = 6$ . TASS3 consists in classifying each tweet in 3 polarity classes (*positive*, *neutral* and *negative*). To this end, tweets

**Table 1.** Distribution of tweets per polarity class in the TASS corpus

TASS5	P+	P	NEU	N	N+	NONE	TOTAL	$k_X$
training	1 764	1 019	610	1 221	903	1 702	7 219	5.6
testing	20 745	1 488	1 305	11 287	4 557	21 416	60 798	4.1

TASS3								
training		2 783	610	2 124		1 702	7 219	3.6
testing		22 233	1 305	15 844		21 416	60 798	3.2

tagged as positive and strong positive are merged into a single category (*positive*), and tweets tagged as negative and strong negative into another (*negative*). This task is called TASS3 but has  $k = 4$ .

Table 2 shows the numeric results of the different metrics on the (a) TASS3 and (b) TASS5 tasks. These data reveal that the EMA is much lower than normal accuracy and that there would be some reordering of the ranking if EMA was the ranking criterion. In particular, some sets of submissions are systematically pushed downwards in the table according to EMA. These phenomenon warrants some postmortem analysis of the results of such systems.

Furthermore, some systems, specifically those with  $\mu_{XY} \approx 1.000$ , essentially took random decisions but their accuracies were well above random. This is a strong result that shows the inadequacy of accuracy for such evaluations.

Figure 3 presents the ET visualization of the performance of the different systems at either task, revealing some interesting results. First, in both tasks four systems are closer to the upper vertex of the triangle implying a better behaviour than the others. However, their distance to the apex of the ET indicates that even these systems are still far from solving the task, that is, being able to model the different polarities captured in the data, even though the best accuracy is 72.3% in TASS3, 67.8% in TASS5. This is another strong hint that *high accuracy does not correlate with high performance in the task*. Furthermore, the triangles show that two systems (correlative submissions in either tasks) are placed very close to the base of triangle, which suggests both random decision and specialization as majority classifiers, despite their achieving an accuracy of around 35% in both tasks. These are the very same systems with  $\mu_{XY} \approx 1.000$ .

Second, while the accuracy of the systems is better in TASS3 than in TASS5 (as expected, since the complexity of the problem increases with the number of classes), the evaluation according to the ET shows that the behaviour of the systems is, in practice, the same in both tasks. In our opinion, the explanation can be found in the evaluation methodology and distribution of classes in the dataset: for TASS3, *positive* and *strong positive* tweets are merged in a single category, and *negative* and *strong negative* tweets are merged in another category. But since the number of tweets in the *positive* and *strong negative* categories is very low in comparison with the number of tweets in the remaining categories, the effect of misclassifying tweets of these two categories in TASS5 is not that marked, in terms of accuracy.

**Table 2. Perplexities, accuracy ( $a$ ), EMA ( $a'_X$ ) and NIT factor ( $q_X$ ) for the TASS test runs.** . The ranking by accuracy (official) and by EMA have some inversions (red=should sink, green=should rise).

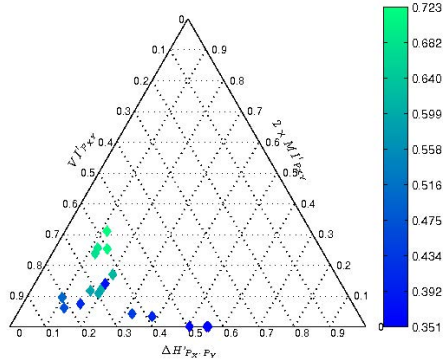
(a) TASS3: $k = 4, k_X = 3.2$					(b) TASS5: $k = 6, k_X = 3.2$						
TASS3 run	$k_{X Y}$	$\mu_{XY}$	$a$	$a'_X$	$q_X$	TASS5 run	$k_{X Y}$	$\mu_{XY}$	$a$	$a'_X$	$q_X$
daedalus-1	2.090	1.539	0.723	0.478	0.385	daedalus-1	2.413	1.705	0.678	0.414	0.284
elhuyar-1	2.265	1.420	0.711	0.441	0.355	elhuyar-1	2.664	1.545	0.653	0.375	0.257
l2f-1	2.258	1.424	0.691	0.443	0.356	l2f-1	2.625	1.567	0.634	0.381	0.261
l2f-3	2.256	1.426	0.690	0.443	0.356	l2f-3	2.620	1.570	0.633	0.382	0.262
l2f-2	2.312	1.391	0.676	0.432	0.348	l2f-2	2.734	1.505	0.622	0.366	0.251
atrilla-1	2.541	1.266	0.620	0.394	0.316	atrilla-1	3.077	1.337	0.570	0.325	0.223
sinai-4	2.706	1.189	0.606	0.370	0.297	sinai-4	3.432	1.199	0.547	0.291	0.200
uned1-1	2.735	1.176	0.590	0.366	0.294	uned1-2	3.505	1.174	0.538	0.285	0.196
uned1-2	2.766	1.163	0.588	0.362	0.291	uned1-1	3.454	1.191	0.525	0.290	0.199
uned2-1	2.819	1.141	0.501	0.355	0.285	uned2-2	3.809	1.080	0.404	0.263	0.180
imdea-1	2.953	1.089	0.459	0.339	0.272	uned2-1	3.395	1.212	0.400	0.295	0.202
uned2-2	3.033	1.061	0.436	0.330	0.265	uned2-3	3.865	1.064	0.395	0.259	0.177
uned2-4	2.900	1.109	0.412	0.345	0.277	uned2-4	3.600	1.143	0.386	0.278	0.190
uned2-3	3.070	1.048	0.404	0.326	0.262	imdea-1	3.674	1.121	0.360	0.272	0.187
uma-1	2.649	1.214	0.376	0.377	0.304	sinai-2	4.107	1.002	0.356	0.243	0.167
sinai-2	3.212	1.001	0.358	0.311	0.250	sinai-1	4.110	1.001	0.353	0.243	0.167
sinai-1	3.213	1.001	0.356	0.311	0.250	sinai-3	4.113	1.000	0.350	0.243	0.167
sinai-3	3.216	1.000	0.351	0.311	0.250	uma-1	3.338	1.232	0.167	0.300	0.205

### 3.2 Reputation Analysis in RepLab 2012

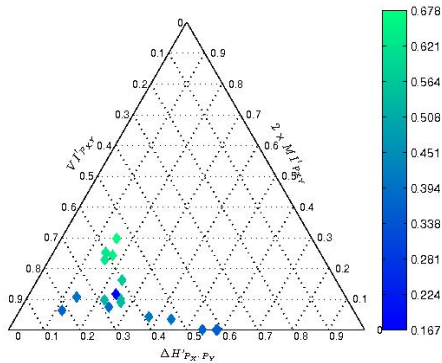
RepLab 2012 was an evaluation campaign aimed at comparing classification systems trained to determine whether a tweet content has positive, negative or neutral implications for corporate reputation [9]. This task is related to sentiment analysis and opinion mining, but differs in some important points: not only opinions or subjective content are being analysed, but also polar facts, i.e. objective information that might have negative or positive implications for a company’s reputation. For instance, “Barclays plans additional job cuts in the next two years” is a fact with negative implications for reputation. Since more than 1 out of 3 tweets are polar facts affecting reputation without containing sentiments or emotions, the number of cases that cannot be correctly captured using sentiment analysis techniques alone is very significant.

Moreover, the focus is set on the decisive role that the point of view or perspective can play since, for example, the same information may be negative from the point of view of the clients and positive from the point of view of investors. For instance, “R.I.P. Michael Jackson. We’ll miss you” has a negative associated sentiment for fans, but a positive implication for the reputation of Michael Jackson.

The data set was manually labelled by experts for 6 companies (for training) and 31 companies (for testing) both in English and Spanish. The distribution of tweets among classes is summarized in Table 3.



(a)



(b)

**Fig. 3. Entropy triangles for the TASS Sentiment Analysis tasks for 3 (a) and 5 (b) polarity degrees. Colormap correlates with accuracy.**

Figure 4 shows the performance of the different systems submitted to the RepLab 2012 evaluation on the Entropy Triangle, whose analysis seems to indicate that classifying reputation polarity is a more complex task than classifying sentiment polarity, since the results in the RepLab 2012 show that most systems present a nearly random behaviour (obtaining very bad performances in the more balanced test distribution). This is further supported on lower accuracies and EMAs.

Only one system (the one above the others) presents results that suggest that, even reporting a low performance, is differentiating correctly between classes. Notoriously, this system is knowledge-supervised, while most of the rest approaches are based in machine learning statistical supervised approaches.

In contrast, the system to the middle of the bottom side of the triangle is specialized returning to every input the label of the majority class. This deduction from the theoretical side was corroborated by its authors declaring that this last system classifies all instances as positive [10], the majority class *in training*. This

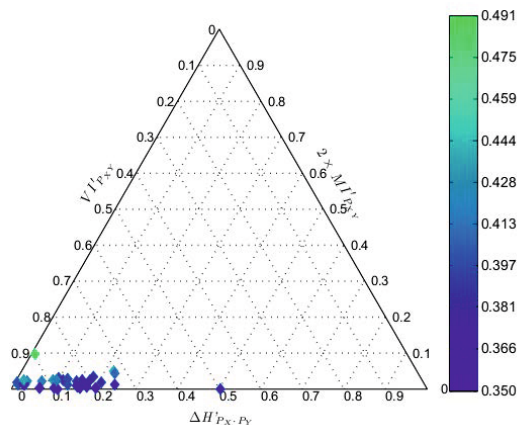


**Table 3. Distribution of tweets per polarity class in the RepLab 2012 corpus.** Effective perplexities are very different for training and testing.

Dataset	P	NEU	N	TOTAL	$k_X$
training	885	550	81	1 516	2.32
testing	1 625	1 488	1 241	4 354	2.98

was a profitable strategy in terms of accuracy according to the training set (see Table 3) but certainly not in the test set where the classes are not that skewed (hence accuracies in the 30%). This extreme behaviour is perfectly identified in the ET and with the NIT factor and it would have been detected irrespective of the test set distribution. In fact, this system is the last in the ranking according to both EMA and NIT whilst holding the 24th position out of 35, according to accuracy. Since many of the systems of the competition were based on statistical modelling, similar behaviours can be observed due to the marked imbalance of the training set classes.

An example of this is the system presented to both evaluations (RepLab 2012 [11] and TASS-SEPLN [12]). This system, based on sentiment analysis techniques [13], achieved a reasonably good performance in TASS3, but was considerably worse in the RepLab 2012. This behaviour seems to corroborate our hypothesis that polarity for reputation and sentiment analysis are substantially different tasks. Finally, it is also worth mentioning that both tasks should take into consideration the presence of irony. Few works have dealt with the effect of irony when analyzing polarity [14, 15], but its correct analysis should increase the performance of SA and RA approaches. Our intuition is that this phenomenon



**Fig. 4. Entropy triangles for the whole population of systems presented to the RepLab2012 Reputation Analysis.** The colormap encodes accuracy. The task is not solved, even as a collective effort, taking the NIT as the criterion.

**Table 4. Relevant perplexities, accuracy  $a(P_{XY})$ , EMA  $a'(P_{XY})$  and NIT factor  $q_X(P_{XY})$  for RepLab 2012 confusion matrices.  $k_X$  is not homogeneous due to the possibility of submitting only part of the results.**

RepLab 2012	$k_X$	$k_{X Y}$	$\mu_{XY}$	$a$	$a'_X$	$q_X$
polarity-Daedalus-1	2.982	2.678	1.113	0.491	0.373	0.371
polarity-HJHL-4	2.775	2.629	1.056	0.439	0.380	0.352
profiling-uned-5	2.982	2.897	1.029	0.436	0.345	0.343
profiling-BMedia-4	2.982	2.899	1.029	0.427	0.345	0.343
profiling-BMedia-5	2.982	2.911	1.024	0.420	0.343	0.341
profiling-uned-2	2.982	2.902	1.027	0.418	0.345	0.342
profiling-uned-4	2.982	2.902	1.027	0.418	0.345	0.342
profiling-BMedia-2	2.982	2.911	1.024	0.415	0.344	0.341
profiling-OPTAH-2.tx	2.981	2.841	1.049	0.408	0.352	0.350
profiling-BMedia-3	2.982	2.924	1.020	0.398	0.342	0.340
profiling-BMedia-1	2.982	2.941	1.014	0.398	0.340	0.338
profiling-OXY-2	2.982	2.938	1.015	0.396	0.340	0.338
profiling-uned-1	2.982	2.892	1.031	0.396	0.346	0.344
profiling-uned-3	2.982	2.892	1.031	0.396	0.346	0.344
profiling-OXY-1	2.982	2.939	1.015	0.394	0.340	0.338
polarity-HJHL-1	2.775	2.685	1.034	0.391	0.372	0.345
profiling-ilps-4	2.982	2.962	1.007	0.391	0.338	0.336
profiling-ilps-3	2.982	2.914	1.023	0.385	0.343	0.341
profiling-ilps-1	2.982	2.962	1.007	0.384	0.338	0.336
profiling-kthgavagai	2.982	2.922	1.020	0.383	0.342	0.340
profiling-ilps-5	2.982	2.876	1.037	0.382	0.348	0.346
profiling-OPTAH-1.tx	2.981	2.904	1.026	0.380	0.344	0.342
polarity-HJHL-3	2.775	2.695	1.030	0.377	0.371	0.343
profiling-GATE-1	2.982	2.982	1.000	0.373	0.335	0.333
profiling-OXY-4	2.982	2.947	1.012	0.369	0.339	0.337
profiling-ilps-2	2.982	2.960	1.008	0.369	0.338	0.336
polarity-HJHL-2	2.775	2.697	1.029	0.369	0.371	0.343
profiling-uiowa-2	2.982	2.937	1.015	0.367	0.340	0.338
profiling-uiowa-5	2.982	2.940	1.014	0.367	0.340	0.338
profiling-OXY-5	2.982	2.967	1.005	0.365	0.337	0.335
profiling-uiowa-1	2.980	2.933	1.016	0.362	0.341	0.339
profiling-uiowa-4	2.982	2.974	1.003	0.360	0.336	0.334
profiling-GATE-2	2.982	2.971	1.004	0.357	0.337	0.335
profiling-uiowa-3	2.980	2.975	1.001	0.355	0.336	0.334
profiling-OXY-3	2.982	2.967	1.005	0.350	0.337	0.335

is more common in RA texts and can explain, to some extent, the remarkable differences in the results.

Table 4 shows the numeric results of the various metrics being compared. The interesting note here is that another system would actually have won the competition if the metric was EMA, specifically “polarity-HJHL-4”. This is one of set of systems marked in green whose EMA is comparable to that which won the competition.

## 4 Conclusions: A Proposal

We have motivated and proposed a combination of two tools as an alternative or a complement to standard accuracy-based metrics for Sentiment Analytics tasks, testing them on two different evaluation runs of Sentiment Analysis (TASS-SEPLN) and Reputation Analysis (RepLab 2012).

On the one hand, EMA is a better motivated, although pessimistic, estimate of accuracy that takes into consideration the dataset being considered and how much a particular system has learnt in the training process. This is to be used for ranking purposes.

On the other hand, the NIT factor is a measure of how efficient the training process of the classifier was, that can be visualized directly with the help of the Entropy Triangle. This is intended as a mechanism for technology development under the heuristic of maximizing the information transmitted in the learning process. It is well-matched to EMA in the sense that maximizing the former maximizes the latter.

We have shown that using both in combination in postmortem system analysis detects incongruencies and shortcomings of rankings based in accuracy.

As future lines of work a more in depth analysis of the learning process can be pursued by interpreting the split entropy diagram of Fig. 1.

The MATLAB<sup>1</sup> code to draw the entropy triangles in Figs. 3 and 4 has been made available at: <http://www.mathworks.com/matlabcentral/fileexchange/30914>

**Acknowledgments.** We would like to thank the organizers of the TASS-SEPLN and RepLab12 evaluations for providing us with the evaluation data.

## References

- [1] Zhu, X., Davidson, I.: Knowledge discovery and data mining: challenges and realities. Premier reference source. Information Science Reference (2007)
- [2] Thomas, C., Balakrishnan, N.: Improvement in minority attack detection with skewness in network traffic. In: Proc. of SPIE, vol. 6973, pp. 69730N–69730N–12 (2008)
- [3] Fernandes, J.A., Irigoien, X., Goikoetxea, N., Lozano, J.A., Inza, I., Pérez, A., Bode, A.: Fish recruitment prediction, using robust supervised classification methods. *Ecological Modelling* 221, 338–352 (2010)
- [4] Valverde-Albacete, F.J., Peláez-Moreno, C.: Two information-theoretic tools to assess the performance of multi-class classifiers. *Pattern Recognition Letters* 31, 1665–1671 (2010)
- [5] Meila, M.: Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* 28, 875–893 (2007)
- [6] Valverde-Albacete, F.J., Peláez-Moreno, C.: 100% classification accuracy considered harmful: The Normalized Information Transfer explains the accuracy paradox (submitted, 2013)

---

<sup>1</sup> A registered trademark of The MathWorks, Inc.

- [7] Mejía-Navarrete, D., Gallardo-Antolín, A., Peláez-Moreno, C., Valverde-Albacete, F.J.: Feature extraction assessment for an acoustic-event classification task using the entropy triangle. In: *Interspeech 2010* (2011)
- [8] Villena-Román, J., García-Morera, J., Moreno-García, C., Ferrer-Ureña, L., Lana-Serrano, S.: TASS - Workshop on sentiment analysis at SEPLN (2012)
- [9] Amigó, E., Corujo, A., Gonzalo, J., Meij, E., Rijke, M.: Overview of RepLab 2012: Evaluating online management systems. In: *CLEF* (2012)
- [10] Greenwood, M.A., Aswani, N., Bontcheva, K.: Reputation profiling with gate. In: *CLEF* (2012)
- [11] Carrillo-de-Albornoz, J., Chugur, I., Amigó, E.: Using an emotion-based model and sentiment analysis techniques to classify polarity for reputation. In: *CLEF* (2012)
- [12] Martín-Wanton, T., Carrillo-de-Albornoz, J.: UNED at TASS 2012: Polarity classification and trending topic system. In: *Workshop on Sentiment Analysis at SEPLN* (2012)
- [13] Carrillo-de-Albornoz, J., Plaza, L., Gervás, P.: A hybrid approach to emotional sentence polarity and intensity classification. In: *Conference on Computational Natural Language Learning, CoNLL 2010*, pp. 153–161 (2010)
- [14] Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation* 47, 239–268 (2013)
- [15] Reyes, A., Rosso, P.: On the difficulty of automatically detecting irony: beyond a simple case of negation. In: *Knowledge and Information Systems*, 1–20 (2013)

# A New Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection

Imene Bensalem<sup>1</sup>, Paolo Rosso<sup>2</sup>, and Salim Chikhi<sup>1</sup>

<sup>1</sup> MISC Lab., Constantine 2 University, Algeria  
bens.imene@gmail.com, chikhi@misc-umc.org

<sup>2</sup> Natural Language Engineering Lab. – EliRF, Universitat Politècnica de València, Spain  
prossso@dsic.upv.es

**Abstract.** The present paper introduces the first corpus for the evaluation of Arabic intrinsic plagiarism detection. The corpus consists of 1024 artificial suspicious documents in which 2833 plagiarism cases have been inserted automatically from source documents.

**Keywords:** Arabic intrinsic plagiarism detection, evaluation corpus, automatic plagiarism generation.

## 1 Introduction

“Plagiarism occurs when someone presents the work of others (data, text, or theories) as if they were his/her own and without proper acknowledgment” [1]. One may uncover plagiarism in a text document by observing similarities between it and other documents (external plagiarism detection), or by noticing a sort of heterogeneity in the writing style (intrinsic plagiarism detection) [2]. Automatic methods of plagiarism detection are inspired by these two traditional approaches. In the external approach, it is necessary to hold a collection of documents representing the source of plagiarism; whereas, in the intrinsic approach, there is no need for source documents. Indeed, the importance of this approach emerges when the plagiarism source is unknown or does not have a digital version. In this paper, we are interested in the intrinsic plagiarism detection in Arabic texts. Concretely, we will describe the first corpus for the evaluation of Arabic intrinsic plagiarism detection. The remainder of the paper is structured as follows: Sections 2 and 3 provide a brief overview of the intrinsic plagiarism detection in English and Arabic languages respectively. In this overview we focus on the evaluation aspect. Section 4 presents the methodology adopted in the construction of our corpus and provides statistics on it. Finally, Section 5 concludes the paper.

## 2 Intrinsic Plagiarism Detection in English Text

In the last years, a great effort has been made to standardize the evaluation of the automatic plagiarism detection with its external and intrinsic approaches. As a result, an evaluation framework has been developed. It consists in a set of quality measures

and a series of evaluation corpora involving automatically created suspicious documents [3]. This evaluation framework was used in the plagiarism detection task of PAN competition<sup>1</sup> from 2009 to 2011 [2] [4] [5]. The part of PAN 2011 corpus, used to evaluate the intrinsic approach, contains 4753 suspicious documents with 11443 plagiarism cases.

In PAN 2012, another evaluation framework has been introduced [6]. Unlike the previous corpora, all the suspicious documents of PAN 2012 corpus were created manually through crowdsourcing. This new corpus was used to evaluate only the external approach, while the intrinsic one has been considered as an authorship clustering problem and therefore, has been evaluated within PAN authorship attribution task using another evaluation corpus, which is very small in comparison with the former (less than 10 documents) [7].

### 3 Intrinsic Plagiarism Detection in Arabic Text

Although the broad spread of plagiarism in the Arab world [8], plagiarism detection in the Arabic text is still in its infancy, especially when it concerns the intrinsic approach. We think that the main reason behind this fact is the lack of an evaluation corpus. Moreover, there are very few works on Arabic authorship analysis [9–11] which is one of the most related disciplines to intrinsic plagiarism detection. To the best of our knowledge, the only work in this area is ours [12] where we used a toy corpus composed of 10 documents with 63 plagiarism cases.

With regard to the external approach, some detection methods were proposed in the last few years. Nonetheless, it is difficult to draw a clear conclusion on the performance of these methods since they were evaluated, using different strategies and corpora. Jadalla and Elnagar [13] compared their web-based system with a baseline method using a number of documents that have been presumed to be suspicious. Alzahrany and Salim [14] as well as Menai [15] evaluated their methods using respectively 15 and 300 suspicious documents constructed by rewording and restructuring sentences. Jaoua et al. [16] created 76 suspicious documents by the manual insertion of text fragments obtained by queries to search engine, using keywords in relation with the subject of the document that will host the plagiarism.

The next section describes the building of the first Arabic corpus for intrinsic plagiarism detection evaluation. We think that the creation of such a corpus will encourage researchers to investigate this unexplored area.

### 4 Methodology

A corpus of plagiarism detection evaluation should be composed of two collections of documents: suspicious documents and source documents. A suspicious document contains fragments of texts plagiarized from one or more source documents. These latter are omitted from the corpus if the evaluation concerns the intrinsic approach.

---

<sup>1</sup> <http://pan.webis.de/>

Due to the difficulty (for ethical and feasibility reasons) of owning a document collection containing actual plagiarism cases, suspicious documents have to be built. Two approaches have been used in the state-of-the-art researches: manual and automatic. The manual approach [17] is the more realistic in terms of simulating the real plagiarist behaviour. It consists in charging people to write essays on designated topics with allowing the text reuse from different references. However, the automatic approach [3] follows two steps: (1) Compilation of target and source documents. Documents of both collections must be tagged with their author names and topics to prepare them for the second step; (2) Insertion of plagiarism: this task tries to simulate the act of plagiarism by borrowing automatically text segments from source documents and inserting them randomly in a target document. The target document and their sources of plagiarism must have the same topic but different authors.

Although the automatic approach is less realistic and suffers from many shortcomings [6], we adopted it to build our corpus for two main reasons. First, the automatic approach is acceptable since it has been used to build PAN 2009-2011 corpora. Second, the manual approach is costly in terms of human and material resources [17]. The following subsections provide details on the steps of our corpus construction which are text compilation and plagiarism insertion.

#### 4.1 Text Compilation

**Criteria of Texts:** We set a number of criteria that should be verified in the target documents (documents where plagiarised fragments will be inserted).

*C1.* Each target document must be written by one author only. Otherwise, the document will contain many writing styles which may complicate the intrinsic plagiarism detection even further.

*C2.* Target documents should not include much of text reuse or many quotations. In fact, this is a feature of Arabic religion books which include many quotations from Holy texts. The purpose of this criterion is to avoid altering the evaluation by texts that are likely to be detected as plagiarism cases, although they are actually legitimate cases of text reuse.

*C3.* Target documents should not be too short. Indeed, we presume that the stylistic analysis becomes unreliable with short Arabic texts as it is with short English text (less than half a page approx.) [18].

*C4.* Texts should be punctuated because they will undergo a style analysis where the punctuation is an important feature. This criterion seemed obvious, but we decided to mention it because in a late stage of the text compilation, we noticed a lack of quality of some Arabic online texts. Effectively, we discarded many of the collected documents because they were poorly edited in terms of punctuation as well as section separations (no new line character between sections)<sup>2</sup>.

---

<sup>2</sup> It is particularly the case of old books which represent an important part of the copyright-free text available online.

**Source of Text.** Since we plan to make the corpus publicly available, it was primordial to gather texts from a copyright-free source. For this reason along with the specific desired criteria, sources of text have become very limited. We finally decided to build our corpus from Arabic Wikisource which is a library of heritage books and public domain texts. Furthermore, most of its documents are tagged with topics and author names (see our paper [19] for further details on the text compilation from Wikisource). We also added some texts from other sources, after making sure that they are without copyright. Table 1 presents the sources of our document collection.

**Table 1.** Our corpus sources of text

Source of text	Percentage of documents in the corpus
Arabic Wikisource <sup>3</sup>	98%
Create your own country blog <sup>4</sup> KSUCCA corpus <sup>5</sup> Islamic book web site <sup>6</sup>	2%

## 4.2 Insertion of Plagiarism

Inspired by the PAN 2009-2011 corpora methodology, the suspicious documents were created automatically according to two parameters: the percentage of plagiarism per document and the length of plagiarism fragments. The main steps of the plagiarism insertion are:

1. Indexing source documents as fragments of different lengths to be used as plagiarism cases.
2. Selection of plagiarism sources for each target document according to its topic and its author name.
3. Random selection of segments from the source documents indices and their insertion in a random position in the target document.
4. Annotation of the plagiarism cases in an XML document following PAN corpora scheme.

To generate the suspicious documents with a variety of the plagiarism percentage and the case lengths, we split the target documents into 6 sets according to the document lengths. Each set was divided arbitrary into 4 equal subsets. Finally, plagiarism was inserted in each subset with a fixed percentage limit and a list of plagiarism case lengths. Statistics on the obtained corpus are provided in Table 2.

<sup>3</sup> <http://ar.wikisource.org>

<sup>4</sup> <http://diycountry.blogspot.com>

<sup>5</sup> Al-Rabiah, M.: King Saud University Corpus of Classical Arabic (KSUCCA), <http://ksucorpus.ksu.edu.sa> (2012).

<sup>6</sup> <http://www.islamicbook.ws>



**Table 2.** Statistics on the Arabic intrinsic plagiarism detection corpus

<b>Document statistics</b>			
<b>Total number of documents</b>		1024	
<b>Plagiarism percentage per document</b>		<b>Document length</b>	
Null (0%)	20%	Very Short (1-3 pages)	46%
Hardly ]0% 10%]	24%	Short (3-15 pages)	37%
Few ]10% 30%]	32%	Medium (15-100 pages)	12%
Medium ]30% 60%]	24%	Long (>100 pages)	05%
<b>Plagiarism cases statistics</b>			
<b>Total number of plagiarism cases</b>		2833	
<b>Plagiarism cases length</b>		<b>Number of plagiarism cases per document</b>	
Very short (some sentences)	09%	Null (0)	20%
Short (some paragraphs)	40%	Few ]0 5]	69%
Medium (around 1 page)	21%	Medium ]5 15]	08%
Long (many pages)	30%	Much ]15 45]	03%

## 5 Conclusion

In this paper we described the first evaluation corpus for Arabic intrinsic plagiarism detection. The corpus was built automatically and it follows standards in the annotation of plagiarism cases. The main difficulty we encountered during the construction of this corpus is the lack of good quality copyright-free Arabic text. This fact has limited the text sources of our corpus. We believe that the release of such a free corpus will foster research in intrinsic plagiarism detection in Arabic.

**Acknowledgements.** This work is the result of the collaboration in the framework of the bilateral research project AECID-PCI AP/043848/11 (Application of Natural Language Processing to the Need of the University) between the Universitat Politècnica de València in Spain and Constantine 2 University in Algeria.

## References

1. Springer Policy on Publishing Integrity. Guidelines for Journal Editors
2. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009), pp. 1–9 (2009)
3. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An Evaluation Framework for Plagiarism Detection. In: Huang, C.-R., Jurafsky, D. (eds.) Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), pp. 997–1005. ACL (2010)
4. Potthast, M., Barrón-cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In: Braschler, M., Harman, D. (eds.) Notebook Papers of CLEF 2010 LABs and Workshops (2010)

5. Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd International Competition on Plagiarism Detection. In: Petras, V., Forner, P., Clough, P. (eds.) Notebook Papers of CLEF 2011 LABs and Workshops (2011)
6. Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., Stein, B.: Overview of the 4th International Competition on Plagiarism Detection. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop –Working Notes Papers (2012)
7. Juola, P.: An Overview of the Traditional Authorship Attribution Subtask Notebook for PAN at CLEF 2012. In: Forner, P., Karlgren, J., and Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop –Working Notes Papers (2012)
8. Yakout, M.M.: Examples of Plagiarism in Scientific and Cultural Communities (in Arabic), [http://www.yaqout.net/ba7s\\_4.html](http://www.yaqout.net/ba7s_4.html)
9. Abbasi, A., Chen, H.: Applying Authorship Analysis to Arabic Web Content. In: Kantor, P., Muresan, G., Roberts, F., Zeng, D.D., Wang, F.-Y., Chen, H., Merkle, R.C. (eds.) ISI 2005. LNCS, vol. 3495, pp. 183–197. Springer, Heidelberg (2005)
10. Shaker, K., Corne, D.: Authorship Attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis. In: 2010 UK Workshop on Computational Intelligence (UKCI), pp. 1–6. IEEE (2010)
11. Ouamour, S., Sayoud, H.: Authorship attribution of ancient texts written by ten arabic travelers using a SMO-SVM classifier. In: 2012 International Conference on Communications and Information Technology (ICCIT), pp. 44–47. IEEE (2012)
12. Bensalem, I., Rosso, P., Chikhi, S.: Intrinsic Plagiarism Detection in Arabic Text: Preliminary Experiments. In: Berlanga, R., Rosso, P. (eds.) 2nd Spanish Conference on Information Retrieval (CERI 2012), Valencia (2012)
13. Jadalla, A., Elnagar, A.: A Plagiarism Detection System for Arabic Text-Based Documents. In: Chau, M., Wang, G.A., Yue, W.T., Chen, H. (eds.) PAISI 2012. LNCS, vol. 7299, pp. 145–153. Springer, Heidelberg (2012)
14. Alzahrani, S., Salim, N.: Statement-Based Fuzzy-Set Information Retrieval versus Fingerprints Matching for Plagiarism Detection in Arabic Documents. In: 5th Postgraduate Annual Research Seminar (PARS 2009), Johor Bahru, Malaysia, pp. 267–268 (2009)
15. Menai, M.E.B.: Detection of Plagiarism in Arabic Documents. *International Journal of Information Technology and Computer Science* 10, 80–89 (2012)
16. Jaoua, M., Jaoua, F.K., Hadrich Belguith, L., Ben Hamadou, A.: Automatic Detection of Plagiarism in Arabic Documents Based on Lexical Chains. *Arab Computer Society Journal* 4, 1–11 (2011) (in Arabic)
17. Potthast, M., Hagen, M., Völske, M., Stein, B.: Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: 51st Annual Meeting of the Association of Computational Linguistics (ACL 2013). ACM (to appear, 2013)
18. Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic plagiarism analysis. *Language Resources and Evaluation* 45, 63–82 (2010)
19. Bensalem, I., Rosso, P., Chikhi, S.: Building Arabic Corpora from Wikisource. In: 10th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2013). IEEE (2013)

# Selecting Success Criteria: Experiences with an Academic Library Catalogue

Paul Clough and Paula Goodale

Information School  
University of Sheffield  
Sheffield, UK  
{p.d.clough,p.goodale}@sheffield.ac.uk

**Abstract.** Multiple methods exist for evaluating search systems, ranging from more user-oriented approaches to those more focused on evaluating system performance. When preparing an evaluation, key questions include: (i) why conduct the evaluation, (ii) what should be evaluated, and (iii) how the evaluation should be conducted. Over recent years there has been more focus on the end users of search systems and understanding what they view as ‘success’. In this paper we consider *what* to evaluate; in particular what criteria users of search systems consider most important and whether this varies by user characteristic. Using our experience with evaluating an academic library catalogue, input was gathered from end users relating to the perceived importance of different evaluation criteria prior to conducting an evaluation. We analyse results to show which criteria users most value, together with the inter-relationships between them. Our results highlight the necessity of conducting multiple forms of evaluation to ensure that search systems are deemed successful by their users.

**Keywords:** Evaluation, success criteria, digital libraries.

## 1 Introduction

Evaluation is highly important for designing, developing and maintaining effective search systems, as it allows the measurement of how successfully the system meets its goal of helping users fulfill their information needs or complete their tasks [1-3]. Evaluation involves identifying suitable success criteria that can be measured in some way. Success might refer to whether a search system retrieves relevant (compared with non-relevant) documents; how quickly results are returned; how well the system supports users’ interactions; whether users are satisfied with the results; how easily users can use the system; whether the system helps users carry out their tasks and fulfill their information needs; whether the system impacts on the wider environment; or, how reliable the system is.

How to conduct IR system evaluation has been an active area of research for the past 50 years and the subject of much discussion and debate [1, 2]. Traditionally in IR there has been a strong focus on measuring system effectiveness: the ability of an IR system to discriminate between documents that are relevant or not relevant for a given

user query. This focus on the system has, in part, been influenced by the focus of the IR community on the development of retrieval algorithms, together with the organization of large IR evaluation events, such as TREC and CLEF. Such events have focused on measuring system effectiveness in a controlled experimental setting [4, 5]. However, the scope of ‘system’ in IR has slowly broadened to include more elements of the retrieval context, such as the user or the user’s environment, which must be included in the evaluation of IR systems [3, 6]. Therefore, instead of focusing on just the system (i.e., its inputs and outputs), a more user-oriented approach can be taken. This may take into account the user, the user’s context and situation, and their interactions with an IR system, perhaps in a real-life operational environment [7, 8].

When planning an evaluation, there are at least three key questions to address [10]: (i) why evaluate; (ii) what to evaluate; and (iii) how to evaluate. These apply regardless of the type of search system being evaluated (e.g. search engine or library catalogue). Saracevic [10] also mentions these when planning an evaluation of a digital library, together with identifying for whom to evaluate (i.e., the stakeholder). In this paper we focus on the issue of *what*; selecting criteria to assess search success. In practice this is a key question as it is often not possible, nor desirable, to run every type of evaluation available, and it is therefore necessary to be selective, both in order to measure the success of the new system, and to make best use of time and scarce resources. Although there is much literature exploring evaluation criteria, factors and measures, there is less on identifying which criteria are the most important. Through our experiences with planning the evaluation of an operational search system (online library catalogue) we investigate the importance of different evaluation criteria as perceived by existing end users. This was a necessary pre-cursor of selecting the evaluation methods and criteria to determine the system’s success. We specifically address the following research questions: (RQ1) What evaluation criteria are viewed as important by end users?; (RQ2) What degree of variation exists between users’ preferences?; and, (RQ3) What inter-relationships exist between different criteria?

The remainder of the paper is structured as follows. Section 2 describes related work on evaluation, particularly in the context of digital libraries where much recent discussion has taken place. Section 3 outlines the methodology used to gather end users’ feedback on the importance of different evaluation criteria. Section 4 analyses results based on various statistical analyses. Section 5 provides a discussion and Section 6 concludes the paper and provides avenues for further work.

## 2 Related Work

Saracevic [10] argues for evaluation of digital libraries to take place at different levels, each with different measurement criteria derived from research objectives. In a similar vein, the Interaction Triptych Framework (ITF) defines three main type of digital library evaluation criteria [11, 12]: *performance* (system performance); *usability* (user interaction and satisfaction measures); and, *usefulness* (support for task and information needs), which should be deployed to measure success in the digital

library context. Tsakonias & Papatheodorou, [12] test the relative preferences of users for evaluation measures in the three ITF categories, finding high scores for usability and usefulness measures, but lower scores for performance measures, as well as correlation between usability and usefulness measures. The usability and usefulness dimensions of evaluation are explored further and again found to be inter-related, with several key attributes of each dimension identified as a basis for an integrated approach to measurement [13].

A holistic and multi-dimensional approach is widely advocated for digital library evaluation, yet ranking of diverse evaluation criteria by degrees of importance is less evident, particularly in advance of designing evaluation protocols. Nielsen [14] rates factors within a range of usability heuristics, concluding they are all very similar in importance, but does not consider other types of evaluation metrics. Toms et al. [15] look at determining the different dimensions of relevance, and Al-Maskari & Sanderson [16] focus on elements of user satisfaction. These are however all studies of individual dimensions of evaluation. Xie [18] considers a broader range of evaluation measures and ranks criteria within several high-level evaluation categories, but does not rank across categories. Conversely, Buchanan & Salako [13] rank a wide range of evaluation criteria across categories. The rankings or relative importance by Xie [17] and Buchanan & Salako [13] are undertaken during (rather than prior to) an active evaluation study, once evaluation experiments have been completed, and can therefore only contribute to the design of future studies of the same systems.

User preferences for evaluation measures have also been largely overlooked. Xie [18] addresses this issue, with a study eliciting evaluation criteria from users, finding that measures suggested by users match those proposed in evaluation models and used in empirical studies. However, it is acknowledged that the user sample for this study is too homogeneous and that more diverse groups may influence the results. Kelly [19] criticizes the use of inadequate user models, citing the over-use of the librarian or search intermediary, and other single types of users, when more detailed and varied user models would be more appropriate, especially for more complex systems.

Differences between users have been studied in many ways, including search success, information needs, user requirements, information seeking behavior, cognitive style, and relevance judgments, amongst others, yet it is much less common for user differences to be taken into account in the selection of appropriate evaluation criteria. One common user classification in digital library and search user studies is the novice/expert dichotomy. In this light, Marchionini [20] defines three aspects of user difference relating to information skills; domain expertise (subject knowledge), search expertise, and system expertise. Similarly, Hölscher & Strube [21] define novices and experts with attributions of domain and search experience, considering the impact of these characteristics on search success. These are then the main characteristics that we consider as differentiators in the current study, with the addition of a three-way classification of user role in line with recommendations by Kelly [19] for consideration of a range of user models. In summary, investigating what criteria users perceive as indicators of search success has been largely under-explored in past research.

### 3 Methodology

In the current study, the search system to be evaluated is a virtual union OPAC (on-line public access catalogue). The UK-based Search25<sup>1</sup> project aimed to develop a significantly updated version of the longstanding and somewhat outdated InforM25 system, retrieving records from the individual OPACs of the 58 member institutions of the M25 Consortium of academic libraries. End users vary from undergraduate and postgraduate students, to academics and research staff, and library professionals, with varying degrees of subject, domain and search expertise, information needs and patterns of search behavior. During the development of Search25, a formative evaluation was planned to assess the success of an existing system known as InforM25.

**Table 1.** How important are the following factors when using academic library catalogues?

#	Statement (criteria)	Criteria group*
1	The catalogue/system is easy to use	Usability
2	The interface is attractive	Usability
3	Results are retrieved quickly	Performance
4	All items on a topic are retrieved	Performance
5	The most relevant items of a topic are identified	Performance
6	Information in the catalogue is easy to understand	Usefulness
7	It is enjoyable to use	Usability
8	Help is available when needed	Usability

\* Criteria group based on the relationships identified in the Interaction Triptych Framework [11, 12]

An online survey of the current InforM25 users was carried out to identify suitable criteria for evaluating the updated system once fully implemented (i.e., to define *what* to evaluate). Specific questions were presented as eight statements (Table 1) covering a range of evaluation criteria drawn from various literature, against which a 5-point Likert scale was used to rate users' level of agreement, from Strongly Disagree (1) to Strongly Agree (5). Criteria were grouped based on the Interaction Triptych Framework [11, 12]. In addition, we asked open-ended questions to gather qualitative responses about what users liked and disliked about the InforM25 system and what they would like to change, in order to corroborate and enrich the quantitative data.

An email invitation to complete an online survey was distributed to mailing lists of the 58 academic institutions known to use InforM25, and a link was placed on the home page. In total, 196 respondents provided responses about their perceived importance of different evaluation criteria. The survey sample comprises library, learning and technical support staff (79%), academic and research staff (5%), and undergraduate and postgraduate students (16%) from across the M25 Consortium 58 institutions. This skewed distribution is representative of users of the system at the time of the study, and one reason for updating the system was to try to broaden the appeal to academic and student users. Respondents came from a broad range of subject domains

<sup>1</sup> <http://www.search25.ac.uk/pazpar2/inform25>

with varying levels of search experience, many reporting more than 10 years experience in using academic library catalogues (65%) and web search engines (71%). In terms of frequency of use, 87% of participants used their own institution's OPAC on at least a weekly basis, whilst only 22% used InforM25 with this regularity.

Data collected through the survey was analysed using descriptive and inferential statistics. One of the main goals of this analysis was to determine if differences between user groups regarding the importance of evaluation criteria could be observed. Relationships between evaluation criteria were assessed using Principal Component Analysis (PCA) and bivariate correlations assessed using Kendall's Tau. Finally, a thematic analysis of qualitative data from the associated open-ended questions and focus groups is used to expand upon and corroborate the quantitative findings.

## 4 Results

Responses relating to the evaluation criteria shown in Table 1 were analyzed first for the whole sample (Section 4.1), and then for differences between users based upon a range of characteristics (Section 4.2). We then explore interrelationships between the criteria in Section 4.3, and associated qualitative results in Section 4.4.

**Table 2.** Frequency of responses for importance of evaluation criteria, all users ( $N=196$ )

Criteria	1 Strong. Disagree	2 Disagree	3 Neutral	4 Agree	5 Strong. Agree	Median score	SD
Easy To Use	1.0%	0.0%	3.1%	17.4%	<b>78.5%</b>	5	0.622
Attractive	3.6%	6.3%	31.3%	<b>35.9%</b>	22.9%	4	1.012
Quick	0.5%	1.5%	5.6%	27.2%	<b>65.1%</b>	5	0.719
Retrieve All	0.5%	3.2%	8.4%	33.7%	<b>54.2%</b>	5	0.812
Relevant	2.6%	3.7%	14.2%	31.6%	<b>47.9%</b>	5	0.988
Understandable	0.5%	0.0%	5.8%	24.7%	<b>68.9%</b>	5	0.646
Enjoyable	10.0%	10.0%	<b>35.8%</b>	27.4%	16.8%	3	1.165
Help	3.2%	11.1%	27.5%	<b>31.7%</b>	26.5%	4	1.081

### 4.1 Analysis of All Users

Table 2 shows results for percentage of responses across all users for each of the evaluation criteria. The results suggest overall importance for some aspects of usability and system performance measures. The most highly rated measures are ease of use (96% rating this at 4 or 5 on the Likert scale); information in the catalogue is easy to understand (94%); results are retrieved quickly (93%); all items on a topic are retrieved (88%); and, the most relevant items are identified (80%). Responses for interface is attractive (59%); help is available (58%); and, enjoyable to use (44%) indicate that overall, these criteria are somewhat less important to users.

## 4.2 Analysis by User Characteristic

To examine the effects of user characteristic (role, subject area, search experience and frequency of use of existing finding aids) on the ratings provided, we divide the data into groups and compare the ratings for each group using a Kruskal-Wallis test (due to the non-parametric nature of the data). When analysing results of the test we test the null hypothesis that the user characteristic makes no difference to the perceived importance of evaluation criteria, rejected at  $p < 0.05$ .

**Table 3.** Kruskal-Wallis test statistics, importance of evaluation criteria grouped by user characteristics ( $N=196$ ). Bold indicates results with  $p < 0.05$

		1	2	3	4	5	6	7	8
User type	K-W (2 df)	3.686	<b>14.046</b>	<b>6.825</b>	1.775	3.299	5.802	2.820	2.364
	p-value	0.158	<b>0.001</b>	<b>0.033</b>	0.412	0.192	0.055	0.244	0.307
Subject	K-W (6 df)	7.881	4.284	9.305	11.663	6.468	9.156	6.502	5.054
	p-value	0.247	0.638	0.157	0.070	0.373	0.165	0.369	0.537
Web search experience	K-W (3 df)	4.401	1.018	7.363	1.406	.952	<b>7.914</b>	0.493	0.574
	p-value	0.221	0.797	0.061	0.704	0.813	<b>0.048</b>	0.920	0.902
Experience of library work	K-W (3 df)	1.249	6.923	<b>11.843</b>	<b>9.692</b>	1.296	<b>10.403</b>	2.358	1.507
	p-value	0.741	0.074	<b>0.008</b>	<b>0.021</b>	0.730	<b>0.015</b>	0.502	0.681
Frequency of using InforM25	K-W (3 df)	11.826	7.760	<b>12.962</b>	5.080	9.649	3.917	10.508	11.814
	p-value	0.066	0.256	<b>0.044</b>	0.534	0.140	0.688	0.105	0.066

Firstly, we find that the null hypothesis is rejected most often for criteria 3) results are retrieved quickly, indicating that the importance of this varies most by user characteristic. Secondly, we observe that experience of library work has an effect on the most criteria (2, 3 and 6), suggesting that this user characteristic causes users to disagree the most about search success. Next we consider each user characteristic in turn:

**User Type.** For the user type (e.g. librarian vs. student) the null hypothesis is rejected for attractiveness of interface and results are retrieved quickly. It can be concluded that there is a difference in perceived importance of these evaluation criteria between the various user roles. Splitting results by different academic subject areas there are no statistically significant differences between the criteria, suggesting that domain does not affect the importance of certain evaluation criteria.

**Search Experience.** This was measured by two characteristics: number of years experience in using an academic library catalogue and experience of using web search engines. The Kruskal-Wallis test showed no significant differences against the level of experience in using library catalogues; however, for experience of web search the null hypothesis is rejected ( $p < 0.05$ ) for criteria 6 (information is easy to understand). Inspection of the medians reveal that those users with only 0-1 years experience have



a lower median of 4, with a median of 5 for all other levels of experience, and the most demanding users (least spread of results) are those with 2-5 years experience. Given the high proportion of library staff in the sample, experience was also considered by the number of years engaged in library work, and for students (the least experienced users), their year of study. No significant differences for student users by year of study were found, but for library staff significance was found for criteria 3 (results are retrieved quickly) at  $p < 0.01$ , for criteria 4 (all items are retrieved) and 6 (information is easy to understand) at  $p < 0.05$ .

**Frequency of Using Existing Finding Aids.** Finally, we analysed ratings based on grouping users by their frequency of use of a variety of library IR systems, including the OPAC at the user's own institution, and the InforM25 virtual union catalogue. For home library OPACs, no significant differences were found against the frequency of use. However, for the frequency of use on InforM25, the null hypothesis is rejected for criteria 3 (items are retrieved quickly) at  $p < 0.05$ . Analysis of the medians reveals that for criteria 1, daily users have a median of 4.5, whilst all other users have a median of 5. For criteria 3, the differences are inconclusive, as users of all frequencies have a median of 4, except those with monthly use (median=3.5). Lastly, the medians for criteria 8 show that the availability of help is more important with less frequent users, with a median of 4 for users with a frequency or monthly, less often or never, and median of 3-3.5 for more frequent users.

### 4.3 Relationships between Evaluation Criteria

To examine relationships between the evaluation criteria in each group, a factor analysis was conducted using Principal Component Analysis (PCA) with varimax rotation (assuming independence between groupings). An initial PCA based upon Eigenvalues  $> 1$  extracted two dimensions, accounting for 56% of total variance. However, the scree plot shows that three or four dimensions might be more appropriate, and therefore revised PCAs specifying three and four fixed factors were undertaken, accounting for 70% and 79% of variance, respectively. Bartlett's Test of Sphericity is found to be highly significant at  $p < 0.001$  and the Kaiser-Olkin measure of sampling adequacy is sufficient at 0.792, indicating that the evaluation criteria are likely to be related and that the sample is of adequate size.

Rotated component matrices show the loading of the components for each of the evaluation criteria (Table 4). With three components extracted, four criteria (easy to use, quick, understandable, relevant) load high on the first component; three criteria (enjoyable, attractive, help) load high on the second component; and, two criteria (help, retrieve all) load high on the third component. With four components extracted the results are similar, with the exception that one criterion (relevant) now loads highly on the fourth component, instead of the first component, and help now only loads highly on the second component, rather than on the second and third components.

It is interesting to note that one evaluation criteria from the usability group in Table 1 (easy to use), and one from the usefulness group (understandable) are close to performance variables, in particular speed of retrieval (quick) and relevant information is returned by the system (relevant). It is perhaps to be expected that relevant and

retrieval all are at opposite ends of this cluster (with four components extracted they occupy their own space) as their importance is likely to vary by task. Two usability variables (enjoyable, attractive) are also clustered together, with the addition of a further usability variable (help). The groupings found using the PCA are further confirmed by inspecting significant ( $p < 0.01$ ) bivariate correlations between criteria with a high ( $> 0.4$ ) Kendall's Tau score: 'easy to use' and 'quick' ( $\tau = 0.502$ ), 'ease of use' and 'understandable' ( $\tau = 0.529$ ); 'enjoyable' and 'attractive' ( $\tau = 0.540$ ), 'enjoyable' and 'help' ( $\tau = 0.402$ ).

**Table 4.** Rotated Component Matrices (varimax), 3 and 4 Components extracted

	3 Components				4 Components			
	1	2	3		1	2	3	4
Easy to Use	<b>0.813</b>		0.224	Quick	<b>0.86</b>	0.159		
Quick	<b>0.81</b>	0.129	0.162	Easy to Use	<b>0.854</b>	0.112	0.125	0.107
Understandable	<b>0.753</b>	0.236	0.284	Understandable	<b>0.731</b>	0.26	0.242	0.215
Relevant	<b>0.653</b>	0.316	-0.108	Enjoyable	0.112	<b>0.86</b>	0.13	0.203
Enjoyable	0.182	<b>0.862</b>	0.156	Attractive	0.203	<b>0.78</b>	-0.102	0.242
Attractive	0.275	<b>0.8</b>		Help	0.262	<b>0.675</b>	0.319	-0.369
Help		<b>0.588</b>	<b>0.553</b>	Retrieve All	0.188		<b>0.933</b>	
Retrieve All	0.242		<b>0.839</b>	Relevant	0.324	0.241	0.141	<b>0.81</b>

The manner in which the variables group into components in Table 4 may suggest that combinations of criteria are particularly important to users. For example, with four components extracted, the first component (easy to use, quick and understandable) might relate to users as they interact with the system and perform specific tasks (*user needs*); whereas the criteria for the second component would highlight aspects related more to the general *user experience* (enjoyable, attractive, help). The third and fourth components reflect individual aspects of *retrieval performance*, which could be measured using recall and precision respectively.

#### 4.4 Qualitative Results

Areas of performance which users responded to positively in open-ended survey questions included time-saving and ease of use from searching multiple OPACs at the same time, as well as support for specific tasks, such as finding items that are unavailable at their home institution. These findings seem to correspond relatively well with components relating to user needs and peripherally with retrieval performance.

In contrast, negative comments focused more on usability and interface design issues, speed of information retrieval, issues with completeness of information, and lack of de-duplication of records for the same item from different institutions, making the information retrieved more difficult to understand. These findings constitute a mix of issues, with the strongest opinions relating to speed, quality and understanding of

information retrieved. Focus group discussions surfaced more on issues relating to system and IR performance. Speed of results was generally seen as less of an issue when feedback on progress is provided on-screen, and the speed demonstrated via the prototype was generally seen as acceptable. However, in a virtual union catalogue speed is largely a factor of the number of records fetched from each institution, and in Search25 a limit has been set to manage the inevitable delays from waiting for all records to be retrieved from each participating institution. Users had mixed opinions about the impact of this, and on balance they preferred to be able to set the number or records fetched themselves, and/or to be able to revert to Retrieve All records when needed. Strongly related to this was a concern that with partial retrieval some of the most relevant records might not be fetched. Interestingly then, when probed in depth, IR performance measures come to the fore and along with speed, and information quality and completeness, are perceived to be the most critical measures of success.

## 5 Discussion

In planning the evaluation of Search25, we asked users about their overall perceived importance of a range of common evaluation criteria (compared to asking users to carry out tasks and then measuring their importance). There are two main limitations to this approach. Firstly, what users *perceive* to be important may change after completing tasks. In this analysis we have involved users who have experience with using the legacy system. Secondly, the importance of criteria that define success may change based on carrying out specific tasks, (e.g., recall would be important for a systematic review where all material on a topic is required), and in different contexts (e.g., speed may be important for completing work tasks compared to leisure tasks). In addition, we recognize that the findings in this paper are related to a specific search system (InforM25) and may not generalize to other types of search setting (e.g. web search). However, these results are still useful in gaining a more general impression of what users view as important criteria against which to evaluate. The results in Section 4 clearly demonstrate that some evaluation criteria are more critical than others, but there are some significant differences in the order of preference by users with different characteristics. Relationships between evaluation criteria are also found, with distinct components identified that group together criteria from diverse (component 1) and similar categories (component 2), according to the Interaction Triptych Framework. In light of these findings, we can summarize answers to the three research questions as follows:

*RQ1: What evaluation criteria are viewed as important by end users?*

Frequency rankings (Table 5) indicate that participants in this study place greater importance on evaluation criteria 1) easy to use, 6) understandable, and 3) quick, related to their *user needs* (corresponding to component 1 in the PCA results), than on criteria 2) attractive, 8) help available, and 7) enjoyable, relating to *user experience* (component 2). *Retrieval performance* measures (components 3 and 4) are placed in the middle of the ranking, but their scores are closer to those for component 1 than for

component 2. These results suggest that a mix of user-centred and system-centred evaluation measures (and methods) are appropriate within the case study context, but that the more subjective experiential and satisfaction type measures may be less critical. The high scores for ease of use and understandable information align with previous results exploring the Interaction Triptych Framework [12], where usability and usefulness criteria rated highly. However, in our findings performance criteria, particularly speed, also rate higher than some of the usability criteria.

**Table 5.** Ranking of evaluation criteria (all users, 4 Agree + 5 Strongly Agree), compared with extracted components

#	Evaluation criteria	Likert 4+5	3 Factor Components	4 Factor Components
1	Easy to use	96%	1	1
6	Understandable	94%	1	1
3	Quick	93%	1	1
4	Retrieve All	88%	1	3
5	Relevant	80%	3	4
2	Attractive	59%	2	2
8	Help available	58%	2	2
7	Enjoyable	44%	2	2

*RQ2: What degree of variation exists between users' preferences?*

User responses were analysed according to a variety of characteristics including role, subject/domain, search experience, and system experience. Rankings for importance of evaluation criteria varied to some degree by each classification of user type, and significant differences between users were found for several characteristics.

User characteristics which demonstrated the greatest amount of difference in preference are user role (three criteria – 2, 3, 6) and experience of library work (four criteria – 2, 3, 4, 6). These evaluation criteria are all in the top half of the ranking by overall importance. These user characteristics identified as a source of significant difference could be used as a basis for user recruitment in future evaluation activities, as well as an aid to interpretation of evaluation results. Evaluation criteria with the most significant differences are 4) retrieve all and 6) understandable, but the user characteristics where differences are found vary for each one. No significant differences were found for criteria 5) relevance and 7) enjoyable, suggesting that these criteria may be less prone to variation by user type, or that the results are inconclusive.

*RQ3: What inter-relationships exist between the different criteria?*

By applying principal component analysis, 3-4 main groupings of evaluation criteria emerge, depending on the number of factors extracted. These groupings correlate well with the frequency rankings. There is a particularly strong inter-relationship between criteria in the *user needs* group (component 1), representing a mix of usability, speed and information quality criteria, that combined together would address some of the primary issues in user and task performance. This is similar to the findings

of [13], where inter-relationships are also found and multiple measures for evaluating success are advocated. However, whilst the system measures 4) retrieve all, and 5) relevant are collocated in the frequency ranking (Table 5), they are situated wide apart and take on separate components (Table 4). This result is possibly explained by the opposing nature of the criteria; one might either want to retrieve everything available, or be interested in only the most relevant results. In system development terms, as raised by our focus group discussions, decisions relating to relevance, recall (retrieve all), and speed of retrieval impact different users in different ways, according to their task and information needs. A note of caution is therefore required in measuring success in these highly ranked, but potentially conflicting evaluation criteria.

It is interesting to note that retrieval speed (quick) correlates highly with ease of use ( $\tau=0.502$ ), as retrieval speed is often a measure that is not assessed in evaluation campaigns such as TREC and CLEF. Previous work has shown significant correlations between user satisfaction and retrieval effectiveness [16], but additionally considering retrieval speed would be an interesting avenue for further investigation as it may suggest that speed, which could be measured using system-oriented approaches (e.g., test collections), may correlate well with user satisfaction that would otherwise have to involve user studies that are more costly to run.

## 6 Conclusions

In this paper we have described our experiences with gathering information from a sample of 196 end users of an operational search system (academic library catalogue) regarding success criteria. This is important when planning an evaluation in deciding what to evaluate, particularly in the case when evaluation methods must be selected due to resource limitations. We find overall that users rate criteria relating to user needs (as confirmed using PCA), such as ease of use, most highly; in contrast with aspects such as whether the system is enjoyable. The results are used to help us select certain kinds of evaluation criteria that will more likely match users' perceived importance. Understanding how users think about success is an important, and often overlooked, aspect of IR evaluation. Future work will carry out similar exercises within different search contexts and explore inter-relationships between criteria.

**Acknowledgments.** We gratefully acknowledge funding from JISC through the Search25 project and contributions from survey and focus group participants. Work also partially funded by the PROMISE network of excellence (contract no. 258191) project as a part of the 7th Framework Program of the European commission (FP7/2007-2013).

## References

1. Saracevic, T.: Evaluation of evaluation in information retrieval. In: Fox, E., Ingwersen, P., Fidel, R. (eds.) Proc. 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, July 9-13, pp. 138-146. ACM Press, New York (1995)

2. Harman, D.: Information retrieval evaluation. Synthesis Lectures on Information Concepts, Retrieval, and Services, vol. 3(2). Morgan & Claypool Publishers, San Raphael (2011)
3. Robertson, S.E., Hancock-Beaulieu, M.: On the evaluation of information retrieval systems. *Information Processing and Management* 28(4), 457–466 (1992)
4. Robertson, S.: On the history of evaluation in IR. *Journal of Information Science* 34(4), 439–456 (2008)
5. Voorhees, E.M., Harman, D.K.: TREC: experiments and evaluation in information retrieval. MIT Press, Cambridge (2005)
6. Ingwersen, P., Järvelin, K.: The turn: integration of information seeking and retrieval in context. Springer, New York (2005)
7. Borlund, P.: User-Centred Evaluation of Information Retrieval Systems. In: Göker, A., Davies, J. (eds.) *Information Retrieval: Searching in the 21st Century*. John Wiley & Sons, Chichester (2009)
8. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3(1-2), 1–224 (2009)
9. van Rijsbergen, C.J.: *Information retrieval*, 2nd edn. Butterworths, London (1979)
10. Saracevic, T.: Digital Library Evaluation: Toward Evolution of Concepts. *Library Trends* 49(2), 350–369 (2000)
11. Fuhr, N., et al.: Evaluation of Digital Libraries. *International Journal on Digital Libraries* 8(1), 21–38 (2007)
12. Tsakonias, G., Papatheodorou, C.: Exploring Usefulness and Usability in the Evaluation of Open Access Digital Libraries. *Information Processing & Management* 44(3), 1234–1250 (2008)
13. Buchanan, S., Salako, A.: Evaluating the Usability and Usefulness of a Digital Library. *Library Review* 58(9), 638–651 (2009)
14. Nielsen, J.: Enhancing the Explanatory Power of Usability Heuristics. In: *Proc. SIGCHI Conference on Human Factors in Computing Systems*, pp. 152–158. ACM Press, New York (1994)
15. Toms, E.G., O'Brien, H.L., Kopak, R., Freund, L.: Searching for relevance in the relevance of search. In: Crestani, F., Ruthven, I. (eds.) *CoLIS 2005*. LNCS, vol. 3507, pp. 59–78. Springer, Heidelberg (2005)
16. Al-Maskari, A., Sanderson, M.: A Review of Factors Influencing User Satisfaction in Information Retrieval. *Journal of the American Society for Information Science and Technology* 61(5), 859–868 (2010)
17. Xie, H.: Users' Evaluation of Digital Libraries (DLs): Their Uses, Their Criteria, and Their Assessment. *Information Processing & Management* 44(3), 1346–1373 (2008)
18. Xie, H.: Evaluation of Digital Libraries: Criteria and Problems from Users' Perspectives. *Library and Information Science Research* 28(3), 433–452 (2006)
19. Kelly, D., et al.: Evaluation Challenges and Directions for Information-Seeking Support Systems. *Computer* 42(3), 60–66 (2009)
20. Marchionini, G.: *Information seeking in electronic environments*. Cambridge University Press, Cambridge (1995)
21. Hölscher, C., Strube, G.: Web Search Behavior of Internet Experts and Newbies. *Computer Networks* 33(1), 337–346 (2000)

# A Dependency-Inspired Semantic Evaluation of Machine Translation Systems

Mohammad Reza Mirsarraf<sup>1</sup> and Nazanin Dehghani<sup>1,2</sup>

<sup>1</sup> CyberSpace Research Institute, Tehran, Iran

<sup>2</sup> ECE Department, University of Tehran, Tehran, Iran  
{mirsarraf, ndehghani}@csri.ac.ir

**Abstract.** The goal of translation is to preserve the original text meaning. However, lexical-based machine translation (MT) evaluation metrics count the similar terms in MT output with the human translated reference rather than measuring the similarity in meaning. In this paper, we developed an MT evaluation metric to assess the output of MT systems, semantically. Inspiring by the dependency grammar, we consider to what extent the headword and its dependents contribute in preserving the meaning of the original input text. Our experimental results show that this metric is significantly better correlated with human judgment.

## 1 Introduction

Evaluating MT systems has necessarily received significant attention alongside the development of MT systems themselves. In recent years, a number of automatic evaluation metrics have been proposed [1–3]. The main assumption behind developing such metrics is that ‘acceptable’ translation tends to share the lexicon with a predefined set of manual reference translations. Though, this assumption works well in many cases, this method of estimating is not a trivial task especially in morphological languages such as Persian and Arabic.

There exists a wide range of lexical based evaluation metrics all perform well in capturing the translation fluency, but in some cases they strongly disagree with human judgment. Koehn [4] itemized BLEU drawbacks as not considering the relative relevance of different words and not considering the overall grammatical coherence, which are common in most lexical-based evaluation metrics. The underlying reason is that lexical similarity does not reflect the similarity in meaning [5]. The aforementioned drawbacks opened a new field of research to semantically evaluate the MT outputs. Towards this, Semantic Role Labeling (SRL) is used as one of the major steps in representing text meaning.

C. Lo and his co-worker in [6] introduced the methodology that unlike conventional n-gram based MT evaluation metrics, it measures the utility of translations using SRL. After that, they proposed MEANT [5] that evaluates the utility of translation more accurately via Propbank-style semantic roles.

In this paper, we propose a dependency-inspired semantic MT evaluation metric to quantify how well the essential meaning of the source is kept in the

translated output by utilizing the concepts of dependency parsing in SRL. Our proposed method differs from [5] in weighting the matched headwords and dependents of each argument. In fact, the headwords of each constituent supply more information about the meaning of a sentence while dependents are less important.

The rest of the paper is organized as follows. In Section 2, we briefly review the semantic role concepts. In Section 3, we propose the formulation of our metric and the experimental results. Finally we conclude the paper in Section 4.

## 2 Semantic Annotation

Semantic role labeling is a process of identifying the semantic arguments associated with the predicate of a sentence and assigning specific roles to them. In fact, the purpose of SRL is to extract the semantic structure of a sentence so that the reader comprehends “who did what to whom, when, where and why” [5].

Although, there exist good resources of English annotated corpus with semantic roles [7,8], to the best of our knowledge, there is no such dataset in Persian. For the task of this paper, we added a number of roles to the role set in [8] to enrich our semantic description. It includes Agent, Patient, Source, Goal, Topic, Percept, Beneficiary, Time, Location, Manner, Reason, and Indefinite.

To perform MT evaluation more precisely, we label the headword with its semantic role while its dependents in a constituent are also determined. During the translation, the head words of each constituent supply more information about the meaning of a sentence while dependents are less important. This fact is the basis of our proposed MT evaluation metric.

## 3 Methodology and Experiments

To evaluate MT systems semantically, we construct an evaluation metric which counts the degree of match between SRL of the human translated reference versus machine translations of sentences. Then, for each correctly translated predicate, we calculate a weighted fraction of correctly translated headwords and their dependents. The parameter  $\alpha$  can be viewed as the importance of meaning preservation for the semantic role of headwords. We define our metric in terms of an f-measure that balances the precision and recall, as follows.

$h_i$  = #of correct headwords of ARGs of PRED  $i$  in MT

$d_i$  = #of correct dependents of ARGs of PRED  $i$  in MT

$H_{M_i}$  = total #of headwords of ARGs of PRED  $i$  in MT

$D_{M_i}$  = total #of dependents of ARGs of PRED  $i$  in MT

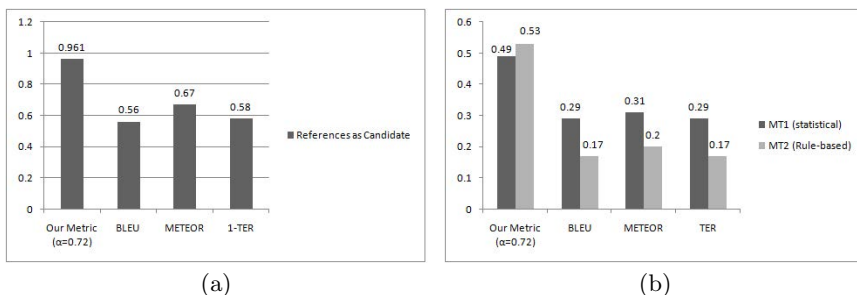
$H_{R_i}$  = total #of headwords of ARGs of PRED  $i$  in reference

$D_{R_i}$  = total #of dependents of ARGs of PRED  $i$  in reference

$$\text{Precision} = \frac{\sum_{\text{matched predicate } i} \frac{\alpha * h_i + (1-\alpha) * d_i}{\alpha * H_{M_i} + (1-\alpha) * D_{M_i}}}{\text{total \#of predicates in MT}}$$

$$\text{Recall} = \frac{\sum_{\text{matched predicate } i} \frac{\alpha * h_i + (1-\alpha) * d_i}{\alpha * H_{R_i} + (1-\alpha) * D_{R_i}}}{\text{total \#of predicates in reference}}, \text{ F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$





**Fig. 1.** a) Evaluation of human translated references with different metrics. b) Semantic evaluation of Persian statistical and rule-based MT outputs versus other metrics.

If all the reconstructed semantic frames in the MT output are identical to the gold standard annotation in the reference translation, and all the arguments in the reconstructed frames are judged to express the same meaning as the corresponding arguments in the reference translations, then the f-score in the above definition will be equal to 1, regardless of the values of  $\alpha$ .

In order to perform our experiments, we included the English corpus with 1031 sentences from news sites consisting of BBC, CNN, Reuters and VOA. We asked professional translators to provide us the Persian translated side of this corpus as reference text. Then, the Persian statistical [9] and rule-based [10] MT systems were given the English corpus to translate. After that, Persian linguistics were given the role set to annotate the reference and Persian outputs of MT systems. They also provide us the headword and its dependents in each argument of a predicate. The F-measure is then calculated to quantify semantic frame match between the aforementioned MT systems outputs and the reference. Before evaluating the MT outputs, we obtained the optimal value 0.72 of parameter  $\alpha$  using a validation set. This parameter regulates the extent to which matching the headwords contribute to the overall score.

Since we are not provided with adequacy nor fluency scores of a candidate file to measure the correlation between human evaluation and the automatic one, we consider another way to determine the effect of our proposed metric. Having four annotated reference texts, we consider one of them as candidate and the other file as reference and then ask the evaluation metric to give us the score. This experiment is repeated for other permutations of four annotated references and the average scores are reported. The higher the evaluation score is, the more the evaluation metric is close to human judgment. This proposed approach is justifiable since the candidate and the reference are all translated by human experts and thus we expect high scores from evaluation metric. Figure 1(a) illustrates the results of this experiment. The obtained results indicate that there exists a far distance between the correlation of our proposed metric and other lexical-based metrics, and the human judgment.

Our second experiment is dedicated to semantically evaluating the Persian SMT output versus the rule-based, and then compare the obtained results with

other evaluation metrics. Figure 1(b) illustrates the scores of these two MT systems given by four evaluation metrics, our proposed metric with  $\alpha = 0.75$ , BLEU, METEOR and TER. Experimental result reveals that the rule-based system is given higher score from our metric while the other metrics give higher score to the statistical system. The main point of this experiment is confirming the fact that lexical based evaluation metrics such as BLEU, operates only on a very local level and does not address overall grammatical coherence and this biases the metric in favour of phrase-based statistical systems, which are good at producing good n-grams, but less able to produce meaningful coherent sentences.

## 4 Conclusion

While the goal of translation is to preserve the full meaning of the original text, many popular methods for MT evaluation are often depends on n-gram and lexical terms which do not take the meaning into consideration. In this paper, we proposed our semantic MT evaluation metric utilizing semantic role labeling. Furthermore, we considered to what extent the headword and its dependents contribute in preserving the meaning of the input text. The results show that our proposed metric is significantly better correlated with human judgment.

## References

1. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311–318 (2002)
2. Snover, M., Dorr, B., Schwartz, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: Proc. of the 7th Conference of the Association for Machine Translation in the Americas, Cambridge, MA, pp. 223–231 (2006)
3. Denkowski, M., Lavie, A.: METEOR-NEXT and the METEOR Paraphrase Tables: Improve Evaluation Support for Five Target Languages. In: Proc. of the ACL WMT/MetricsMATR 2010 (2010)
4. Koehn, P.: What is a Better Translation? Reflections on Six Years of Running Evaluation Campaigns. Tralogy (2011)
5. Lo, C., Wu, D.: MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 220–229 (2011)
6. Lo, C., Wu, D.: Evaluating Machine Translation Utility via Semantic Role Labels (2010)
7. Kingsbury, P., Palmer, M.: Propbank: the next level of treebank. In: Proc. of Treebanks and Lexical Theories (2003)
8. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proc. of the COLING-ACL Conference, Montreal, Canada, pp. 86–90 (1998)
9. Targoman, <http://targoman.com/smt.php>
10. Farazin, 217.218.62.239

# A Turing Test to Evaluate a Complex Summarization Task

Alejandro Molina<sup>1</sup>, Eric SanJuan<sup>1,2</sup>, and Juan-Manuel Torres-Moreno<sup>1,2,3</sup>

<sup>1</sup> LIA, Université d'Avignon et des Pays de Vaucluse,  
339 chemin des Meinajaries, Agroparc BP 1228, F-84911 Avignon Cedex 9, France  
alejandromolina.villegas@alumni.univ-avignon.fr,  
{eric.sanjuan,juan-manuel.torres}@univ-avignon.fr

<sup>2</sup> Brain & Language Research Institute,  
5 avenue Pasteur, 13604 Aix-en-Provence Cedex 1, France

<sup>3</sup> École Polytechnique de Montréal,  
2900 Bd Edouard-Montpetit Montréal, QC H3T1J4, Canada

**Abstract.** This paper deals with a new strategy to evaluate a Natural Language Processing (NLP) complex task using the Turing test. Automatic summarization based on sentence compression requires to assess informativeness and modify inner sentence structures. This is much more intrinsically related with real rephrasing than plain sentence extraction and ranking paradigm so new evaluation methods are needed. We propose a novel imitation game to evaluate Automatic Summarization by Compression (ASC). Rationale of this Turing-like evaluation could be applied to many other NLP complex tasks like Machine translation or Text Generation. We show that a state of the art ASC system can pass such a test and simulate a human summary in 60% of the cases.

## 1 Introduction

Alan Turing predicted that computers will be better at playing complex board games like chess than to chat with humans in an open world. Natural Language Processing (NLP) appeared in 1951 to be one of the greatest challenges for computers. Surprisingly, some tasks like automatic summarization appeared to be easier than anticipated when considering extracts instead of abstracts [1]. Summarization by extraction often consists in segmenting the text to be summarized into sentences and to apply scoring methods to rank sentences by decreasing informativity. In this simplified task, resulting short summaries are often readable because they use real sentences. The main difficulty when dealing with longer summaries involving ten or more sentences is to avoid breaking anaphora. This is handled using simple heuristics like displaying top ranked sentences in the order they appear in the original text. Since local text grammaticality is ensured by keeping entire sentences, resulting summaries often give the illusion that they were written by a human. Moreover, under the assumption that the produced summary is readable, summary informativeness can be evaluated using measures like ROUGE given on a set of reference summaries or Jensen-Shannon/Kullback-Leibler metrics if no reference summary is available [2–4].

The task becomes much more complex if computer cut and compress sentences like humans do since this implies the ability to understand and modify inner sentence structures. Discourse structure among other implicit semantic relations play a key role [5]. Moreover there are usually several correct ways to compress a sentence and human experts often disagree on which is the best one. When trying to build a reference corpus of compressed sentences, inter agreement between annotators is low, even to decide if a sentence should be shortened in the summary or not. Automatic Summarization by Compression (ASC) requires to handle a high level of uncertainty in the decision process since there is not a best way to compress a sentence, only observations that sometimes humans prefer one way rather than another one [6]. Not only the task itself is difficult but it cannot be evaluated using existing methods. Using sentence compression to produce a summary not always improve informativeness scores and can produce unreadable summaries. Therefore, actual state of the art evaluation metrics for automatic summarization discourage thorough investigations if a computer can handle or not ASC.

In this paper we show that coming back to the original idea of a Turing test, it is possible to set up a simple imitation game to evaluate ASC. We also show that a state of the art system that learns human behavior using simple regression analysis [6] can pass this test on short summaries and give the illusion to human referee that the summary was written by a human. Moreover this imitation game is clearly adapted to crowd sourcing through Internet and can be used to evaluate large amount of systems at a reasonable cost.

The rest of the paper is organized as follows. Section 2 goes back to the general definition of a Turing test. Section 3 details the imitation game that we propose to evaluate ASC in a pragmatic way. Section 4 shows statistical evidence that a state of art ASC system can pass the test. Finally, section 5 opens perspectives on how this evaluation methodology can contribute to the improvement of effective ASC systems.

## 2 Back to Turing Test

As suggested by Alan Turing, a test to evaluate the ability of a computer to handle a human mind task should involve:

- an interaction with humans where the computer tries to give the illusion that it is human,
- a clear evaluation metric that allows the reproducibility of the experiment,
- a gateway to the open world to explore beyond restricted contexts and closed world assumptions.

Our main motivation relies on the fact that, to the best of our knowledge, there is no summarization evaluation methodology that encourages research on advanced NLP tasks like summarization by sentence compression. We therefore suggest to come back to Turing’s initial motivations[7] when imaging imitation games to answer the controversial philosophical question “do computers have a mind?” without having to define what “mind” means. The question then becomes “what are the common human intellectual tasks that a computer can handle?” These are the roots of theoretical computer science where tasks almost useless for technical applications can be fundamental to understand computers’ real limits. ASC can have many applications in our interconnected

world but we claim that its main interest relies on the theoretical study of computer capabilities.

In a the original imitation game defined by Turing in [7], there are two players and one assessor. The first player is a human ( $A$ ) and the second a computer ( $B$ ). Another human ( $C$ ) plays the role of the assessor and has to guess the real nature (human or computer) of the two other players. The assessor cannot see the other players, he can just interact with them through a more or less restricted interface that at least allows to exchange written messages. The assessor asks questions through the interface and has to distinguish between answers given by the human player and those sent by the computer.

Turing imagined advanced imitation games to study the spectrum of Artificial Intelligence and compare it to the human mind. However, as pointed out by [8], Turing entrusted interaction through natural language. In our case, we intend to study the method of interacting itself related to NLP and its linguistic functionalities based on summary generation. Indeed, in the general case of a Turing test, the assessor is not allowed “to see” the players. This is to ensure that he focus on functional aspects and not on appearances. It then seems natural to adapt the imitation game to NLP tasks that try to reproduce human ability to handle texts like summarization. We do not consider tasks that cannot be carried out without computer assistance like Information Retrieval from large collections. Only intellectual tasks that can easily be accomplished by non experts meanwhile there are real challenges for an automatic system.

### 3 Imitation Game to Evaluate ASC

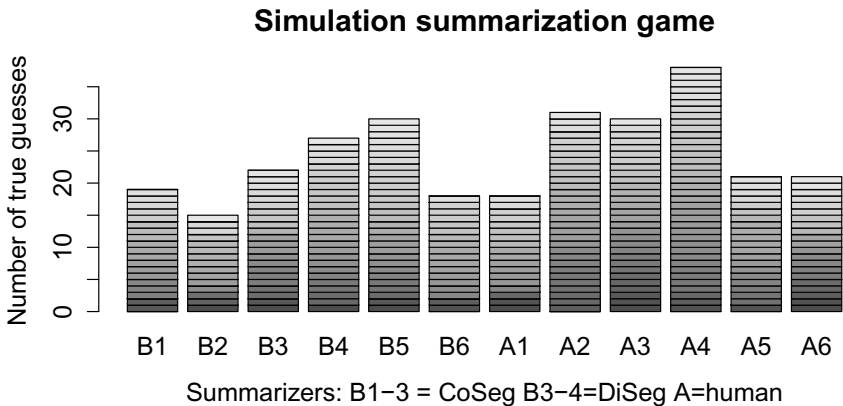
We consider the following imitation game involving a human player ( $A$ ), a computer ( $B$ ) and a human assessor ( $C$ ).  $A$  and  $B$  are asked to write one summary for each of some texts. After some time, an interface between  $C$  and the players dispatches the summaries at random, just checking that each player have the same number of texts. So  $C$  does not know who between  $A$  and  $B$  wrote each summary and has to guess the correct author for each text.

This setting follows Turing’s idea of an interactive game between two humans and a computer. However, one difficulty to carry it out is that humans need time to write a summary meanwhile it is necessary to reproduce the same experiment at least 30 times to expect some statistical evidence if there is a regular winner between  $A$  and  $B$ . To adapt this game to standard crowd-sourcing evaluations, we decided to consider a team of  $n \geq 30$  extra assessors ( $C_1, \dots, C_n$ ), a team of different human players ( $A_1, \dots, A_k$ ) and a set of different computational strategies or systems ( $B_1, \dots, B_m$ ). The main drawback of this adaptation is the lack of real interactivity. The main advantage is that this rationale could be adapted to many other domains using machine learning to simulate particular human brain functionality (for instance, NLP complex tasks).

Let us give some details about the way we implemented this game to evaluate ASC systems. 60 post-graduated students accepted to participated in this simulation game, 6 of them were asked to write summaries ( $A_1, \dots, A_6$ ) and the 54 other participated as assessors ( $C_1, \dots, C_{54}$ ). It must be note that all of them, team ( $A$ ) and team ( $C$ ), expected team ( $B$ ), the systems, to fail. 12 texts were selected from the RST Spanish Tree

Bank[9] at random. Summaries of these texts have been written down by team (*A*). We chose the ASC systems derived from [6] as the team (*B*). These summarizers are based on machine learning techniques that emulate the way annotators agree or not with a sentence compression using two discourse segmentation strategies: DiSeg [10]<sup>1</sup> (*B*<sub>1</sub>) and CoSeg (*B*<sub>2</sub>). It has been shown in [5] that humans tend to remove complete discourse units from sentences when they try to compress them. As anticipated for a so subjective task, inter agreement between assessors was very low but enough to carry out a regression analysis and learn to predict the probability of a particular sentence compression to be accepted by humans. Three summaries of different length (short, medium and long) were generated using DiSeg (*B*<sub>1</sub>), and three other ones also of different length were generated using CoSeg (*B*<sub>2</sub>). All assessors read the 12 summaries and for each they tried to guess if the author of the summary was a human or a computer. They did not know that exactly half of the summaries were automatically generated.

## 4 Results

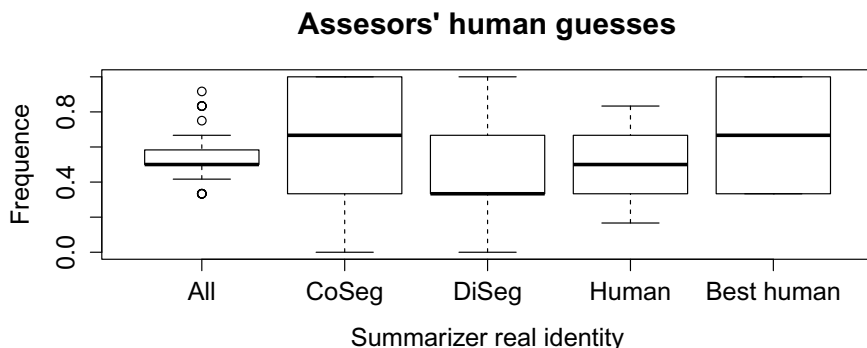


**Fig. 1.** summarization simulation game: each bar shows the number of correct guesses (Human or Computer) for each summarizer

Figure 1 shows the results of the simulation game. where numbers next to (*B*) team represent the three different lengths (1:short, 2:medium, 3:expanded). In this figure, bars for summaries written by team (*A*) are expected to be higher if they are good quality summaries meanwhile bars for team (*B*) are expected to be low since they intent to mislead the assessor. It appears that over the six authors of summaries, only three manage to write summaries that more that 60% of the assessors think they can not be automatically generated. Meanwhile, player (*B*<sub>1</sub>), the automatic system DiSeg, manage to mislead the assessors on 60% of long summaries and player (*B*<sub>2</sub>), CoSeg system on short and medium summaries.

<sup>1</sup> <http://diseg.termwatch.es>

Plot 2 shows the median normalized frequency of times that an assessor thinks the summary has been written by a human. The first boxplot shows it over the twelve summaries each assessor has to read. The second and third boxplots over the three summaries generated using CoSeg and DiSeg respectively. The fourth is over the six summaries by humans and the last one is restricted to the three best authors  $A_2$ ,  $A_3$  and  $A_4$ . These boxplots suggest that summary quality by three best authors (last box-



**Fig. 2.** Boxplots showing the median number of times that an assessor thought it was a summary produced by a human for each set of six summaries and each subset of three summaries automatic/human

plot) is above average among summaries written by real authors (fourth boxplot) and among overall summaries (first boxplot). However, according to a Wilcoxon test with a  $p$ -value lower than 0.01, only the differences between best human summaries and all human summaries is statistically significant. The difference between best human summaries and overall summaries is not. Similarly, CoSeg summaries outperform DiSeg summaries since the median frequency it misleads assessors is significantly higher ( $p$ -value  $< 0.05$ ) meanwhile all other differences are not statistically significant. In particular there is not statistical evidence based on Wilcoxon rank sum test with continuity correction that an assessor thinks that the summary has been done by a human author when reading a summary generated by one of the automatic summarizers tested here, than one really done by a human author.

## 5 Discussion

We have use a Turing test to evaluate two state of the art automatic summarizers where usual evaluation protocols failed to differentiate between quality levels among the two system outputs. The principal argument is that if human and machine productions could not be differentiated, then they might have similar quality.

The experiment set up here with 60 human players gives statistical evidence that one system outperforms the other. But we also find out that human juges cannot differentiate between written by an author abstracts and automatically generated summaries when

using sophisticated methods as ASC that goes beyond sentence extraction and ranking. Results are promising, though this to be checked out by setting up a larger crowdsourcing experiment and testing some enhancements. For instance, this first experiment cannot quantify the gap in quality between the good and bad summaries. However, mixing human and machine outputs using Turing test adapted to specific tasks could represent a new evaluation paradigm that need to be more explored. Even more, we think that it could have broader applications.

**Acknowledgments.** We would like to thank our contributors for their help with the corpus annotation, specially to Gerardo Sierra and the Linguistics Engineering Group UNAM. This work was partially supported by the CONACyT grant 211963 and the ANR Imagweb project (ANR-12-CORD-0002).

## References

1. Tratz, S., Hovy, E.: Summarisation Evaluation Using Transformed Basic Elements. In: Workshop Text Analysis Conference (TAC 2008), Gaithersburg, MD, USA (2008)
2. Louis, A., Nenkova, A.: Automatically Evaluating Content Selection in Summarization without Human Models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Singapour, August 6-7, pp. 306–314. ACL (2009)
3. Saggion, H., Torres-Moreno, J.-M., da Cunha, I., SanJuan, E.: Multilingual summarization evaluation without human models. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING 2010), Beijing, Chine, pp. 1059–1067. ACL (2010)
4. Torres-Moreno, J.-M., Saggion, H., da Cunha, I., SanJuan, E.: Summary Evaluation With and Without References. *Polibits: Research Journal on Computer Science and Computer Engineering with Applications* 42, 13–19 (2010)
5. Molina, A., Torres-Moreno, J.-M., SanJuan, E., da Cunha, I., Sierra, G., Velázquez-Morales, P.: Discourse segmentation for sentence compression. In: Batyrshin, I., Sidorov, G. (eds.) MICAI 2011, Part I. LNCS, vol. 7094, pp. 316–327. Springer, Heidelberg (2011)
6. Molina, A., Torres-Moreno, J.-M., SanJuan, E., da Cunha, I., Martínez, G.E.S.: Discursive sentence compression. In: Gelbukh, A. (ed.) CICLing 2013, Part II. LNCS, vol. 7817, pp. 394–407. Springer, Heidelberg (2013)
7. Turing, A.M.: Computing machinery and intelligence. *Mind* 59(236), 433–460 (1950)
8. Harnad, S.: Minds, Machines and Turing. *Journal of Logic, Language and Information* 9(4), 425–445 (2000)
9. da Cunha, I., Torres-Moreno, J.-M., Sierra, G.: On the Development of the RST Spanish Treebank. In: Linguistic Annotation Workshop, pp. 1–10. The Association for Computer Linguistics (2011)
10. da Cunha, I., SanJuan, E., Torres-Moreno, J.-M., Lloberes, M., Castellón, I.: DiSeg 1.0: The first system for Spanish Discourse Segmentation. *Expert Systems with Applications* 39(2), 1671–1678 (2012)



# A Formative Evaluation of a Comprehensive Search System for Medical Professionals

Veronika Stefanov<sup>1</sup>, Alexander Sachs<sup>2</sup>, Marlene Kritz<sup>2</sup>, Matthias Samwald<sup>3</sup>,  
Manfred Gschwandtner<sup>2</sup>, and Allan Hanbury<sup>1</sup>

<sup>1</sup> Information & Software Engineering Group, Institute of Software Technology and Interactive Systems, Vienna University of Technology, Vienna, Austria

<sup>2</sup> Society of Physicians in Vienna, Vienna, Austria

<sup>3</sup> Section for Medical Expert and Knowledge-Based Systems, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria

**Abstract.** Medical doctors need rapid and accurate answers, which they cannot easily find with current search systems. This paper describes a formative evaluation of a comprehensive search system for medical professionals. The study was designed to guide system development. The system features included search in text and 2D images, machine translated summaries of search results, as well as query disambiguation and suggestion features, and a comprehensive search user interface. The study design emphasizes qualitative user feedback, based on realistic simulated work tasks and data collection with spontaneous and prompted self-report, written and spoken feedback in response to questionnaires, as well as audio and video recordings, and log files. Results indicate that this is a fruitful approach to uncovering problems and eliciting requirements that would be harder to find in a component-based evaluation testing each feature separately.

## 1 Introduction

Medical doctors need rapid and accurate answers — a search of MEDLINE takes on average 30 minutes [1], while doctors have on average 5 minutes available for such a search [2]. Furthermore, over 40% of searches do not yield the information required [3].

We report on an evaluation of a system that aims to bridge this gap. It is a case study where we studied a small number of participants in greater detail. The study is positioned at the human end of the research continuum of interactive information retrieval research [4]. The main goals of this evaluation were to find problems with the current state of the system, and to collect requirements and desired features for the next version of the system. The study reported here is part of a longer, iterative process. For such an evaluation only a handful of test participants are required to find the larger part of the errors and bugs of a system. This study is noteworthy as it uses the actual target users (medical doctors) to test the system, rather than the surrogates often used, e.g. library and information science students instead of intelligence analysts in [5].

The paper has two contributions: (1) It reports the process of designing and organising user-centred evaluations for busy professional users, which is useful knowledge for researchers undertaking a similar task. Such evaluations provide important information for developing a professional search system, but are not often done, due to the large effort involved. (2) It reports some of the key findings obtained in the evaluations. In particular, the findings related to working in a multilingual environment provide useful insight.

After a short introduction to the KHRESMOI system and a description of the study design, we report findings and our future research plans.

## 2 KHRESMOI

KHRESMOI – “Knowledge Helper for Medical and Other Information participants” – is an information access and retrieval system for biomedical information, targeting the general public, physicians, and specifically radiologists, developed within the framework of an EU research project; it integrates and extends open-source components from several partners:

**User Interface:** comprehensive search user interface, based on ezDL [6]

**Text Search:** search capabilities over **annotated text**, based on Mimir [7]

**Image Search:** content-based 2D image search, based on ParaDISE [8]

**Query Disambiguation:** query disambiguation and suggestion services, based on OWLIM [9]

**Translation:** Multilingual tools for query and document translation (based on MOSES)

**Spell-Checking Tools:** spelling suggestion and correction service

The evaluations reported in this paper were done at the prototype stage around the midpoint of the project. At this point, the system had a simple web interface and a comprehensive desktop client, both of which offered integrated text/2D image search, and a radiology interface with 3D image search. The evaluation reported in this paper applies only to the desktop search client for medical doctors. Data sources included HONcode certified medical websites [10], scientific journals, and images. The client software had features for unstructured search of text and 2D images, machine translated excerpts for some languages, result classification and filters.

## 3 Methodology

### 3.1 Participants

The “KHRESMOI for professionals” search client is meant for medical professionals. Participants were recruited from the Society of Physicians in Vienna, whose members are required to have completed a medical degree. In total, the study involved 14 tests and 5 pilot tests, for a total of 19 participants, performing one-hour sessions each. Participants received an Amazon voucher worth 50

Euros as a token of appreciation. We aimed to have various age groups, specialists and general practitioners, physicians in training and medical professors participating, but overall young physicians in their first two years after graduation comprised the majority of participants. 57% were male and 43% female, and only one a general practitioner.

### 3.2 Task Scenario Development

The tasks were written based on the findings of a survey on the use of online medical resources and search tools by European physicians reported in [11]. Initially, separate tasks for different groups of physicians were defined, as the survey results lead to the expectation that they would behave differently. But during the pilot tests, the limitations of such an approach at this stage of development became obvious. We would not have obtained reliable results. The tasks were consequently simplified and adapted to the available resources and functionality, and their number was reduced to four.

In the final evaluations, all participants were asked to solve the same tasks: a treatment decision, diagnosis based on a textual description, diagnosis of an x-ray image, and a scientific task.

The tasks are simulated work tasks [12]. Each task had a context story attached to it and asked the participant to employ one or more of the KHRESMOI tools during the tasks. For example, task 4 is based on the scenario that the participant is collaborating with a colleague and should therefore try to use the export function to share findings.

To illustrate, Figure 1 shows the description of task 1.

### 3.3 Procedure

The study was administered in single-mode, with each participant working alone, one at a time. The researcher had a checklist that guided through the test protocol to ensure consistency between different runs of the test.

**Pilot Testing.** In order to identify problems with the instructions, schedule or test management software, a number of pilot tests were performed.

The first pilot test already revealed that the tasks could not fit the proposed time frame. In response to the second and third pilot tests, time management was adapted and confirmed: 10 minutes for the introduction and demographics, 40 minutes for four tasks (10 minutes each) and 10 minutes for the overall feedback at the end of test.

The pilot tests also showed that for two tasks, not enough resources were available in the system to allow participants to find the correct answers. For one of these tasks, more data was added, and for the other task, the topic was changed. This made it possible to solve all tasks, but already showed that the system was not ready for a real life medical situation with unpredictable information needs.

<p><b>Atrial Fibrillation (10 minutes)</b></p> <p>Is it ok for a 69 year old women with a history of atrial fibrillation and cardioversion to stop anticoagulation due to recent rhythm stability?</p> <p><b>Case scenario:</b> A 69 year old woman, diagnosed 4 years ago with atrial fibrillation has successfully received cardioversion. That time she felt elevated heart rate and palpitations and is taking oral anticoagulants. Since then she is symptom free. She is health conscious and regularly measures her heart rate, which seems ok. She is otherwise healthy, her heart has a normal structure, only the left ventricle shows a moderate enlargement. She wants to stop oral anticoagulants.</p> <p>From your knowledge: Is it ok for her to STOP taking oral anticoagulants?</p> <p><input type="checkbox"/> Yes</p> <p><input type="checkbox"/> No</p> <p><input type="checkbox"/> I don't know / I require further information to answer this question</p> <hr/> <p>Please use KHRESMOI to find the (evidence to support your) answer and cite at least 3 websites (or more until you are confident in your answer) that you consider supportive.</p>
---

**Fig. 1.** Task 1: Treatment Question (abridged version)

During the pilot tests, all problems were immediately reported to the development teams. Feature and data requests were initially collected, and weekly online meetings were held between the evaluation team and the people responsible for the software components, reviewing findings and deciding on the urgency of changes.

After 5 pilot tests, it was decided to postpone the evaluations and instead make a number of larger changes to the prototype system. All issues marked as critical were solved before the evaluations. Changes included:

- added more datasets for the task scenarios
- removed dead links, double results, corrected empty metadata
- added a redundant way to access the translation feature
- various bug-fixes to the spelling correction, result preview presentation (tool tips, missing titles), and the UI in general (moved scrollbars, icons)

**Schedule.** Each session consisted of three parts: In the first part the participant was introduced to the test. After a brief introduction to KHRESMOI and the goals of the session a consent form was signed. All information collected was used exclusively for the purpose of the study and was kept confidential. The participant was then introduced to the interface of the search engine for physicians and led through a short tutorial case with basic information on how to use the interface. This tutorial was given in person by the researcher coordinating the

session. There was also a short tutorial available within the client user interface, but as expected, people did not read this. After the introduction, the recording with the test management software was started. The software then led the participant through the steps of the test, while the tasks were started manually by the researcher. As a first step, the participant had to fill in the demographics questionnaire. This part took 10 minutes.

In the second part the participants were asked to perform four different search tasks (treatment, diagnosis, image diagnosis, scientific), representing real-life scenarios. 10 minutes were attributed to each task. For each task, the initial knowledge (without external help) was collected from the participant. Then the KHRESMOI search engine was used to find the answer, or to support existing knowledge. Finally, feedback was provided in a short questionnaire and/or verbally to the researcher.

In the third part of the session the participant was asked to complete a questionnaire evaluating how satisfied he or she was with the search system. This questionnaire consisted of questions from the standard SUS questionnaire [13] plus additional questions specific to KHRESMOI. This overall feedback represented the end of the test and had to be given in 10 minutes. The recording was terminated and stored for future analysis. In total the session required 60 minutes to complete.

All sessions took place in the offices of the Society of Physicians in Vienna. The researcher who provided the introduction remained in the same room as the participant during the test to be available for questions and comments. However, he remained in the background and pretended to work on something else, to avoid distracting the participant. During the recording with the test management software, a second researcher was connected to the participant's computer from another room, watching the session via video and audio live streaming and adding observations and comments as annotations to the recording. Participants were aware of this setup.

## Collecting Data

*Questionnaires.* Similar to the tasks, the questionnaires improved with the pilot tests. The wording was clarified, and the number of questions was optimized to strike a balance between the needs of researchers and participants. Table 1 gives an overview of the questionnaires, which were displayed by the test management software. The session started with a demographics questionnaire. After each task participants were asked for the answer, supporting website(s)/images, and three feedback questions: what they liked, what they disliked, and what functionality or aspect they had missed. This allowed instant feedback in relation to solving a concrete task. After completing the last task, participants were asked to complete a longer questionnaire with overall feedback on the system.

*Observation, Logging, and Self-Report.* Apart from observation with the recording and annotation software described above, we also had access to the logfiles

**Table 1.** Overview of questionnaires

Questionnaire	How often	When	Number of questions
Demographics	once	before the first task	12
Task feedback	for each task	after the task	3 (all freetext)
Overall feedback	once	after the last task	21 (20 Likert scale, 1 freetext)

of the searches, which included queries, clicked results, accepted spelling corrections, and invocations of the translation component.

For direct user feedback, we employed spontaneous and prompted self-report [4]. This methodology is not as ‘unnatural’ as think-aloud, but still useful for gathering detailed feedback from the users as they encounter noteworthy situations. The researcher prompted the participants once during each task, and most spoke quite a lot, explaining and justifying their actions to the researcher. Additionally, in most sessions a conversation ensued after finishing each task.

**Test Management and Recording Software.** As test management and recording software, that combines capturing and recording features with displaying task descriptions and reminders and prompting participants with questionnaires at predefined steps, Morae usability testing software [14] was used for all tests. The recording component of the software ran on the computer used by the participant and recorded the screen, various interaction events such as mouse clicks or window dialogues, and the participant’s face and voice via webcam and microphone. The sequence of tasks and questions as predefined by the experimenters was advanced manually or on autopilot. Additionally, comments and observations by the experimenters, so-called markers, were integrated into the recording, either from a second computer during the recording, or at any time later.

The recordings proved valuable for deciphering the comments made by the participants in the feedback forms. Due to the language situation, with German speaking participants expressing their feedback in written English, many comments in the questionnaires are very brief and/or ambiguous. But since many participants spoke quite a lot during the tests, either prompted by the experimenter, or spontaneously when encountering a noteworthy situation, they provided an additional much richer level of feedback in spoken form. To make use of this source of information, all recordings were reviewed in detail at least one more time, to improve the probability that no comments were lost or misinterpreted.

### 3.4 Data Analysis

One of the sessions encountered severe server problems, and the results had to be discarded, leaving 13 sessions to be evaluated. Two of these included only the first 3 of the 4 tasks.

For each participant session, we read the notes taken during the session and watched the recording and transcribed any additional comments of the

participants. During weekly online meetings with project team members from the development and other evaluation teams initial findings were discussed. After the sessions were complete, we analysed the logfiles and evaluated the answers given by the participants, and grouped and summarized their comments and our observations. In combination with the recordings, the logfiles proved useful for finding similar events in other sessions, which had not been annotated earlier because their significance had not yet been known.

During the evaluation phase, no changes were made to the prototype. Additional problems encountered were added to the bug-tracking software and then solved by the development teams, but nothing was deployed to the evaluation system until after the evaluations reported in this paper were completed.

## 4 Findings

This section presents some of the key findings obtained in the evaluation, as well as their implications for further development and design of the search system.

### 4.1 Answers to the Task Questions

All tasks were phrased as yes/no or either/or questions, which makes the answers clearly right or wrong. Participants were asked to first give an answer from their knowledge before beginning the search. Table 2 compares the answers before and after the search. Of the 50 cases where participants completed a task, 14 (28%) ended with the participant not finding an answer, 17 (34%) reported a correct answer that was unknown before searching, 5 (10%) reported a wrong answer (in 1 (2%) of these cases the searcher had known the correct answer before searching but was persuaded to change their mind), and in 14 (28%) cases the answer remained the same, 12 right, 2 wrong. Overall, in 29 (58%) of the cases, searchers responded with the correct answer, and a wrong answer was reported 7 (14%) times.

**Table 2.** Answers given by participants before and after searching

Answer before/after search	Task 1	Task 2	Task 3	Task 4	Total
unknown → unknown	2	2		1	5
wrong → unknown	3				3
correct → unknown	3	3			6
unknown → correct	1	4	2	9	16
wrong → correct		1			1
unknown → wrong	1	1	2		4
correct → wrong			1		1
correct → correct	2	1	8	1	12
wrong → wrong	1	1			2
Total	13	13	13	11	50

## 4.2 Search

**Query Formulation.** Several users requested a kind of helping function that would support them with the choice of queries, and with the search process as a whole, as they struggled with finding the optimal balance between more general and more specific queries. The multi-level search possibilities with a fulltext search first, followed by various filtering options (search within results, classification by topic, source, type, etc.) left many users feeling unsure about their search strategy. For the scientific task, which was the last one, one user mentioned that the two text query fields (one for search and one for results filtering) finally made sense.

Many participants had problems deleting the old query before entering a new one. The query text field did not respond to clicks as expected (which would be two clicks to select a word, and three for the whole query), so these users proceeded to use the keyboard to delete every letter separately. There was an icon on the right end of the query text field for resetting the query, but these users apparently didn't expect such a functionality and did not ask.

*Implications for design:* Template-based structured queries could help users to express their information need in such a way that the search system can support their intentions. Such a feature is already planned and under development for the upcoming version of KHRESMOI.

**Image Search.** The data sources used in the prototype included images. One of the tasks required the participants to choose between two diagnoses for an x-ray image of a lung. Some participants expected to be able to use the image directly as a query, and to be able find similar images for a result.

## 4.3 Result Selection and Use

**Relevance Judgements.** The participants of this study were biased towards scientific journals. Only one user liked the idea of finding reliable webpages to recommend to her patients, whereas most other users found websites “not scientific enough” and sought to avoid them, even though it was known to them that only websites with a HONcode certification [10] are included in the KHRESMOI prototype.

There might be a difference in the views of general practitioners and specialists, but the participants of this round of evaluations do not allow us to make this distinction. Also, the comparatively young professional age of most participants might have produced a bias towards resources used at university.

**Number of Results.** The settings of the prototype restricted the number of result documents displayed to 200, since participants would not have time to look at more documents. Many participants believed that this meant the system had found only 200 documents for their query, and insisted that it must be so even when the researcher explained the setting and the reasoning behind it to them.



*Implications for design:* Display the total amount of results for a query, also to allow a distinction between those that return a few dozen, and those that return thousands of documents.

**Tools.** The task scenarios asked the participants to use various tools for the tasks, e.g., to store the results they selected as relevant for this task in the Personal Library of their account. The tools were still rather basic: Results could be temporarily put into a Tray, or permanently stored in a Personal Library, where they could be attributed with tags, or exported to various formats. Overall, the participants were quite eager to use these features, and made many suggestions for improvement, such as sorting in the Library, additional ways for accessing the stored items, or for sharing them with other people. The statement “I would find the personal library (tray) a helpful tool for my work.” received a median answer of “strongly agree” in the final feedback questionnaire.

*Implications for design:* These tools for managing search results need to work in every incarnation of the software, mobile, web or desktop client. For transfer between devices they require users to login to accounts.

**Classification and Filtering of Results.** Opinions on the classification and filtering options available for the result list were very divided. Several users achieved very fast, good search results by combining more general text queries with a click on the right classification subgroup. The classifications thus were mentioned by several users as a feature they liked and as a highlight of their search experience. Other users were not so lucky, and thus there are also comments that the filters do not work, should be reduced, removed, or offered as advanced option only by user request.

*Implications for design:* The classification and filtering options in the user interface need to be more self-explanatory. The classification box can be hidden when there are only a few results, and different backgrounds can be used to make the filters more distinguishable

#### 4.4 Issues Arising from Multilinguality

All participants had German as mother tongue. The user interface of the KHRESMOI prototype, the task descriptions and the questionnaires were all in English, as well as the sources in the search engine. The participants’ user accounts were set to German, to enable the translation feature into German.

The introduction given verbally by the researcher at the beginning of the tests was in German. All participants spoke German during the test, asking questions and providing feedback to the researcher. The written feedback in the questionnaires was given in English by some test participants and in German by others. Some participants had difficulties expressing their thoughts in written English and used dictionary websites to find the words they were looking for while filling out the feedback forms. Some popup windows from the test management software had the question text in English and the answer buttons in German. Overall,

the test participants seemed to be very familiar with and relaxed about such mixed language situations, where they read in English and discuss in German.

Users use more than one language in a query (and often are not aware of it). The system's design expects participants to have a language and to use that language in their interactions with the system. Therefore, queries are expected to be either completely in English, or completely in German, which would then be translated in a pre-processing step before submission to the retrieval system. But many queries typed by the participants consisted of several terms in different languages:

differentiating pneumonia from atelektasis xray

diabetes mellitus typ 2 elevated risk of cancer

The word 'atelektasis' is a mixture between 'Atelektase' (German) and 'atelectasis' (English), while 'typ' is the German word for 'type'. During the tests, the experimenter occasionally prompted participants to re-examine their spelling, but they did not see the mistakes even when pointed to them and chose to keep their spelling. Since many words only differ in one character between the languages, these might also be seen as spelling mistakes. Indeed, the spelling correction feature, which provides as-you-type suggestions, often reacted to these mistakes, but in many cases the participants ignored the hints.

The system was configured to expect only terms in one language and did not automatically correct the spelling, so these queries produced inferior results, i.e., no results were found for "atelektasis". Since users were not aware of their spelling, they believed that no more or better results existed.

*Implications for design:* Language detection needs to be done on single term level, and spelling correction suggestions adapted to the results. If one or more terms are not recognized, suggested alternatives should be offered together with the result list ("did you mean?"), and if no results are found at all, an alternative query can be executed immediately and offered as a suggestion to the user.

The machine translated result snippets in the result preview were used by many participants. Some participants used the translations only occasionally to check a single word or to confirm their understanding, but others needed more support. Since not all items could be translated, these participants were at a disadvantage, even if they were able to understand the task descriptions and questionnaires. The text in images could not be translated, yet some searches returned text-heavy images of charts or graphs, which contained useful summaries. Also, for websites, the result summary snippet could be translated, but the link to the whole page led to the original page in English.

*Implications for design:* When offering translation help, the feature needs to span the whole search process. For participants with below average English skills, the system should allow them to filter results by language and availability of translation, to be able to ignore results they cannot use.

## 4.5 Common Themes

**Data Sources.** The data sources accessible through the prototype were a major topic which most users commented on/asked about during the tasks. Overall, many users had doubts about the quantity of the data and repeatedly asked the facilitator whether certain sources were contained in the prototype and supposed to be findable. For some tasks, users knew that a specific journal exists for the topic and would have liked to prioritize results from that source. Several users stumbled over external pages where access was denied.

*Implications for design:* Provide a way for users to check the data sources. For sources that require a (paid) account for viewing the full text, users will need to login to an account. For others, provide a way to hide links to the full text when it is not available.

**Speed.** Even when the retrieval of results took a few seconds, this was perceived as normal by the participants. They found the search speed to be fast enough, but were instead very sensitive to small lags in the user interface. Non-responsive scrollbars or mouse-over tool-tips that remained visible for too long were perceived as very annoying and as obstacles to efficient work.

**Overall Satisfaction and Feedback.** Participants gave the system very good marks in the feedback questionnaire at the end of the test session. They found the system easy to use, and its functions helpful for their work, and they did not find it too complex or difficult to use.

This is contrary to our impressions during the sessions and while watching the recordings. There is research to show that people tend to use the higher (better) end of the scale when evaluating systems [15].

## 5 Conclusion and Future Work

Overall, this evaluation has been very successful in uncovering bugs and errors, and discovering requirements and desirable features for the next development steps. Our study design allowed us to identify which components are already effective in supporting users in their tasks, and which features need to be improved or added. It also allowed us to identify interesting characteristics of professional users working in a multilingual environment.

We tested and improved our study design, and revised the protocols and plans for future tests, including many valuable details such as which information needs to be logged, or how much time is needed to schedule test sessions with busy physicians.

As new features become available in the prototype, they will be tested on an ongoing basis, to allow an iterative development and improvement. We plan these future tests to distinguish between use cases for different groups of physicians, such as general practitioners vs. specialists, together with a widened range of tasks. Based on our experiences with the evaluations reported here, to achieve the goal of more frequent, more focused, and smaller test rounds we are planning small evaluation sessions at medical conferences, starting in autumn of 2013.

## References

1. Hersh, W.R., Hickam, D.H.: How well do physicians use electronic information retrieval systems? a framework for investigation and systematic review. *JAMA* 280(15), 1347–1352 (1998)
2. Hoogendam, A., Stalenhoef, A.F.H., de Vries Robbé, P.F., Overbeke, A.J.P.M.: Answers to Questions Posed During Daily Patient Care Are More Likely to Be Answered by UpToDate Than PubMed. *J. Med. Internet Res.* 10(4) (2008)
3. Ely, J.W., Osheroff, J.A., Maviglia, S.M., Rosenbaum, M.E.: Patient-care questions that physicians are unable to answer. *JAMA* 14, 407–414 (2007)
4. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, vol. 3(1-2) (2009)
5. Zhang, P., Plettenberg, L., Klavans, J.L., Oard, D.W., Soergel, D.: Task-based interaction with an integrated multilingual, multimedia information system: A formative evaluation. In: *JCDL 2007*, Vancouver, Canada, June 17-22. ACM (2007)
6. Beckers, T., Dungs, S., Fuhr, N., Jordan, M., Kriewel, S.: ezDL: An interactive search and evaluation system. In: *SIGIR 2012 Workshop on Open Source Information Retrieval*, OSIR 2012 (2012)
7. Cunningham, H., Tablan, V., Roberts, I., Greenwood, M.A., Aswani, N.: Information Extraction and Semantic Annotation for Multi-Paradigm Information Management. In: Lupu, M., Mayer, K., Tait, J., Trippe, A.J. (eds.) *Current Challenges in Patent Information Retrieval*. The Information Retrieval Series, vol. 29. Springer (2011)
8. Garcia Seco de Herrera, A., Markonis, D., Eggel, I., Müller, H.: The medGIFT Group in ImageCLEFmed 2012. In: *CLEF (Online Working Notes/Labs/Workshop)* (2012)
9. Kiryakov, A., Ognyanov, D., Manov, D.: OWLIM – A Pragmatic Semantic Repository for OWL. In: Dean, M., Guo, Y., Jun, W., Kaschek, R., Krishnaswamy, S., Pan, Z., Sheng, Q.Z. (eds.) *WISE 2005 Workshops*. LNCS, vol. 3807, pp. 182–192. Springer, Heidelberg (2005)
10. Health On the Net Foundation: The HON Code of Conduct for medical and health Web sites, HONcode (2013), <http://www.hon.ch/HONcode/Patients/Conduct.html>
11. Kritza, M., Gschwandtner, M., Stefanov, V., Hanbury, A., Samwald, M.: Utilization and perceived problems of online medical resources and search tools among different groups of European physicians. *J. Med. Internet Res.* (forthcoming, 2013)
12. Borlund, P.: The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research* 8(3), 152 (2003)
13. Brooke, J.: SUS: A “quick and dirty” usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) *Usability Evaluation in Industry*. Taylor and Francis, London (1996)
14. TechSmith Corporation: Morae usability testing software, version 3.3.2 (2013), <http://www.techsmith.com/morae.html>
15. Hornbaek, K., Law, E.L.C.: Meta-analysis of correlations among usability measures. In: *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 617–626 (2007)

# Exploiting Multiple Translation Resources for English-Persian Cross Language Information Retrieval

Hosein Azarzonyad, Azadeh Shakery, and Hesham Faili

School of Electrical and Computer Engineering,  
College of Engineering,  
University of Tehran, Tehran, Iran  
{h.azarzonyad, shakery, hfaili}@ut.ac.ir

**Abstract.** One of the most important issues in Cross Language Information Retrieval (CLIR) which affects the performance of CLIR systems is how to exploit available translation resources. This issue can be more challenging when dealing with a language that lacks appropriate translation resources. Another factor that affects the performance of a CLIR system is the degree of ambiguity of query words. In this paper, we propose to combine different translation resources for CLIR. We also propose two different methods that exploit phrases in the query translation process to solve the problem of ambiguousness of query words. Our evaluation results on English-Persian CLIR show the superiority of phrase based and combinational translation CLIR methods over other CLIR methods.

**Keywords:** Cross Language Information Retrieval, English-Persian CLIR, Phrase Based Query Translation, Combining Translation Resources for CLIR.

## 1 Introduction

Cross Language Information Retrieval (CLIR) deals with the problem of expressing queries in one language (source language) and retrieving the related documents in another language (target language). The most important problem in CLIR is the difference between source and target languages, which makes it impossible to directly match queries and documents. Query translation is the most common approach to solve this problem. This paper focuses on query translation for English-Persian CLIR.

Different types of resources can be used for query translation such as bilingual dictionaries, machine translators, parallel corpora, and comparable corpora[6]. Because of limitations and high cost of machine translators, corpora based and dictionary based methods have been widely used in CLIR[6]. Query translation using each resource has its advantages and disadvantages. Translations that are extracted from parallel corpora are usually more accurate than those extracted from other translation resources. However, parallel corpora are expensive resources which may not be available in all language pairs. Unlike parallel

corpora, which are clearly defined as translated texts, comparable corpora are composed of aligned related documents that are not exact translations of each other. Therefore, translations that are mined from comparable corpora may contain translation candidates that are only related to query terms and are not their direct translations. This can guide the translation process to query expansion. However, as aligned documents in comparable corpora are not exact translations of each other, they could be very noisy in order to be employed in query translation. Dictionaries have also some benefits and deficiencies for query translation. Dictionaries are available for many language pairs and their usage is simple, but they have a main shortcomings: limited coverage. For example dictionaries usually do not contain proper nouns.

In order to take advantage of all the translation resources and reduce their shortcomings, we propose to combine different resources. For example, by combining comparable and parallel corpora, we can benefit from accurate translations of query words extracted from parallel corpora, as well as related words to the query extracted from comparable corpora. In this paper, we exploit different translation resources simultaneously for translating queries. We assign different weights to translations extracted from different translation resources and linearly combine the translations to obtain a more accurate translation for the query.

Also, in order to increase the accuracy of CLIR, we propose two different methods for exploiting phrases extracted from parallel corpora in query translation process: Phrase Indexing method and Phrase Based Translation Re-ranking method. Phrase Indexing method considers each phrase as a term and indexes phrases as well as single terms. Phrases are then used for query translation. Phrase Based Translation Re-ranking method considers phrases as bags of words and uses the phrase translation probabilities extracted from the parallel corpus for re-ranking translations of query words.

To evaluate the proposed methods, we use the Hamshahri collection[1] and do the CLIR task of CLEF-2008 and CLEF-2009: Retrieving Persian documents in response to queries in English. Also, we use TEP parallel corpus[8], UTPECC comparable corpus[4], and Arianpour bilingual English-Persian dictionary<sup>1</sup> for implementing different CLIR methods. Our results show that the proposed phrase based CLIR methods outperforms all of single resource based methods. Also, the combinational method outperforms single resource based method as well as phrase based CLIR methods.

The remainder of the paper is organized as follows: In Section 2, we describe the methods using single translation resources for CLIR, phrase based translation methods, and an intuitive method for combining translation resources. Section 3 explains the experiments for English-Persian CLIR, based on different translation resources. Finally, Section 4 concludes the paper with a brief discussion on the impact of different translation resources for Persian-English CLIR.

---

<sup>1</sup> <http://www.aryanpour.com/>

## 2 Query Translation Approaches

In this research, we use parallel corpora, comparable corpora, and dictionary for query translation. In the rest of this section, we present details of the different CLIR methods using different translation resources.

Parallel corpora are valuable resources for obtaining translation knowledge. Among different methods of employing parallel corpora for extracting translation relations between source and target language words, IBM model-1[2] is the most used method. We use the method proposed in [7] for employing the translation probabilities obtained using IBM model-1 and constructing queries in the target language. Using this method, we select the top N translation candidates which have the highest translation probabilities among the candidate translations of all query words as the translated query. The translation probability of a given target language word  $f$  for the source language query  $Q_e$  could be calculated as follows:

$$P(f|Q_e) = \sum_{e \in Q_e} p(f|e) * P(e|Q_e), \quad (1)$$

where  $p(f|e)$  is the probability of translating source language word  $e$  to target language word  $f$  and is extracted using IBM model-1 from the parallel corpus. We consider  $P(e|Q_e)$  to be uniform regarding to all query words. Using Okapi BM25 method, we score the documents in response to the constructed query and rank the results.

One of the main drawbacks of IBM model-1 is that it considers the words to be independent of each other when extracting translations. This assumption is not realistic. In this paper, to consider the relations of words in query translation process, we use phrases extracted from parallel corpora. We use Koehn method[5] for extracting the probabilistic phrase table. We have exploited phrases in two different ways:

**Phrase indexing:** This method has three main steps:

1. Indexing step: In this step, we find noun phrases in the target language. Then, in document indexing, we index phrases in target language as well as unique words.
2. Query translation step: In this step, queries are considered as combinations of a number of units, where a unit could be a unique word or a phrase. We consider each combination of consecutive query words as well as unique words as query units. To translate a query, we first find noun phrases in the query. Phrases in a query could be translated to a number of phrases in the target language. To construct a query in the target language, we use top N translation candidates that could be words or phrases with the highest translation probabilities for the whole query, which are achieved by Equation 1.
3. Retrieval step: Finally, in the retrieval step, we employ the translated query to calculate BM25 score of documents in response to the query. The translated queries contain phrases and words which are indexed by our method in the indexing step. So, using the index, we can easily calculate the BM25 scores of documents for each query.

**Phrase Based Translation Re-ranking:** This method uses phrases for re-weighting the translation probabilities extracted using IBM model-1 and considering the phrasal weight of query words in query translation. This method has two main steps:

1. Candidate translation extraction step: In this step, we use Equation 1 to calculate the translation probabilities of target language words regarding the query.
2. Re-weighting and re-ranking step: In this step, we detect phrases in queries and translate them using the phrase table. After translating phrases, we consider the candidate phrases as bags of words and use them to re-rank candidate translations of query terms. Phrases in the phrase table have translation candidates with translation probabilities. We calculate phrasal score of a target language word  $f$  as follow:

$$S_{ph}(f, Q_e) = \frac{\sum_{\substack{ph \in Q_e \\ f \in ph}} P(ph|Q_e)}{\sum_{v \in Can} \sum_{\substack{ph \in Q_e \\ v \in ph}} P(ph|Q_e)}, \quad (2)$$

where,  $ph$  and  $P(ph|Q_e)$  are a candidate phrase and the translation probability of the candidate phrase to the source language query respectively and  $Can$  is the set of translation candidate words that are extracted in step 1. This Equation gives a high phrasal score to the words that are contained in more phrases. After calculating the score of words with regard to phrases, we add these scores to the translation probabilities of words, which is calculated by Equation 1 and then we re-rank the translation candidates using the new scores.

We can use comparable corpora for extracting translations of query words. We use the method described in [9] for extracting translation knowledge from comparable corpora. The output of this method is a lexicon in which for each source language word, its translation candidates and their correlation scores are specified. We use the method described in [3] for transforming these scores to translation probabilities. Also, we use the method proposed in [3] for selecting the number of translations for each query word.

An intuitive method for combining translations extracted from different resources is to directly merge translation probabilities before the translation step. To do so, we assign a weight to each translation resource regarding its accuracy and use the weighted sum of translation probabilities for calculating the final translation probabilities. We calculate the probability of translation of a target language word  $f$  to a source language word  $e$  as:

$$P(f|e) = \lambda * P_{R1}(f|e) + (1 - \lambda) * P_{R2}(f|e), \quad (3)$$

where  $R1$  and  $R2 \in \{\text{Parallel corpus, Comparable corpus, Dictionary}\}$  and  $R1 \neq R2$  and  $\lambda$  is a parameter which controls the effect of  $R1$  and  $R2$  in translation. After calculating these probabilities, again we use Equation 1 for selecting the top  $N$  translation candidates and translating query. We use the methods described in this section for extracting translation probabilities from parallel and comparable



corpora. Since dictionaries do not contain translation probabilities, if a source language word has  $M$  candidate translations in the dictionary, we consider the translation probability of each candidate to be  $1/M$ .

### 3 Experimental Results

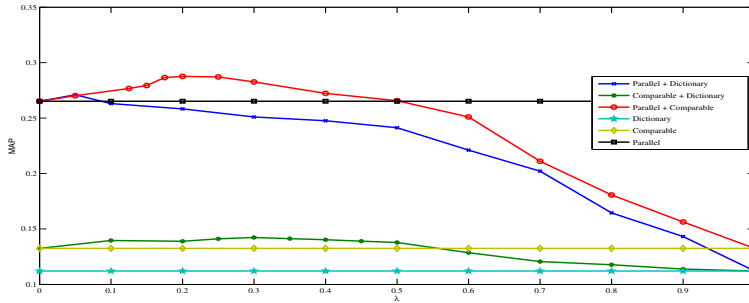
For evaluating different CLIR methods, we do the CLIR task of CLEF-2008 and CLEF-2009: Retrieving Persian documents in response to queries in English. We use Hamshahri[1] dataset, which contains about 166,000 documents in Persian and 100 queries in English and Persian. We employ TEP[8] parallel corpus for extracting translations of query terms. Also, we use this corpus for extracting phrases used for implementing phrase based CLIR methods. For implementing comparable corpora based CLIR method, we use UTPECC[4] comparable corpus. Also, we employ Arianpour dictionary for implementing dictionary based CLIR method.

The results of different CLIR methods are shown in Table 1. For the monolingual experiment, we use the Persian queries for retrieving Persian documents. As can be seen from Table 1, parallel corpus based CLIR has the best performance among single resource based CLIR methods. The best result of parallel corpus based method is achieved when we set  $N = 20$  (the size of constructed target language query). Also, in dictionary based method, we use top  $N$  translations of each query word for constructing query in the target language. The best result is achieved when we set  $N = 6$ . Thus, in Table 1, we only report the results of dictionary based method for  $N = 6$ . From Table 1 we can see that comparable corpus does better than dictionary in translating queries.

The results of phrase based query translation methods are also shown in Table 1. As can be seen from this table, phrase based methods outperform the pure parallel corpora based method. In addition, Phrase Based Translation Re-ranking method has better performance in CLIR. Our experiments show that in cases where queries contain phrases, the proposed phrase based query translation methods significantly outperform other parallel corpora based methods. We conducted statistical significant test (t-test) on the improvements of the best performing phrase based method over pure parallel corpora based method. Our results show that the improvements of this method in terms of MAP (Mean Average Precision) for queries that contain phrases are statistically significant.

**Table 1.** CLIR results using different translation resources

Method	MAP	%Mono	P@5	%Mono	P@10	%Mono
Monolingual IR	0.4126	-	0.702	-	0.643	-
Dictionary Based CLIR	0.1383	34	0.216	31	0.204	32
Comparable Corpus Based CLIR	0.1485	36	0.288	40	0.266	41
Parallel Corpus Based CLIR	0.2652	64	0.44	62	0.418	65
Phrase Indexing	0.2721	66	0.446	64	0.429	67
Phrase Based Translation Re-ranking	0.2813	68	0.452	64	0.431	67



**Fig. 1.** Performance of combining different translation resources for different values of  $\lambda$  in terms of MAP

In Figure 1, the results of combining different translation resources are shown. In these experiments we set  $N = 25$ , where  $N$  is the number of translation words obtained by Equation 3. We use different resources for extracting translations using Equation 3. We test our method for the different values of  $\lambda$ . As can be seen from Figure 1, the best result of combinational method is achieved from combination of parallel and comparable corpora. In this experiment, we consider R1 as comparable corpus and R2 as parallel corpus. We achieved the best performance when  $\lambda = 0.2$  and the MAP for this  $\lambda$  is 0.2832. When the value of  $\lambda$  is less than 0.2, MAP is less than 0.2832, but increasing  $\lambda$  leads to improvement in MAP. In fact, for the low values of  $\lambda$ , comparable corpus cannot influence the translation process, but by increasing  $\lambda$  comparable corpus is affecting the translations and due to query expansion occurred by comparable corpus translations, the performance is increasing. When the value of  $\lambda$  is becoming larger than 0.2, the performance decreases. In other words, for big values of  $\lambda$  the influence of comparable corpus in translation is increased, which increases the noise in translations. Also, from Figure 1, we can see that using other combinations of resources also outperforms single resource based CLIR methods.

## 4 Conclusion and Future Works

In this paper, we studied the effect of different translation resources in English-Persian CLIR. We employed different translation resources to translate English queries to Persian. Also, we proposed two context based methods that employ phrases extracted from parallel corpora for query translation. We examined these methods in English-Persian CLIR and realized that these methods outperform other English-Persian CLIR methods. Our results showed that phrase based CLIR method improves the MAP of basic parallel corpus based method by 6%. Furthermore, we proposed a combinational method to combine translations that are mined from different resources. We examined the proposed method by combining different resources. Our results showed that the combinational translation method outperforms single resource based translation method.

The best results of combinational method is achieved when we combine parallel and comparable corpora. Our results showed that this method improves comparable corpus based method by 106% and parallel corpus based method by 7%.

In our future work we are going to improve the proposed CLIR method by employing more, higher quality translation resources. It will also be interesting to exploit phrases and other contextual information like mutual information of translation candidates in query translation process.

## References

1. AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., Oroumchian, F.: Hamshahri: A standard persian text collection. *Know.-Based Syst.* 22(5), 382–387 (2009)
2. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.* 19(2), 263–311 (1993)
3. Hashemi, H.B.: Using Comparable Corpora for English-Persian Cross-Language Information Retrieval. Master's thesis, University of Tehran, Tehran, Iran (2011)
4. Baradaran Hashemi, H., Shakery, A., Faili, H.: Creating a persian-english comparable corpus. In: Agosti, M., Ferro, N., Peters, C., de Rijke, M., Smeaton, A. (eds.) *CLEF 2010. LNCS*, vol. 6360, pp. 27–39. Springer, Heidelberg (2010)
5. Koehn, P.: Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In: Frederking, R.E., Taylor, K.B. (eds.) *AMTA 2004. LNCS (LNAI)*, vol. 3265, pp. 115–124. Springer, Heidelberg (2004)
6. Nie, J.Y.: *Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies.* Morgan & Claypool Publishers (2010)
7. Nie, J.Y., Isabelle, P., Plamondon, P., Foster, G.: Using a probabilistic translation model for cross-language information retrieval. In: 6th Workshop on Very Large Corpora, pp. 18–27 (1998)
8. Pilevar, M.T., Faili, H., Pilevar, A.H.: TEP: Tehran english-persian parallel corpus. In: Gelbukh, A. (ed.) *CICLing 2011, Part II. LNCS*, vol. 6609, pp. 68–79. Springer, Heidelberg (2011)
9. Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., Laurikkala, J.: Focused web crawling in the acquisition of comparable corpora. *Inf. Retr.* 11(5), 427–445 (2008)

# ALQASIM: Arabic Language Question Answer Selection in Machines

Ahmed Magdy Ezzeldin, Mohamed Hamed Kholief, and Yasser El-Sonbaty

College of Computing and Information Technology, AASTMT  
Alexandria, Egypt  
a.magdy@alworks.com, {kholief,yasser}@aast.edu

**Abstract.** This paper presents “ALQASIM”, a question answering system that focuses on answer selection and validation. Our experiments have been conducted in the framework of the main task of QA4MRE @ CLEF 2013. ALQASIM uses a novel technique by analyzing the reading test documents instead of the questions, which leads to a promising performance of 0.31 accuracy and 0.36 C@1, without using the test-set background collections.

**Keywords:** Question Answering, QA4MRE, Machine Reading Evaluation, Answer Selection, Answer Validation.

## 1 Introduction

“ALQASIM” is a Question Answer (QA) selection and validation system that aims at answering the multiple choice questions of QA4MRE @ CLEF 2013 test-set. It could also be used as a part of the answer validation module of any ordinary Arabic QA system. In the upcoming sections, the related works, system architecture, evaluation and discussion and the future work of ALQASIM are demonstrated.

## 2 Related Works

In CLEF 2012, Arabic QA4MRE was introduced for the first time. Two Arabic systems participated in this campaign. The first system, IDRAAQ [1], achieved a 0.13 accuracy and a 0.21 c@1. It used the Distance Density N-gram Model and semantic expansion using Arabic WordNet, and did not use the CLEF background collections.

The second system by Trigui et al. [6] achieved the accuracy and c@1 of 0.19 with their system. They used semantic expansion using inference rules on the background collection. They also determined the question focus and aligned the retrieved passages with the multiple answer choices of the question. However, these systems do not compare to the system created by Bhaskar et al. [2] for English QA4MRE @ CLEF 2012 that has an accuracy of 0.53 and c@1 of 0.65.

### 3 ALQASIM Architecture

Most QA systems are composed of three main phases, which are: Question Analysis, Passage Retrieval and Answer Extraction. However, these systems are mainly targeted at searching for answers in a large collection of documents or on the Internet, which makes passage retrieval efficient [3]. QA4MRE is different in that aspect because the answer to a question is found in only one document, so there is not enough information redundancy to help the IR statistical approaches of passage retrieval. Thus, the ordinary QA pipeline is not the best approach to QA4MRE; the best approach is the one used by human beings in reading tests. A person would normally read and understand a document thoroughly, and then begins to tackle the questions. So, the suggested approach divides the QA4MRE process into three phases: (i) Document Analysis, (ii) Locating Questions & Answers, and (iii) Answer Selection.

#### 3.1 Document Analysis

In the Document Analysis phase, the reading test documents are analyzed using MADA+TOKAN [3] morphological analyzer to stem each word in the documents and get its Part-of-Speech (PoS). Then, stop words are removed, and an inverted index of the remaining words stems is created, which contains the locations of each stem and its weight. Arabic WordNet (AWN) is then used to expand the words semantically by adding the synonyms of each word to the inverted index of that document. The weight of each word in the inverted index is assigned according to its PoS and repetition. So, nouns, verbs, adjectives, adverbs, proper nouns and the other parts of speech are assigned different weights. These weights mark word importance and are assigned according to our experiments with QA4MRE @ CLEF 2013 questions, by assigning the weights that yield the best results. Then the weight of a word is divided by its count in the document, thus, the more a word is repeated the less its weight will be. Thus, if the word is repeated many times in the target document, it is less likely to mark a question/answer snippet, because it appears in many sentences.

$$K_i = \frac{W_i \times S}{C_i} \quad (1)$$

$K_i$  : the Weight of the word ( $i$ ) saved in the inverted index

$W_i$  : the weight of the word ( $i$ )

$S$  : the synonym multiplier if the word is semantically expanded using AWN

$C_i$  : the number of repetitions of the word ( $i$ ) in the document

#### 3.2 Locating Questions and Answers

In the second phase, every question and answer choice is handled as follows. Keywords are identified by stemming and removing stop words. The inverted index is then searched to find the best scoring three snippets locations for each question and

answer choice keywords. This score is calculated according to: (i) the number of keywords found within a distance threshold, (ii) the weights of all found keywords and (iii) the distance between these keywords. The impact of keywords count and weights is positive while the impact of distance is negative which means that snippets locations scores are penalized for higher distance among its keywords.

$$S_n = \left( \sum_{i=1}^N K_i \right) + N - \left( \sum_{i=2}^N d_i \right) \quad (2)$$

$S_n$  : the score of snippet ( $n$ ) which is found keywords for a question or answer choice.

$N$  : the number of found keywords for the snippet

$K_i$  : the weight of the keyword ( $i$ ) as found in the inverted index. See equation 1.

$d_i$  : the distance between the found keywords ( $i$ ) and ( $i-1$ )

### 3.3 Answer Selection

By now, the question and its five answer choices have three scored snippets locations each. In this phase, answer choices snippets locations are scored by summing the scores of one question location and one answer choice location and subtracting the distance between them. The maximum of these scores is selected as the answer choice score. The best scoring answer choice is then selected as the question answer. If there is more than one best scoring answer choice, the question is marked as unanswered.

$$\text{Score}_{nA_i} = QS_n + A_iS_n - D_{QS_nA_iS_n} \quad (3)$$

$Score_{nA_i}$  : the score of one question snippet with one answer choice snippet.

$D_{QS_nA_iS_n}$  : the distance between the question snippet and the answer choice location.

$$\text{Score}A_i = \max \left( \text{Score}_{nA_i} \right) \quad (4)$$

$ScoreA_i$  : the maximum score of all Answer Choice ( $i$ ) snippets.

$Score_{nA_i}$  : the score of one question snippet with one answer choice snippet.

## 4 Evaluation and Discussion

The test-set, used by ALQASIM, is the set of questions and answers provided by CLEF 2011 [5] and translated to Arabic in 2012. ALQASIM uses Accuracy and C@1 [4] as evaluation metrics. It performs at an Accuracy of 0.31 and a C@1 of 0.36, which is considered promising, as it did not use any background collections.

Our system performs better than the other two Arabic QA4MRE systems from CLEF 2012 mainly because it analyses the reading test documents instead of the questions and answers. Documents have much more words than questions and answers, which gives context for morphological analyzers to produce more accurate analyses.

This explains why ALQASIM performs better on questions and answers with more keywords. On the other hand, many incorrectly answered questions are causative and list questions and questions that were incorrectly translated due to erroneous automatic translation. It is also noticed that sometimes the correct answer choice has fewer keywords than the other choices, which misleads the system into selecting an incorrect answer choice with more keywords, thus higher weight.

**Table 1.** Performance of ALQASIM and QA4MRE systems

	<b>Accuracy</b>	<b>C@1</b>
IDRAAQ [1]	0.13	0.21
Trigui et al. [6]	0.19	0.19
Bhaskar et al. [2]	0.53	0.65
<b>ALQASIM</b>	<b>0.31</b>	<b>0.36</b>

## 5 Conclusion and Future Work

This paper presents “ALQASIM” a Question Answer Selection and Validation system that can answer the multiple choice questions of QA4MRE @ CLEF 2013 test-set with an accuracy of 0.31 and a C@1 of 0.36. We are currently working on integrating Named Entity Recognition (NER), anaphora resolution, and temporal inference. We are also working on handling cause/effect relationship, and building an ontology from the background collections to expand questions and answers keywords.

## References

1. Abouenour, L., Bouzoubaa, K., Rosso, P.: IDRAAQ: New Arabic Question Answering System Based on Query Expansion and Passage Retrieval. In: CLEF 2012 Workshop on Question Answering For Machine Reading Evaluation, QA4MRE (September 2012)
2. Bhaskar, P., Pakray, P., Banerjee, S., Banerjee, S., Bandyopadhyay, S., Gelbukh, A.: Question Answering System for QA4MRE@CLEF 2012. In: CLEF 2012 Workshop on Question Answering For Machine Reading Evaluation, QA4MRE (September 2012)
3. Habash, N., Rambow, O., Roth, R.: MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, PoS Tagging, Stemming and Lemmatization. In: Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, pp. 102–109 (2009)
4. Peñas, A., Rodrigo, A., del Rosal, J.: A simple Measure to Assess Non-response. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 1415–1424 (2011)
5. Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Forascu, C., Sporleder, C.: Overview of QA4MRE at CLEF 2011: Question answering for machine reading evaluation. Working Notes of CLEF (2011)
6. Trigui, O., Belguith, L.H., Rosso, P., Amor, H.B., Gafsaoui, B.: Arabic QA4MRE at CLEF 2012: Arabic Question Answering for Machine Reading Evaluation. In: CLEF 2012 Workshop on Question Answering For Machine Reading Evaluation, QA4MRE (September 2012)

# A Web-Based CLIR System with Cross-Lingual Topical Pseudo Relevance Feedback

Xuwen Wang<sup>1</sup>, Xiaojie Wang<sup>1</sup>, and Qiang Zhang<sup>2</sup>

<sup>1</sup> Beijing University of Posts and Telecommunications, Beijing, China  
xw.y.wang@gmail.com, xjwang@bupt.edu.cn

<sup>2</sup> State Grid Electric Power Research Institute, Beijing, China  
zhangqiang7@sgepri.sgcc.com.cn

**Abstract.** This paper presents the performance of a Chinese-English cross-language information retrieval (CLIR) system, which is equipped with topic-based pseudo relevance feedback. The web-based workflow simulates the real multilingual retrieval environment, and the feedback mechanism improves retrieval results automatically without putting excessive burden on users.

**Keywords:** Pseudo Relevance Feedback, Latent Dirichlet Allocation, Cross Language Information Retrieval, Query Expansion.

## 1 Introduction

It is important to find a semantically consistent translation for the source language query in CLIR tasks. One convenient method for optimizing query translation is pseudo relevance feedback (PRF), which has been widely applied to query expansion in monolingual IR tasks. Traditionally, the cross-lingual PRF is performed on the basis of top-ranked retrieval results before or after the query translation step. The relativity of expansion terms are commonly calculated on the document-level. However, it is necessary to perform PRF on a fine-grained text, since a document may include multiple topics, but only a part of them relate to the user's query.

Topic modeling techniques such as LDA (Latent Dirichlet Allocation) represent a document as a mixture of multinomial distributions with Dirichlet priors [1]. LDA-based PRF models were proposed and applied to monolingual IR as well as CLIR tasks, and outperformed document-based PRF methods [2, 3]. An extension of the standard LDA model was polylingual LDA, which has been presented in many research works on utilizing multilingual information [4-6]. Under the assumption that bilingual retrieval results of the source language query and its translation share relevant topics, we further developed a cross-lingual PRF model based on bilingual topics [7].

This paper introduces a web-based Chinese-English CLIR system integrated with the PRF function. Section 2 describes the system framework and details the cross-lingual topical PRF model; section 3 shows the performance of cross-lingual query expansion experiments based on different PRF methods; section 4 outlines the conclusions and further work.



## 2 Design of PRF-Based CLIR system

The system framework is shown in figure 1, with additional PRF-based query expansion and web corpus mining mechanism. Firstly, we translate the original user queries into target language and perform CLIR on web-derived multilingual corpora. Secondly, the retrieved bilingual documents are analyzed by the cross-lingual topical PRF module, then bilingual relevant terms are selected for query expansion. Finally, the second retrieval process is performed based on new queries.

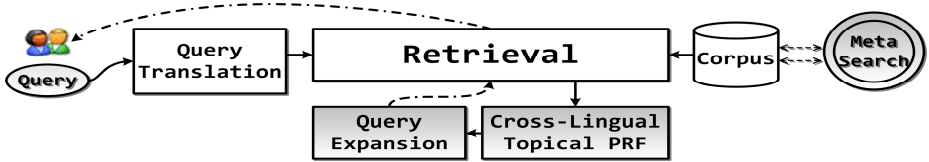


Fig. 1. Functionality frame of CLIR system with topical cross-lingual PRF

**Web-Based Multilingual Corpora.** Multilingual corpora were constructed by crawling webpages through a meta search engine which employed multiple search APIs. A self-constructed spider was applied to grab webpages in case the API access was unavailable. The crawled webpages were processed into plain texts via webpage cleaning, Chinese segmentation, English stemming, stop words removing, etc.

**Query Translation.** Source language queries were translated into target language based on a Chinese-English machine readable dictionaries (MRD). If a query was beyond the coverage of local MRD, we acquired its translation by accessing the Microsoft Translator Service (MTS), so as to alleviate the OOV problem.

**Retrieval Model.** The Indri toolkit (version 5.2) was employed as our fast indexing and retrieval model [8]. A local index base was built by the Indri indexer on the basis of web-derived multilingual corpora, and the retrieval process was executed on local index library for the submitted queries.

**Cross-Lingual PRF and Query Expansion.** Multilingual documents retrieved for a specific query and its translation were assumed to share the same topic space. A topic could generate source language terms as well as target language terms, but the distribution of different language terms under the same topic could be different. This situation was modeled by the bilingual LDA, see the graph model in figure 2. We applied Gibbs Sampling method for model inference and estimated the document-topic distribution  $\theta$  along with the topic-word distributions  $\phi_s$  and  $\phi_t$ .

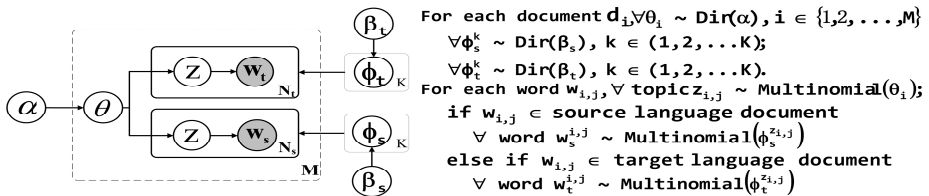


Fig. 2. Graph model of bilingual LDA

Two sorts of topics were used for expanding the original query and its translation. One was the most related topic that generate query terms with highest probability; the other topic represented the bilingual constraints which contained all the source language query terms and target language query terms, see equation 1. Expansion terms were filtered according to the word distribution under the selected relevant topics, usually pick top 20 to 40 terms for query expansion.

$$Z_{R-CTP} = z_{\max} \cup z_{bilingual} = (\operatorname{argmax}_z \sum_{i=1}^m p(q_i | z)) \cup (\bigcup_z p(Q_s, Q_t | z)) \quad (1)$$

### 3 Experiment Results

#### 3.1 Data

The Chinese query set included 54 Chinese terms chosen from the CNKI science and technology concepts, each contain 1 to 3 tokens. The English queries were translated from Chinese query terms via query translation model. Table 1 lists the scale of polysemy and ambiguity in the query set. Bilingual documents were collected from the real-time CLIR results of Google search API. Top 10 source language pages were collected for each Chinese query, since most web users were only concerned with top-ranked retrieval results. The cross-lingual retrieval results based on **Google’s query translation** were adopted as target language documents. Totally 540 Chinese pages and 540 English pages were collected. The hyper parameters were set to  $\alpha = 0.1$ ,  $\beta_s = 0.01$ ,  $\beta_t = 0.02$ . We used nDCG to evaluate the ranking effect of retrieval results. 27 bilingual language speakers were invited to judge the relevance of retrieval results.

**Table 1.** Polysemy and ambiguity of query set

Query Set	Unambiguous	Polysemy	Multiple translation	Translation ambiguity
Query	51	3	21	11
Query Translation	36	18	(38.9%)	(20.4%)

#### 3.2 Results

The ranking effect reflected the slight distinction between different methods on limited data. Compared with the general CLIR system without PRF, the Cross-lingual Topical PRF (CTP)-based CLIR showed significant performance (Two-Tailed Paired Samples T-Test,  $p = 0.0239 < 0.05$ ), see table 2. Both of the LDA-based PRF and VSM-based PRF were applied as monolingual expansion in different stage of CLIR, such as pre-translation PRF, post-translation PRF, and combined-translation PRF. The topical PRF strategies performed better than the VSM-based ones. The CTP-based CLIR has outperformed other PRF methods, indicating the usefulness of bilingual feedback information in cross-lingual query expansion. However, this observation

**Table 2.** Comparison of CLIR performances with different PRF strategies

CLIR	no-PRF	CTP	VSM-pre	VSM-post	VSM-comb	LDA-pre	LDA-post	LDA-comb
nDCG	0.8926	<b>0.9146</b>	<b>0.8930</b>	<b>0.8768</b>	<b>0.8744</b>	<b>0.8932</b>	<b>0.8959</b>	<b>0.9031</b>

does not hold for all the situation. For one thing, queries sharing the same bilingual topic with their translations account for only 60% cases in our experiment, while the rest of queries were average in performance. For another, the common  $\theta$  implies parallelism between bilingual texts, which does not always fit the real webpages.

## 4 Conclusion

This paper introduces a CLIR system which is equipped with cross-lingual topical PRF function on the basis of a small scale of bilingual retrieval results. The PRF mechanism is convenient to be modified and integrated with any retrieval model. Although it is currently a prototype system, there would be further work on mining aligned topics for cross-lingual query expansion. More experiments on larger corpora will also be conducted and discussed in the future.

## References

1. Blei, D.M., Ng, A.Y., Jordan, M.J.: Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 993–1022 (2003)
2. Ye, Z., Huang, X., Lin, H.: Finding a good query-related topic for boosting pseudo relevance feedback. *Journal of the American Society for Information Science and Technology* 62(4), 748–760 (2011)
3. Wang, X., Zhang, Q., Wang, X., et al.: LDA based pseudo relevance feedback for cross language information retrieval. In: *IEEE International Conference on Cloud Computing and Intelligence Systems*, Hangzhou, pp. 1993–1998 (2012)
4. Mimno, D., Wallach, H.M., Naradowsky, J., et al.: Polylingual topic models. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pp. 880–889. ACL, Singapore (2009)
5. Vulić, I., Smet, W.D., Moens, M.: Identifying word translations from comparable corpora using latent topic models. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, pp. 479–484 (2011)
6. Ganguly, D., Leveling, J., Jones, G.J.F.: Cross-Lingual Topical Relevance Models. In: *COLING* (2012)
7. Wang, X., Wang, X., Sun, Y.: Cross Language Pseudo Relevance Feedback Based on Bilingual Topics. *Journal of Beijing University of Posts and Telecommunications* 36(4) (2013)
8. <http://www.lemurproject.org/indri.php>

# A Case Study in Decompounding for Bengali Information Retrieval

Debasis Ganguly, Johannes Leveling, and Gareth J.F. Jones

CNGL, School of Computing, Dublin City University, Dublin 9, Ireland  
{[dganguly](mailto:dganguly@computing.dcu.ie), [jleveling](mailto:jleveling@computing.dcu.ie), [gjones](mailto:gjones@computing.dcu.ie)}@computing.dcu.ie

**Abstract.** Decompounding has been found to improve information retrieval (IR) effectiveness for compounding languages such as Dutch, German, or Finnish. No previous studies, however, exist on the effect of decomposition of compounds in IR for Indian languages. In this case study, we investigate the effect of decompounding for Bengali, a highly agglutinative Indian language. The standard approach of decompounding for IR, i.e. indexing compound parts (constituents) in addition to compound words, has proven beneficial for European languages. Our experiments reported in this paper show that such a standard approach does not work particularly well for Bengali IR. Some unique characteristics of Bengali compounds are: i) only one compound constituent may be a valid word in contrast to the stricter requirement of both being so; and ii) the first character of the right constituent can be modified by the rules of *Sandhi* in contrast to simple concatenation. As a solution, we firstly propose a more *relaxed decompounding* where a compound word is decomposed into only one constituent if the other constituent is not a valid word, and secondly we perform *selective decompounding* by ensuring that constituents often co-occur with the compound word, which indicates how related the constituents and the compound are. We perform experiments on Bengali ad-hoc IR collections from FIRE 2008 to 2012. Our experiments show that both the relaxed decomposition and the co-occurrence-based constituent selection proves more effective than the standard frequency-based decomposition method, improving mean average precision (MAP) up to 2.72% and recall up to 1.8%, compared to not decompounding words.

## 1 Introduction

Vocabulary mismatch between a query and the documents in a collection is an inherent problem in information retrieval (IR), as a result of which relevant documents comprising words different to query words may be retrieved at low ranks (or not at all). In many languages, word compounding is one of the main reasons for such vocabulary mismatch problems. To illustrate with an example, if a query comprises of the term *land*, a document predominantly containing the term *farmland* may be retrieved at a lower rank, than a document containing the terms *farming* and *land*. Decomposition or decompounding of the word *farmland*

in a document into the constituents *farm* and *land* can potentially result in more hits with the query and hence improve its ranking.

Compound splitting has become a standard preprocessing step for compounding languages such as Finnish, Dutch or German, where decomposition typically increases IR effectiveness [1–3]. While the effect of decompounding has been well researched for a number of European languages, there has been comparatively less IR research on the decompounding of agglutinating Indian languages, such as Bengali and Hindi. In this paper, we explore the effect of decompounding on IR effectiveness for an agglutinating Indian language, namely Bengali.

Existing approaches to decompounding mainly select splitting positions based on the maximum combined frequency of the candidate constituents [4, 5]. While such approaches have proven useful in increasing retrieval effectiveness for European languages [1–3], our experiments reported in this paper show that such approaches do not work particularly well for Bengali IR. This is due to the very different inherent characteristics of compounding in an Indian language such as Bengali, as compared to European languages. To understand this issue, let us briefly look at the compounding characteristics of the Bengali language.

Compounds can be decomposed into their constituent parts, which are then indexed in addition to the compound form. For example, assuming a compositional semantics, the German compound *Nasenspitze* (EN: tip of the nose) can be split into *Nase* (nose) and *Spitze* (tip). In Bengali, two words can be concatenated to represent a different concept. For example, the words *lok* (EN: people)<sup>1</sup> and *sabhA* (EN: assembly) can be compounded to form the word *loksabhA* (EN: parliament). In this case, therefore, it is not beneficial to split the compound word *loksabhA* into the constituents. Note that this is conceptually similar to phrases in English, where the phrase *House of Commons* represents a different concept than the constituents *house* and *common*, as a result of which an IR system should treat this phrase as one indexing unit instead of two. A frequency-based approach of decompounding such as [4], however, can split the compound word *loksabhA* into the constituents *lok* and *sabhA*, because both of these constituents are commonly occurring words and are thus likely to have a high frequency in a Bengali document collection. This in turn can potentially reduce retrieval effectiveness. Thus, a decompounding algorithm has to be selective in its decision making as to whether to split a word or not.

The second inherent characteristic of Bengali compounding is that one of the constituents of a compound may not be a valid dictionary word (or a rarely used or archaic word, thus less likely to occur in documents). For example, the word *upanagar* (EN: town) has *up* and *nagar* (EN: city) as its constituents. The prefix *up* expresses in some sense the equivalent concept of *small* in English, but is not a valid Bengali word. In this case, however, it may help to decompound the word *upanagar* into one constituent *nagar* (EN: city), since these words represent similar concepts.

---

<sup>1</sup> For every Bengali word, the transliteration in ITRANS notation followed by its English meaning is provided in this paper.

Another challenge in Bengali compounding arises due to the presence of complex compounding rules in Bengali, known as the *Sandhi*<sup>2</sup> rules. According to the Sandhi rule, the first character of the right (tail) constituent can appear in a modified form in the compound. An illustrative example is *suryoday* (EN: sun-rise) = *surya* (EN: sun) + *uday* (EN: rise), where it can be seen that the first character of the tail constituent, e.g. “u” is changed to “o” in the compound word. While it is easy to directly apply a Sandhi rule to the constituents and derive the compound, the reverse direction is more complex because one may need to apply the rules of Sandhi at each candidate split position and then check whether the modified second constituent appears in a dictionary of valid words.

In this paper, we propose a decompounding method addressing each of the problems introduced above as follows. To address the first issue, our proposed decompounding approach takes into account how closely related the compound word and its constituents are. To address the second problem, we relax the decompounding process by allowing decomposition of a compound word into constituents when at least one constituent is a valid word. The third issue is taken care of by applying Sandhi rules during decompounding. Our experiments show that for Bengali, indexing compounds together with their constituents can improve IR effectiveness considerably.

The rest of this paper is organized as follows: Section 2 presents a brief overview of related work. Section 3 provides a general overview of the compounding process and also introduces our proposed approach to decompounding in Bengali. Section 4 describes and discusses our IR experiments. Section 5 concludes with suggested directions for future work.

## 2 Related Work

Compounding is a word formation process joining two (or more) constituent words into a new word, the compound. This process can include the simple concatenation of constituent words, joining constituents together by linking elements, or other modifications. The reverse process is called decompounding, decomposition, or compound splitting.

Koehn and Knight [4] proposed a compound splitting approach for decompounding German words to find correct translations of compounds and improve machine translation (MT) quality. They examine all possible candidate splits and select the split with the highest probability, which is estimated by the product of constituent frequencies. They allow a few linking elements between compound constituents, e.g. an additional “s” or “es” between constituents. Braschler and Ripplinger [2] investigated stemming and decompounding for German IR, comparing different decomposition approaches, ranging from language independent methods to linguistic methods, including freely available and commercial solutions. They found that stemming and careful decomposition boosts IR performance.

---

<sup>2</sup> The word *Sandhi* literally means *compounding*.

Bengali compounding is derived from Sanskrit compounds and the analysis of Bengali compounds has a long history. Dash [6] attempted to capture lexico-semantic properties of Bengali compounds to describe syntactic and semantic properties of compound constituents and their change over time. Decompounding for Bengali IR has not been researched in detail, but there is some previous research on word formation and morphology in Bengali. Dasgupta et al. [7, 8] present a brief overview of morphological analysis of compound words in Bengali. They apply a unification-based morphological analysis to parse and split compound words while resolving ambiguities and handling inflectional variation. Roy [9] explores NLP for Bengali MT and investigates decompounding as a means for increasing the coverage for lower resourced languages such as Bengali. He observed that decompounding Bengali can decrease the word error rate and increase the BLEU score for MT.

Deepa et al. [10] generate a lexicon of Hindi compounds for speech synthesis. (Hindi compounding is very similar to Bengali compounding). Their approach involves searching a trie-based prefix dictionary to look for the candidate suffixes that can be appended to the current word to form a potential compound. For example, in order to split the word *loksabhA* (EN: the parliament), the algorithm traverses the trie, discovers that the prefix *lok* (EN: people) is a valid word, and that the suffix *sabhA* (EN: assembly) also exists in the dictionary as an independent word, and decompounds the word *loksabhA* into the constituents *lok* and *sabhA*. However, their approach is relatively simple because they did not consider the rules of Sandhi when splitting a compound word. According to the rules of Sandhi, the first character of the suffix constituent may change in the compounded word, which is not handled by the approach described in [10]. Another major difference between [10] and our work in this paper is that their evaluation was performed only on a small corpus of 50 words by comparing their results with respect to manually decompounded words, whereas our approach is applied for IR in Bengali language and thus evaluated using standard IR metrics. This ensures that our decompounding approach is tested on a much larger vocabulary of words. Also, we are able to observe the effect of the decompounding approach on IR effectiveness.

Indexing compound constituents is a linguistically motivated technique. There are several other approaches which aim at relaxing the requirement that index terms have to be words. McNamee et al. [11] and Leveling et al. [12] performed experiments on indexing character  $n$ -grams and subwords for Bengali IR. They found that indexing terms on a subword level, an approach similar to indexing compound constituents, can outperform other approaches based on stemming all words. The morpheme extraction task (MET) at FIRE<sup>3</sup>, the Forum for Information Retrieval Evaluation, was introduced in 2011 with an aim of evaluating and comparing different IR preprocessing techniques (with a focus on stemming), and to provide the corresponding software tools. The task shows that there is a growing interest in scientific evaluation of Bengali IR and natural language processing, but a lack of corresponding software tools.

---

<sup>3</sup> <http://www.isical.ac.in/~fire/morpho/MET.html>

There are very few software tools supporting Bengali decomposition. Sandhi splitter<sup>4</sup> is a computational tool which shows all possible splittings of a given Sanskrit string. In addition, PC-Kimmo has been extended to process Bengali compounds [8].

### 3 Bengali Compounding

In this section, we introduce some of the characteristics of compounding in Bengali. Compounds in Bengali are typically formed by concatenation of two (in rare cases more) constituent words, which can be modified in the compounding process. The compounding rules for Bengali are derived from Sanskrit and are called Sandhi rules. For the experiments described in this paper, we consider hyphens as word delimiters and do not consider decomposing hyphenated words as a problem. In contrast, Roy [9] considers splitting Bengali words at hyphen characters whereas we view hyphens as word delimiters by default.

Let the compound word  $w$  be formed of a left constituent (usually called *modifier*), denoted by  $w_L$ , and a right constituent, denoted by  $w_R$  (usually called *head*). Words are concatenated together (without hyphens), with possible morphological inflections and modification of characters on  $w_R$ . Inflections on the constituent  $w_L$  are not allowed. In European languages, compounds are predominantly endocentric, i.e. a compound  $w = w_L + w_R$  denotes a special kind of  $w_R$ . For example,  $w = \text{“darkroom”}$  means that  $w$  is a special kind of *“room”*. In Indian languages, exocentric compounds (Bahuviri compounds, where  $w_L + w_R$  denotes a special kind of an unexpressed semantic head) could be more frequent. For example, *“skinhead”* refers to a person (unexpressed).<sup>5</sup> We consider four possible cases when splitting a compound:

- Both  $w_L$  and  $w_R$  are valid dictionary words.
- $w_L$  is a valid dictionary word, and the first character of  $w_R$  is modified according to *Sandhi* rules. An example Sandhi rule is that if the first character of  $w_R$  is an independent vowel (e.g. *Aa*), and the last character of  $w_L$  is a consonant, then the independent vowel is changed to a dependent one and is appended after the last character of  $w_L$ .
- $w_L$  is not a valid dictionary word, but  $w_R$  is. For example,  $w_L$  could be a bound morpheme or a word prefix that does not occur independently in the dictionary.
- $w_R$  is not a valid dictionary word, but  $w_L$  is.

Table 1 shows an example of each case along with the frequencies in the FIRE 2008 document collection<sup>6</sup> of newspaper articles for ad hoc IR. The frequencies in the left-most column of the table show that a high percentage of words in this Bengali collection can be compound words (39.54%), out of which 29.83% +

<sup>4</sup> [http://tdil-dc.in/san/Sandhi\\_splitter/index\\_dit.html](http://tdil-dc.in/san/Sandhi_splitter/index_dit.html)

<sup>5</sup> Our proposed decomposing approach would leave this word unchanged, as *“skinhead”* rarely co-occurs with *“head”*.

<sup>6</sup> <http://www.isical.ac.in/~clia/>



**Table 1.** Compound examples in Bengali. The frequencies are reported on the FIRE 2008 document collection. Each Bengali word (transliterated in ITRANS) is accompanied by its translation into English.

Freq.	Conditions	$w$ (Compound)	$w_L$	$w_R$
3.3%	$inDict(w_L) \wedge inDict(w_R)$	<i>mulyabridddhi</i> (EN: price-hike)	<i>mulya</i> (EN: price)	<i>briddhi</i> (EN: hike)
29.8%	$\neg inDict(w_L) \wedge inDict(w_R)$	<i>upanagar</i> (EN: town)	<i>up</i> (EN: vice)	<i>nagar</i> (EN: city)
3.9%	$inDict(w_L) \wedge \neg inDict(w_R)$	<i>moshAri</i> (EN: mosquito net)	<i>moshA</i> (EN: mosquito)	<i>ari</i> (EN: enemy)
2.5%	$inDict(w_L) \wedge inDict(\text{applySandhi}(w_R))$	<i>purbAnchal</i> (EN: eastern region)	<i>purba</i> (EN: east)	<i>anchal</i> (EN: region)

**Table 2.** Selected vowel Sandhi types

Sandhi	Rule	Bengali Example / English translation
Dirgha	(a + a = A)	<i>sUrja + asta = sUrjAsta</i> (EN: sun + set = sunset)
Dirgha	(a + A = A)	<i>mAdak + Asakta = mAdakAsakta</i> (EN: drug + addicted = drug addict)
Dirgha	(A + A = A)	<i>vidyA + Alaya = vidyAlaya</i> (EN: education + house = school)
Guna	(a + i = e)	<i>shrabaN + indriya = shrabaNendriya</i> (EN: hearing + organ = ear)
Guna	(a + u = o)	<i>sUrja + udaya = sUryodaya</i> (EN: sun + rise = sunrise)

3.86% = 33.69% of the words are representative of the cases where only one constituent is a valid dictionary word.

The decomposition process can be complex. Firstly, there may be more than one viable splitting point and the decomposing process has to take into consideration all possible splitting points in a word. Secondly, it has to choose the most likely split in a set of candidate splits. Thirdly, it can be necessary to modify the first character of the constituent  $w_R$  by applying the rules of Sandhi. In the next section, we describe our approach to decomposing which considers all of these steps.

### 3.1 Proposed Decomposing Algorithm

Before describing our proposed algorithm, we first outline its two auxiliary procedures.

- $inDict(w)$  is a unary predicate which returns true if the stem of the word parameter  $w$  is found in the dictionary. The dictionary, in our case, comprises the vocabulary of the indexed document collection.

- *applySandhi*( $w_L, w_R$ ) transforms the first character of the right constituent into another character according to the rules of Sandhi. The *applySandhi* method handles the most frequent Sandhi rules.

Consonant Sandhis occur rarely in the corpus. Examples for the vowel Sandhi rules (Dirgha and Guna Sandhi) are shown in Table 2. We list the steps of our algorithm for splitting a candidate compound word  $w$  as follows.

```
// initialization
·  $mw = \text{min. word length}$  // words comprise at least 2 consonants and 1 vowel
·  $splits = \{\}; result = \{w\}$ 
// generate candidate splits
· FOR  $i = mw - 1$  TO  $length(w) - mw - 1$ 
  · split  $w$  into  $w_L$  and  $w_R$  at position  $i$ 
  ·  $w'_R = \text{applySandhi}(w_R)$ 
  · IF  $\text{inDict}(w_L)$  AND  $\text{inDict}(w_R)$  THEN  $splits = splits \cup \{w_L, w_R\}$ 
  · IF  $\text{inDict}(w_L)$  THEN  $splits = splits \cup \{w_L\}$ 
  · IF  $\text{inDict}(w_R)$  THEN  $splits = splits \cup \{w_R\}$ 
  · IF  $\text{inDict}(w_L)$  AND  $\text{inDict}(w'_R)$  THEN  $splits = splits \cup \{w_L, w'_R\}$ 
· END FOR
// select best split
· let  $w_L$  and  $w_R$  represent the element in  $splits$  with the highest value of  $cf(w_L) + cf(w_R)$ .
· IF  $\text{overlap}(c, w) > \tau$  // see Equation 1, where  $c \in \{w_L, w_R\}$ 
  THEN  $result = result \cup c$ 
· RETURN result
```

Our proposed decomposition process is similar to that of [4] and [5] in the sense that we consider all possible candidate splits, and score the candidate splits based on the corpus frequency of compound constituents. However, there are three major differences as follows. The decomposing approach in [4] considers only those decompositions where  $w_L$  and  $w_R$  are both valid dictionary words. In contrast, due to the linguistic characteristics of Bengali, we needed to consider different cases as described in Section 3.

The second difference is that since decomposing in [4] is performed to improve MT performance, the decision of whether to split a compound word or not was motivated by comparing the collection frequency of the compound with the sum of the frequencies of its constituents. More specifically, a word  $w$  is split into the constituents  $w_L$  and  $w_R$  only if  $cf(w_L) + cf(w_R) > cf(w)$ . The reason for this is that it is more likely to find a translation of a highly frequent word in a corpus parallel to the current one. Thus, if the constituents occur more frequently in the corpus, decomposing a compound word can increase their frequencies even more. In IR however, highly frequent words, due to low inverse document frequency (*idf*), do not play a significant role in determining retrieval output. It is rather the addition of the high *idf* terms which can boost the retrieval score of a document significantly in response to a given query. Thus, a selection rule such as the one proposed in [4] may not be particularly suitable

**Table 3.** Document/Query characteristics

Data	#Documents	Topics	Avg. #rel	Avg. qry length	
				T	TD
FIRE 2008	123,047	26-75	37.26	3.64	13.44
FIRE 2010	123,047	76-125	10.20	4.84	14.18
FIRE 2011	500,122	126-175	55.50	3.30	9.90
FIRE 2012	500,122	176-225	49.08	3.54	10.14

for IR. Our proposed algorithm thus does not involve such a check, and we allow decomponding of a word  $w$  into  $w_L$  and  $w_R$  even if  $cf(w_L) + cf(w_R) < cf(w)$ .

The third difference is that we attempt to estimate the *relatedness* between each constituent  $w_L$  and  $w_R$  and the compound word  $w$ , to avoid the cases where the constituents individually may represent concepts unrelated to the compound word. Some examples in Bengali are *dhAnbAd* (the name of a place) = *dhAn* (EN: rice) + *bAd* (EN: kept out), and *jalpai* (EN: olive) = *jal* (EN: water) + *pai* (EN: get). Adding the constituent words in such cases may be harmful, e.g. retrieval after decomponding can retrieve non-relevant documents on *Dhanbad* (a place) when the added constituent *dhan* (rice) is a query term. We investigate a co-occurrence based measure to selectively apply the decomposition rules only if the co-occurrence between a constituent and the compound is higher than a particular threshold. The intuition is that if a constituent word co-occurs frequently with the compound word, then they represent related concepts, whereas if the co-occurrence is low, then the constituent word is likely to represent a different concept. In the latter case, the compound should not be split. In the last step of the algorithm, we thus employ a co-occurrence check, which adds  $w_L(w'_R)$  only if its co-occurrence with  $w$  is higher than a threshold  $\tau$ . The co-occurrence measure used is the *overlap coefficient* between the set of documents  $D(c)$  containing the constituent term  $c$ , with that of  $D(w)$  containing the compound, as defined in Equation 1 [13].

$$overlap(w, c) = \frac{|D(w) \cap D(c)|}{\min\{|D(w)|, |D(c)|\}} \quad (1)$$

The cardinalities of the document lists  $D(c)$  and  $D(w)$  can differ hugely in which case a standard metric, such as the Jaccard coefficient, may be too small and thus difficult to threshold. The overlap coefficient on the other hand determines the ratio of the overlap compared to the minimum of the set sizes and hence is easier to threshold.

## 4 Experiments and Results

In this section, we describe the evaluation experiments for our proposed decomponding method. We start with a brief description of the dataset and tools, which is followed by a description of the different retrieval settings, and finally we present the results and a comparison between the approaches.

## 4.1 Dataset and Tools

To test the effectiveness of our proposed decomposing approach, we performed IR evaluations on the FIRE monolingual Bengali data used in ad hoc IR evaluations from 2008 to 2012 (see Table 3). Our IR experiments are performed using SMART<sup>7</sup>, with an extension to support language modelling (LM) with Jelinek Mercer smoothing [14]. The smoothing parameter  $\lambda$  was set to 0.4 by optimizing on the FIRE 2008 data. We employed stopword removal using a list of Bengali stopwords<sup>8</sup>. For stemming, we used our rule-based Bengali stemmer<sup>9</sup> [15], which produced the second best retrieval effectiveness in the morpheme extraction task (MET) in FIRE 2012. Note that stemming was applied prior to decomposing.

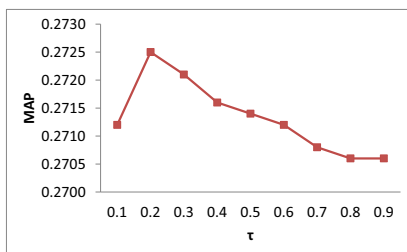


Fig. 1. Optimization of the correlation threshold  $\tau$  on FIRE 2008

We chose the topic set 2008 as the training set to optimize the parameter  $\tau$ , the correlation threshold of Equation 1. The variation of MAP with  $\tau$  for the FIRE 2008 data is shown in Figure 1, which shows a peak at 0.2. The optimal value of  $\tau = 0.2$  was set for the other topic sets as well.

## 4.2 Run Description

We investigated four different decomposing variants and compared them to a baseline experiment *BL* which uses no decomposing:

1. **CF**: We add the constituents with the highest probability estimate based on the sum of constituent frequencies as in [5]. Here,  $w$  is split into  $w_L$  and  $w_R$  only if  $cf(w) < cf(w_L) + cf(w_R)$ .
2. **CF<sub>2</sub>**: Similar to *CF*, with the additional constraint that decomposing is done only if two valid constituents are found, i.e. restricting *CF* to cases where both  $w_L$  and  $w_R$  are dictionary words. This is the standard decomposition technique for IR on European languages.

<sup>7</sup> <ftp://ftp.cs.cornell.edu/pub/smart/>

<sup>8</sup> [http://www.isical.ac.in/~fire/data/stopwords\\_list\\_ben.txt](http://www.isical.ac.in/~fire/data/stopwords_list_ben.txt)

<sup>9</sup> <http://www.computing.dcu.ie/~dganguly/rbs.tar.gz>

**Table 4.** Results for topic title (T) queries

Topics	<i>BL</i>		<i>CF</i>		<i>CF<sub>2</sub></i>		<i>DC<sub>0</sub></i>		<i>DC<sub>0.2</sub></i>	
	MAP	rel <sub>ret</sub>	MAP	rel <sub>ret</sub>	MAP	rel <sub>ret</sub>	MAP	rel <sub>ret</sub>	MAP	rel <sub>ret</sub>
2008	.2686	1605	.2699	1619	.2684	1604	.2706	1609	<b>.2725</b>	<b>1624</b>
2010	.3415	463	.3505	464	.3488	465	.3455	464	<b>.3508</b>	<b>468</b>
2011	.2410	2257	.2401	2251	.2407	2259	.2452	2253	<b>.2496</b>	<b>2270</b>
2012	.2026	1438	.2016	1429	.2018	1433	<b>.2043</b>	<b>1441</b>	.2039	1436

**Table 5.** Results for topic title and description (TD) queries

Topics	<i>BL</i>		<i>CF</i>		<i>CF<sub>2</sub></i>		<i>DC<sub>0</sub></i>		<i>DC<sub>0.2</sub></i>	
	MAP	rel <sub>ret</sub>	MAP	rel <sub>ret</sub>	MAP	rel <sub>ret</sub>	MAP	rel <sub>ret</sub>	MAP	rel <sub>ret</sub>
2008	.3118	1686	.3124	1687	.3111	1687	.3064	1687	<b>.3148</b>	<b>1696</b>
2010	.4315	<b>500</b>	.4348	<b>500</b>	.4325	499	<b>.4352</b>	498	.4336	498
2011	.3201	2464	.3202	2467	.3194	2474	.3245	2480	<b>.3279</b>	<b>2482</b>
2012	.2961	1763	.2966	1767	.2975	1765	.2966	1765	<b>.2985</b>	<b>1769</b>

- DC<sub>0</sub>**: Decompose words using the algorithm described in Section 3.1 with  $\tau$  set to 0, i.e. we decompose every word at the most likely splitting point, irrespective of any co-occurrence check. The major difference of this approach to *CF* is that *CF* does not decompose a word  $w$  if  $cf(w_L) + cf(w_R) < cf(w)$ , whereas *DC<sub>0</sub>* involves a more aggressive decomponing in the sense that we always decompose the word  $w$ . The objective of evaluating this approach is to see whether decomponing a word only to one constituent proves beneficial for retrieval.
- DC<sub>0.2</sub>**: Decompose by the algorithm in Section 3.1 with the co-occurrence threshold  $\tau = 0.2$  (cf. Figure 1), thus ensuring that a constituent is added only if its overlap coefficient with that of the compound is higher than 0.2.

It is worth textslasizing that Sandhi rules are applied on the tail constituent  $w_R$  for all the above approaches described while computing collection frequencies.

### 4.3 Results

Mean average precision (MAP) and the number of relevant documents retrieved (rel<sub>ret</sub>) are reported in Table 4 and Table 5 for the T and TD queries respectively. The results show that decomponing approaches in general can increase effectiveness for Bengali IR, in comparison to the baseline approach of no decomponing (BL). There is a consistent improvement in IR effectiveness when indexing compounds together with their constituents. The improvements, however, are not statistically significant, as measured by Wilcoxon signed rank test with 95% confidence measure.

The results also show that the standard strategy of decomposing based on the collection frequency estimate,  $CF$ , does not perform the best for Bengali. This can be seen by the lower MAP values in the second, third and the last row of Table 4 corresponding to title topics of 2010, 2011 and 2012. The fact that  $DC_0$  outperforms  $CF$  shows that an aggressive approach of decomposing proves beneficial for Bengali.

Moreover, the strategy of decomposing only if all constituents are valid words, i.e.  $CF_2$  performs worse than  $CF$ , as can be seen by comparing the MAP columns of  $CF$  and  $CF_2$  in Table 4 and 5. This suggests that for Bengali, it is beneficial to employ a relaxed decomposition and index at least one compound constituent (see the second and third row of Table 1).

Furthermore, we see that the method of selective decomposing based on the overlap coefficient consistently outperforms the selective decomposing with collection frequencies  $CF$  and  $CF_2$ , or decomposing without threshold ( $DC_0$ ). The only two cases where  $DC_0$  outperforms  $DC_{0.2}$  are the runs on the T query of FIRE 2012 and the TD query of FIRE 2010.

The best percentual improvement in MAP is 2.72% (on FIRE 2010 title queries) using the  $DC_{0.2}$  approach, which is lower than what has been reported for Dutch or German IR. For comparison, Monz et al. report 6.1% and 9.6% improvement for Dutch and German, respectively [3].

Our experiments show some promising results so far. Clearly, simply using approaches that have been proven successful for languages such as Dutch or German and applying them to Bengali does not produce the same improvements (see the results  $CF_2$  in Tables 4 and 5). In summary, the standard collection frequency based decomposing approach can yield some improvement in MAP. However, our proposed approach of selective decomposing shows a more consistent and typically higher improvement in the experiments, due to the more careful choice of decomposing a word using the degree of co-occurrence of the constituents with that of the compound.

## 5 Conclusions and Future Work

In this paper, we investigated the effect of decomposing on IR effectiveness for a relatively little researched Indian language, namely Bengali. This paper reviewed compounding characteristics of Bengali and differences compared to European languages. The major differences in compounding characteristics arise due to the rules of Sandhi where the first character of the second constituent appear in a modified form in the compound, and due to the fact that constituents may not be valid dictionary words.

The very different characteristics of Bengali compounding led us to propose a selective decomposition method based on the co-occurrence of the constituents and the compound. We observe that for Bengali, selective decomposing with a co-occurrence threshold works best, improving MAP up to 2.72%. We also find that a relaxation of the decomposition process, i.e. allowing decomposition even if only one constituent is a valid word, proves beneficial to improve retrieval quality.

As part of future work, we want to investigate the effect of compounding in other Indian languages, such as Hindi and Marathi. We also want to investigate the effect of our co-occurrence based constituent selection approach for non-Indian languages such as Dutch or German.

**Acknowledgments.** This research is supported by the Science Foundation of Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (<http://www.cngl.ie/>).

## References

1. Alfonseca, E., Bilac, S., Pharies, S.: Decomponing query keywords from compounding languages. In: ACL/HLT 2008, HLT-Short 2008, pp. 253–256 (2008)
2. Braschler, M., Ripplinger, B.: Stemming and decomponing for German text retrieval. In: Sebastiani, F. (ed.) ECIR 2003. LNCS, vol. 2633, pp. 177–192. Springer, Heidelberg (2003)
3. Monz, C., de Rijke, M.: Shallow morphological analysis in monolingual information retrieval for Dutch, German, and Italian. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 262–277. Springer, Heidelberg (2002)
4. Koehn, P., Knight, K.: Empirical methods for compound splitting. In: EACL 2003, pp. 187–193. ACL, Stroudsburg (2003)
5. Chen, A., Gey, F.C.: Multilingual information retrieval using machine translation, relevance feedback and decomponing. *Inf. Retr.* 7(1-2), 149–182 (2004)
6. Dash, N.S.: The morphodynamics of Bengali compounds – decomposing them for lexical processing. *Language in India* 6 (2006)
7. Dasgupta, S., Khan, M.: Morphological parsing of Bangla words using PC-KIMMO. In: ICCIT 2004 (2004)
8. Dasgupta, S., Ng, V.: High-performance, language-independent morphological segmentation. In: Sidner, C.L., Schultz, T., Stone, M., Zhai, C. (eds.) Proceedings of NAACL HLT 2007, April 22-27, pp. 155–163. ACL, Rochester (2007)
9. Roy, M.: Approaches to handle scarce resources for Bengali statistical machine translation. PhD thesis, School of Computing, Simon Fraser University (2010)
10. Deepa, S.R., Bali, K., Ramakrishnan, A.G., Talukdar, P.P.: Automatic generation of compound word lexicon for Hindi speech synthesis. In: LREC 2004 (2004)
11. McNamee, P.: N-gram tokenization for Indian language text retrieval. In: FIRE 2008, Kolkata, India (2008)
12. Leveling, J., Jones, G.J.F.: Sub-word indexing and blind relevance feedback for English, Bengali, Hindi, and Marathi IR. *TALIP* 9(3) (September 2010)
13. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999)
14. Hiemstra, D.: Using Language Models for Information Retrieval. PhD thesis, Center of Telematics and Information Technology, AE Enschede, The Netherlands (2000)
15. Ganguly, D., Leveling, J., Jones, G.J.F.: DCU@FIRE 2012: Rule-based stemmers for Bengali and Hindi. In: FIRE 2012, pp. 37–42. ISI, Kolkata (2012)

# Context-Dependent Semantic Annotation in Cross-Lingual Biomedical Resources

Rafael Berlanga<sup>1</sup>, Antonio Jimeno-Yepes<sup>2</sup>,  
María Pérez-Catalán<sup>1</sup>, and Dietrich Rebholz-Schuhmann<sup>3</sup>

<sup>1</sup> Department of Languages and Computer Systems  
Universitat Jaume I, Castelló, Spain  
{berlanga,maria.perez}@uji.es

<sup>2</sup> National ICT Australia  
antonio.jimeno@gmail.com

<sup>3</sup> Department of Computational Linguistics,  
University of Zürich, Ch  
rebholz@ifi.uzh.ch

**Abstract.** This paper presents a study about the impact of contexts in automatic semantic annotation over cross-lingual biomedical resources. Semantic annotation consists in associating parts of document texts to concepts described in some knowledge resource (KR). In this paper, we propose an unsupervised method for semantic annotation that regards contexts for validating the annotations. We test the method with two cross-lingual corpora, which allows us extracting correct annotations in the languages in the aligned corpora. Results show that annotated cross-lingual corpora provides grounds for qualitative comparison of semantic annotation algorithms.

## 1 Introduction

Automatic semantic annotation is becoming more and more popular in Life Sciences as newer and bigger knowledge resources become available [1, 2]. Extracting lexicons from these knowledge resources is a first step to perform the semantic annotation of free-texts. Relying on these lexicons, automatic semantic annotators perform dictionary look-up to find the concepts that fit better with the target text. Additionally, as some knowledge resources provide lexicons for different languages, automatic semantic annotators can be used as a valuable cross-lingual tool for integrating documents written in different languages. Unfortunately, most of the current semantic annotators disregard the annotation context, leading to ambiguous and incomplete annotations. In this paper, we investigate the impact of context-free annotation in cross-lingual scenarios. We study two main issues of context-free annotation: the ambiguity of annotations, and wrong annotations due to missing senses.

## 2 Methods

Performing the semantic annotation of a document  $D$  consists in finding mappings between text chunks of  $D$  (i.e., sequences of adjacent terms), and the



concepts provided by a knowledge resource (KR) that best semantically describes the content of  $D$ . In order to find out candidate concepts for each identified text chunk, the KR must provide a lexicon describing its concepts. We assume that there exists a function  $lex_{KR}^{lang}(C)$  that returns the set of strings describing the concept  $C$  in the language  $lang$ . This set of strings can contain different lexical variants of  $C$ , synonyms of these variants, and a short definition of the concept. We adopt the IR-based approach described in [1], which maps text chunks  $T$  to the KR lexicon strings of each concept  $C$  according to the following information-theoretic measure:

$$sim(T^{lang}, C) = \max_{s \in lex_{KR}^{lang}(C)} \frac{info(s \cap T) - info(T - s)}{info(s)}$$

The function  $info(s) = \sum_{w \in s} -\log(p(w|Background))$  estimates the information of a string  $s$  in terms of its probability in a background corpus (e.g., Wikipedia).

Current automatic annotation is performed independently from the context in which concepts are identified, assuming that the lexicons are well suited to the corpus to be annotated. However, the semantics of a concept may not fit with the context in which it occurs. Additionally, we have the problem of assigning a wrong concept to a text chunk because the correct concept is not present in the KR.

In order to take into account contexts, we use the disambiguation method presented in [3], which measures the similarity between the context words surrounding the annotation and the profile of the annotated concept. A concept profile is a vector of terms weighted by their relevance to the concept contexts.

### 3 Results

For semantic annotations, we have used the UMLS Metathesaurus<sup>®1</sup> (version 2012AB) as KR. This resource has more than a million concepts along with lexical items in several languages. As parallel corpora, we have used MEDLINE<sup>®</sup>, a bibliographic database containing more than 21 million biomedical citations, and the EMEA<sup>2</sup> corpora, a parallel corpus from EMA (European Medicines Agency) data available in several European official languages.

Table 1 shows the main features of the annotated collections. Annotations are calculated as the number of text chunks having associated some concept. The average size of an annotation is the average of the number of words of annotated text chunks. It is worth mentioning that around 30% of the annotations have more than one word. We also measure the percentage of ambiguous annotations, which are those having more than one entity type associated. In general, English collections generate more ambiguous annotations than the Spanish ones. However, this result is mainly due to the higher noise of the English lexicon.

Regarding the lexicons extracted from UMLS, the number of concepts of the English version is considerably much larger than the Spanish one. Around 52%

<sup>1</sup> <http://www.nlm.nih.gov/research/umls>

<sup>2</sup> <http://opus.lingfil.uu.se/EMEA.php>

**Table 1.** Features of the annotations generated for the selected datasets

Collection	Documents/Units	Annotations	Ann. Avg. size	Ambiguity
EMEA EN	879/364005	373971	1.3	5.2%
EMEA ES	895/140552	433671	1.5	5.6%
MEDLINE EN	1593546/1593546	3529800	1.5	5.6%
MEDLINE ES	247655/247655	610636	1.5	7.0%

of concepts in English have no translation to Spanish in the KR. For the tested corpora, the percentage of missing translations is around 30%, which decreases to 21% by generating lexical variants from Spanish to English.

Concerning to the overlap of concepts, and therefore the capacity of performing cross-lingual tasks, we report in Table 2 the results at collection and aligned unit levels. It must be noticed that overlap at collection level is much higher than at unit level due to numerous discrepancies at unit level like word coordination order and alignment errors.

**Table 2.** Overlap of concepts at collection and aligned unit levels

Collection	Collection level	Unit level
EMEA EN/ES (Ed)	79.7%	58%
EMEA EN/ES	76.9%	52%
MEDLINE EN/ES (Ed)	42.0%	51%

Looking only at the context-free EMEA annotation, the top concepts in both languages have a high correlation and denote medical concepts, including terms related to *patients*, *medicines*, *doctors* and population groups and drug related terms. Despite their similarity in annotation, identified by similar UMLS concept identifiers, there are as well differences since English terms like *injection* seem to be expressed differently in Spanish *solución inyectable*, which does not exactly match the English one. The same happens to the Spanish ones like *niños*, which is expressed in English using both *children* and *adolescents* depending on the age range, so linked to different UMLS concepts.

Results for the proposed disambiguation method over this benchmark are shown in Table 3. For the EMEA we have considered the unit and the document as the context. This is not possible with the MEDLINE corpus since documents and units are equivalent. Disambiguation results show that for the EMEA corpus the document provides a better disambiguation context. We find as well that for both corpora the disambiguation results are better for Spanish.

The lexicon used for annotation is a subset of the UMLS, the disambiguation method considers the whole UMLS, thus looking for missing senses that did not appear in the lexicon. Disambiguation performance on the lexicon senses is quite high, denoted by *Correct Accepted* and *Incorrect Accepted*, but the disambiguation performance of missing senses in our lexicon, denoted by (*Correct Discarded* and *Incorrect Discarded*), is much lower.

**Table 3.** Disambiguation results

	EMEA		EN unit EN doc		MEDLINE	
	ES unit	ES doc	EN unit	EN doc	ES	EN
Correct Accepted	190617	195713	184264	189028	248220	264575
Incorrect Accepted	14397	10338	21690	16928	20399	24703
Correct Discarded	26809	30868	40516	45278	23011	40083
Incorrect Discarded	33028	27932	39667	34903	34897	52184
Total	264851	264851	286137	286137	326527	381545
Accuracy	0.8209	0.8555	0.7856	0.8189	0.8307	0.7985

## 4 Conclusions

We have explored the semantic annotation of cross-lingual corpora in the biomedical domain for English and Spanish languages. The multi-lingual lexicon is a subset of the UMLS covering the most relevant entities in the biomedical domain. We have cross-checked the annotations to qualitatively evaluate the performance of the semantic annotator, and evaluated if this could be used to improve the annotator performance. The evaluation could be extended to multiple languages, even though this might be limited to the coverage of multi-lingual resources like the UMLS.

**Acknowledgements.** This work has been partially funded by the Spanish National R&D Programme project with contract number TIN2011-24147 of the “Ministerio de Economía y Competitividad”, and the EU STREP project grant 296410 (“Mantra”) under the 7th EU Framework Programme within Theme “Information Content Technologies, Technologies for Digital Content and Languages” [FP7-ICT-2011-4.1]. National ICT Australia (NICTA) is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

## References

1. Pérez, M., Berlanga, R., Sanz, I., Aramburu, M.J.: A semantic approach for the requirement-driven discovery of web resources in the Life Science. *Knowledge and Information Systems* 34(3), 671–690 (2013)
2. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., Jimeno-Yepes, A.: Text processing through Web services: calling Whatizit. *Bioinformatics* 24(2), 296–298 (2008)
3. Jimeno-Yepes, A., Aronson, A.: Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC Bioinformatics* 11, 565 (2010)

# A Comparative Evaluation of Cross-Lingual Text Annotation Techniques

Lei Zhang<sup>1</sup>, Achim Rettinger<sup>1</sup>, Michael Färber<sup>1</sup>, and Marko Tadić<sup>2</sup>

<sup>1</sup> Institute AIFB, Karlsruhe Institute of Technology, Germany

<sup>2</sup> Faculty of Humanities and Social Sciences, University of Zagreb, Croatia  
{l.zhang, rettinger, michael.farber}@kit.edu,  
{marko.tadic}@ffzg.hr

**Abstract.** In this paper, we study the problem of extracting knowledge from textual documents written in different languages by annotating the text on the basis of a cross-lingual knowledge base, namely Wikipedia. Our contribution is twofold. First, we propose a novel framework for evaluating cross-lingual text annotation techniques, based on annotation of a parallel corpus to a hub-language in a cross-lingual knowledge base. Second, we investigate the performance of different cross-lingual text annotation techniques according to our proposed evaluation framework. We perform experiments for an empirical comparison of three approaches: (i) *Cross-lingual Named Entity Annotation* (CL-NEA), (ii) *Cross-lingual Wikifier Annotation* (CL-WIFI), and (iii) *Cross-lingual Explicit Semantic Analysis* (CL-ESA). Besides establishing an evaluation framework, our results show the differences between the three investigated approaches and demonstrate their advantages and disadvantages.

## 1 Introduction

Text annotation is about attaching additional information such as attributes, comments, descriptions, tags or links to a document or to textual units like words and phrases. In contrast to linguistic processing of natural language text, such as part-of-speech (POS) tagging and named entity recognition and classification (NERC), text annotation studied in this paper goes one level deeper. It enriches unstructured text with links to a knowledge base. In this regard, text annotation helps to bridge the gap between the ambiguity of natural language text and the corresponding formal representations in knowledge bases.

Text annotation as it is understood in this paper is defined in two ways: (i) linking entity mentions in documents to their corresponding representations in the knowledge base; (ii) linking the documents by topics to the relevant resources in the knowledge base. *Cross-lingual* text annotation becoming more and more popular goes beyond general annotation, as it faces the task of linking entities and topics across the boundaries of languages. Here, the text to be annotated and the resources in the knowledge base might be of different languages. In order to manage this new situation, a central knowledge base, where all entities are ultimately linked to, is needed. In our case, Wikipedia was chosen, as it is the

largest on-line encyclopaedia up to date. Its articles are contributed by millions of users over the Web and cover any entity or topic of interest for most end users over the world. In addition, Wikipedia articles that provide information about the same concept in different languages are connected through cross-language links. A wide range of applications can benefit from its multilingualism.

Within the context of globalization, mainly driven by the digital revolution, institutions of any kind can no longer focus only on documents written in one language, but instead operate in various markets in different languages. In such a globalized and multilingual society, cross-lingual text annotation is crucial for processing natural language text in many different tasks. The following scenarios illustrate its application potentials:

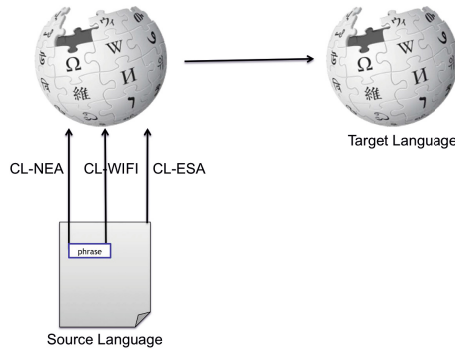
- *Entity Tracking*: A business news website provides current statistics about companies around the world. For each company a dedicated web page displays a list of up-to-date relevant news articles that mention the company. It is essential to detect mentions of each company in the real-time multilingual news streams and to provide the latest relevant company news, preferably from their home markets. This is the task called *entity tracking*.
- *Topic Detection*: For a press agency, it is extremely important to determine the topic coverage of its news articles. As such, detecting the current topics from the global news streams, especially in different languages, is a task of great significance called *topic detection*. It can provide the editors with better understanding of recent developments in the global news topics and will indicate demand on the publishing market – i.e., what the publisher should write about because it is relevant to their audience and not yet or poorly covered from a global perspective.
- *Cross-lingual Recommendation*: An on-line news delivery service recommends relevant articles to its users around the world using materials previously read by the users as the context. To cater for its global customer readership, this service processes the multilingual news streams and provides *cross-lingual recommendations*, the task of finding relevant articles in different languages.

These scenarios described above motivate our study of cross-lingual text annotation in this paper. Regarding the *entity tracking* scenario, due to the general applicability of Wikipedia which contains an enormous number of entities in diverse domains, there is no problem to define the interests of the customers as a set of Wikipedia pages<sup>1</sup>. As a consequence, statements about whether specific newswire articles written in different languages are of interest can be made by linking entity mentions to the corresponding Wikipedia pages. In addition, Wikipedia covers a wide range of topics<sup>2</sup>. Therefore, cross-lingual text annotation can also be employed for *topic detection* by linking articles to their Wikipedia topics. In the case of *cross-lingual recommendation*, a measure to compute the

---

<sup>1</sup> E.g. [http://en.wikipedia.org/wiki/Deutsche\\_Bank](http://en.wikipedia.org/wiki/Deutsche_Bank) represents Deutsche Bank AG, the German global banking and financial services company.

<sup>2</sup> Topics such as, but not limited to, arts, history, events, geography, mathematics, and technology.



**Fig. 1.** Approaches for Cross-lingual Text Annotation

similarity of texts in different languages is needed. However, due to the vocabulary mismatch problem, we cannot compare them directly. Through the annotation with Wikipedia, the documents in different languages will be first mapped to the entities or topics in a hub language in the knowledge base, e.g. English Wikipedia, before they can be compared.

The remainder of the paper is structured as follows: In Section 2, we present the approaches for cross-lingual text annotation. In Section 3, we describe our data, evaluation setting, and results followed by conclusions in Section 4.

## 2 Techniques for Cross-Lingual Text Annotation

In this section, we present three approaches: (i) *Cross-lingual Named Entity Annotation* (CL-NEA) based on named entity recognition and classification (NERC) techniques, (ii) *Cross-lingual Wikifier Annotation* (CL-WIFI) based on the state-of-the-art wikification system, and (iii) *Cross-lingual Explicit Semantic Analysis* (CL-ESA) based on the Explicit Semantic Analysis (ESA) method. It should be noted that for CL-NEA the NERC systems are trained for each language individually on the annotated data. In contrast, CL-WIFI and CL-ESA are directly trained on Wikipedia. Fig. 1 illustrates these three approaches mentioned above. It is observed that all of them make use of the cross-language links in Wikipedia to find the corresponding Wikipedia pages in the different target languages. In the following, we briefly describe these approaches.

### 2.1 Cross-Lingual Named Entity Annotation

Named entity recognition and classification (NERC) is the task within the field of information extraction (IE) of detecting specific information units within text such as names of persons, organizations, and locations. Since its beginnings in the early 1990s, NERC tools primarily have focused on these few classes: PER, LOC, ORG, and MISC. During this time span, the focus evolved from rule-based algorithms to more and more machine learning techniques. In the following,

**Table 1.** Excerpt of the CoNLL 2003 data set. The first item on each line is a word, the second the corresponding part-of-speech (POS) tag, the third a syntactic chunk tag and the fourth the named entity tag.

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

we confine ourselves to supervised machine learning NERC techniques. They can be differentiated by the underlying model they use: Hidden Markov Model (HMM) [1], Decision Tree [2], Maximum Entropy Model (MEM) [3], Support Vector Machine (SVM) [4], or Conditional Random Field (CRF) [5].

For all supervised learning methods, appropriate training data is needed. Table 1 gives an impression of how such a training corpus for NERC can look like. For each term in a sentence, annotation in the form of a POS tag, a syntactic tag, and a NE tag has to be provided.

In our case, NERC for English and Spanish is performed by using AdaBoost on decision trees as described by Carreras et al. [6]. Carreras' approach has obtained best results in the CoNLL-2002 named entity extraction task and treats named entity recognition (NER) and named entity classification (NEC) as two separate tasks which are processed sequentially and independently. NER is performed as a combination of three local classifiers. These classifiers test simple hold decisions on each word in the text. For each target word several features such as lexical, syntactic, orthographical, and affix features are used. The task of NEC is to assign an entity type to an already found named entity and the multiclass multilabel AdaBoost.MH algorithm [7] is used. NEC is modeled here as a four-class classification problem with the four classes PER, ORG, LOC, and MISC. Training was performed by using the CoNLL 2003 data set<sup>3</sup> for English and an updated version of the CoNLL 2002 shared task data set for Spanish (today included in the corpus Ancora<sup>4</sup>). NERC for the German language is performed by using the Stanford NERC tool which is based on the conditional random field model. For training, the CoNLL 2003 data set was used again. For more information, see [8].

On top of the standard monolingual NERC processing, a straight-forward approach for finding the corresponding Wikipedia page in another language is deployed: at first, the NE string is used for a keyword search for the Wikipedia article in the same language having the NE as title; then the cross-language links of this Wikipedia page are used to find the corresponding Wikipedia article of the target language (here, English). NERC is used here as computationally

<sup>3</sup> <http://www.cnts.ua.ac.be/conll2003/ner/>

<sup>4</sup> <http://clic.ub.edu/ancora>

inexpensive, but viable way for entity recognition and classification and as a prerequisite for cross-lingual entity linking.

## 2.2 Cross-Lingual Wikifier Annotation

The process of augmenting phrases in text with links to their corresponding Wikipedia articles (in the sense of Wikipedia article annotation) is known as *wikification*. Training can here be performed on a Wikipedia dump directly. This means that we do not need any gold standard for the annotation of POS, syntactic chunk or NE tags for training, but only Wikipedia as corpus.

While Mihalcea and Csomai [9] met the challenge of wikification by using link probabilities obtained from Wikipedia's articles and by a comparison of features extracted from the context of the phrases, Milne and Witten [10] could improve the wikification service significantly by viewing wikification even more as a supervised machine learning task: Wikipedia is used here not only as a source of information to point to, but also as training data used to find always the appropriate link. Due to the richness of intra-wiki links and the large size of the English Wikipedia, evaluation showed better performance.

Entity linking in general consists of two main steps: entity detection and disambiguation. While disambiguation ensures that the detected phrases link to the correct entity (here: Wikipedia article) and therefore normally has to be done after entity detection, Milne and Witten let the disambiguation training phase be a prerequisite for detection.

Regarding training for disambiguation, three features are used: commonness, relatedness, and goodness of the context. The commonness of a candidate phrase is representing the proportion of linkage to the corresponding Wikipedia page in comparison to other link targets. With the help of the relatedness feature, the semantic context of the candidate phrase is taken into consideration. The relatedness is measured by the Google similarity distance (GSD) [11]. Since not all context terms are equal, but instead some are more meaningful, each context term is given a specific weight. By summing up the weights of the context terms, a feature context quality representing the goodness of the context can be generated. Based on these features, a classifier can be trained for disambiguation. The machine learning based link detection makes use of several features: link probability, relatedness, disambiguation confidence, generality, and location and spread. In this way, the context terms are used for learning what terms should and what should not be linked to.

As already presented before, linkage into another language is done by the cross-language links in Wikipedia.

## 2.3 Cross-Lingual Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) has been proposed recently as an alternative approach for semantic modeling of natural language by exploiting unstructured or semi-structured text corpora instead of the traditional hand-crafted resources such as WordNet, taxonomies, or ontologies. Based on a given set of concepts



with textual descriptions, ESA defines the representation of documents with respect to these concepts. Various knowledge sources for concept definitions have been used. One of the most prominent examples is Wikipedia [12,13]. Concepts are hereby defined by Wikipedia articles, each of which comprises a comprehensive exposition of a topic.

ESA has been successfully applied to compute semantic relatedness between texts [12] or in text categorization tasks [13]. In the context of the cross-language information retrieval (CLIR) task, ESA has been extended to a cross-lingual setting (CL-ESA) by mapping the semantic document representation from one Wikipedia space to a Wikipedia space of another language [14,15]. This is achieved by exploiting language links in Wikipedia. As we use this approach as our third one for cross-lingual annotation, we briefly describe the underlying theory in the following:

Essentially, CL-ESA takes as input a document  $d_s \in D_s$  in the source language  $L_s$  and maps it to a high-dimensional real-valued vector space spanned by a Wikipedia database  $W_t = \{a_1, \dots, a_n\}$  in the target language  $L_t$  such that each dimension corresponds to an article  $a_i$  acting as a concept. In this sense, the semantic representation of document  $d_s$  defined by concepts in  $W_t$  is given by the mapping function

$$\Phi(d_s) = [\phi(\tau_{t \rightarrow s}(a_1), d_s), \dots, \phi(\tau_{t \rightarrow s}(a_n), d_s)]^T$$

where  $\tau_{t \rightarrow s}(a_i)$  maps the Wikipedia article  $a_i$  in language  $L_t$  to the corresponding article in Wikipedia database  $W_s$  for language  $L_s$ .  $\phi(a, d)$  denotes the strength of association between the document  $d$  and the Wikipedia article  $a$  in the same language, which can be defined using a tf-idf function based on the bag-of-words model [14]. Due to the large number of Wikipedia articles, in practice we consider only the top- $k$  dimensions of the vector yielded by CL-ESA with the highest values. In our experiments, we set  $k = 100$ .

### 3 Experimental Evaluation

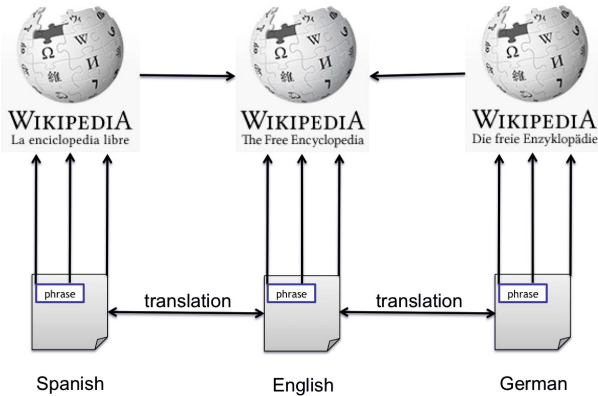
In this section, we propose a novel framework for evaluating cross-lingual text annotation techniques. According to this framework, we perform experiments to investigate the performance of the three approaches (CL-NEA, CL-WIFI and CL-ESA). Our focus is on an empirical comparison of these approaches w.r.t. the annotations (links) of documents in the source language (English, German and Spanish) to Wikipedia articles in the target language (English).

#### 3.1 Evaluation Setting

For the purpose of evaluation, we make use of a random sample of documents in English, German and Spanish from a parallel corpus<sup>5</sup> as test collection. While the evaluation of CL-NEA and CL-WIFI is focused on annotating word phrases

---

<sup>5</sup> Parallel corpus contains translated equivalents of documents in different languages.

**Fig. 2.** Evaluation setting**Table 2.** Statistics about Wikipedia

(a) Number of articles.

	English Wikipedia	German Wikipedia	Spanish Wikipedia
<i>#Articles</i>	4,014,643	1,438,325	896,691

(b) Number of cross-language links.

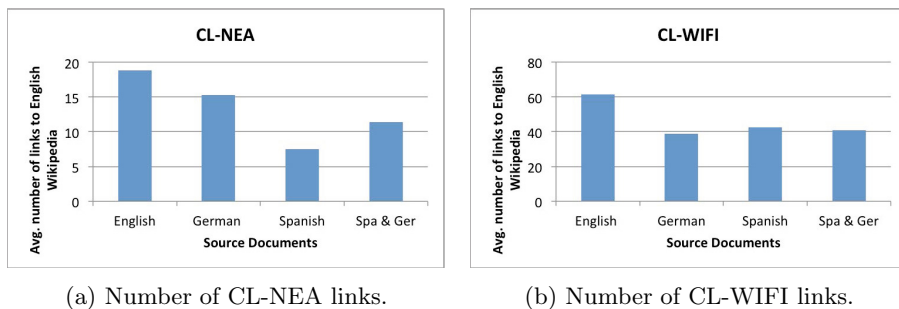
	English-German	English-Spanish	German-Spanish
<i>#Links</i> ( $\rightarrow$ )	721,878	568,210	295,415
<i>#Links</i> ( $\leftarrow$ )	718,401	581,978	302,502
<i>#Links</i> ( <i>merged</i> )	722,069	593,571	307,130

in the test documents and linking each phrase to a single Wikipedia article describing it, CL-ESA is evaluated by linking each test document to a certain number of Wikipedia articles which are topically relevant. The evaluation setting is illustrated in Fig. 2.

To provide the test documents, we use the parallel corpus JRC-Acquis<sup>6</sup>, which consists of legislative documents from the European Union and is widely used in cross-lingual research fields. The corpus is available in 22 European languages and comprises of approximately 23,000 documents in each language. In our experiments, we randomly select 88 parallel English-German-Spanish documents, each of which contains the translations of the same document in the above three languages.

Wikipedia is currently the largest knowledge base on the web and various editors develop it constantly, therefore its breadth and depth are expanding continually. The Wikipedia articles are available in approximately 270 languages

<sup>6</sup> <http://langtech.jrc.it/JRC-Acquis.html>



**Fig. 3.** Number of links detected by different approaches

and they are linked to each other via cross-language links in case they describe the same topic. Most Wikipedia articles are available in English (currently more than 4 million pages). The advantage of Wikipedia is that the articles are not only available in a vast amount with regard to the number of pages per language, but also with regard to the number of different domains in its different languages. That is why we use Wikipedia as our nucleus<sup>7</sup>.

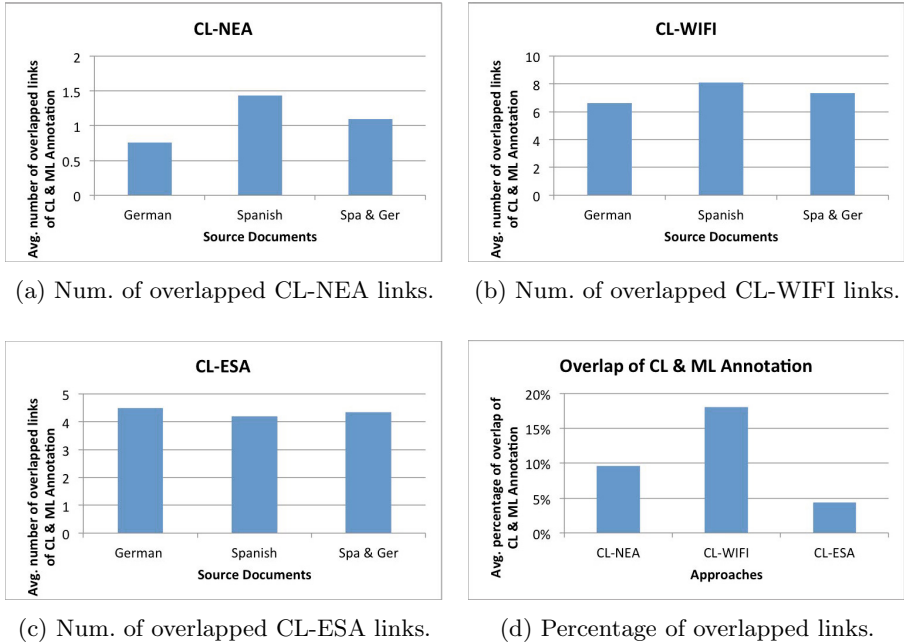
Table 2 shows some statistics of the Wikipedia articles in English, German and Spanish as well as the cross-language links between the articles in these languages extracted from Wikipedia snapshots of May 2012, which are used in our experiments. We analyze cross-language links between Wikipedia articles for each pair of supported languages in both directions and keep only articles for which aligned versions exist at least in one direction. For instance, we have extracted 721,878 cross-language links from English to German, and 718,401 links from German to English. By merging them together, we obtain 722,069 cross-language links, which are used to construct the cross-lingual knowledge base of the English-German language pair.

### 3.2 Evaluation Results

At first, we count the number of links to the English Wikipedia detected by each approach. Fig. 3a shows the average number of links per document detected by CL-NEA for different source languages. The results of CL-WIFI are shown in Fig. 3b. Concerning CL-ESA, we study whether the top-100 linked English Wikipedia topics are relevant to each test document. Therefore, the average number of detected links for each source language is 100.

It is expected that monolingual annotation of English documents detects more links than cross-lingual annotation of German/Spanish documents. This is due to the imbalance in the contents of Wikipedia in different languages and the missing cross-language links. In other words, English Wikipedia contains more articles, and not all Wikipedia articles in other languages are connected with their corresponding English versions. As shown in Fig. 3, for both CL-NEA

<sup>7</sup> The Wikipedia database dumps are available at <http://dumps.wikimedia.org/>.



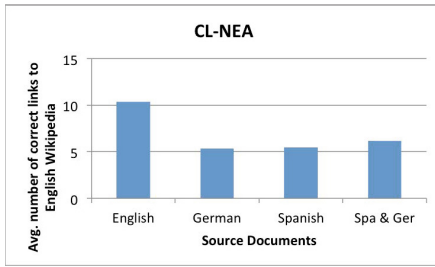
**Fig. 4.** The gap between cross-lingual and monolingual annotation

and CL-WIFI, more links are detected in English documents by monolingual annotation compared to cross-lingual annotation of German/Spanish documents, which conforms to our expectation.

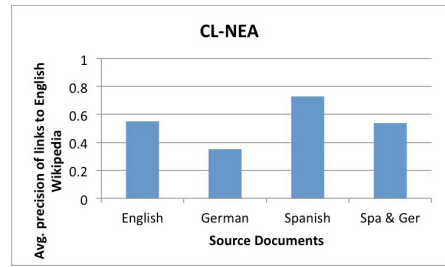
It should be noted that CL-WIFI produces many more annotations than CL-NEA. The reason for that as we believe is that CL-WIFI is trained directly on Wikipedia, while CL-NEA is firstly trained on some other data sets before the detected entities are grounded in Wikipedia in a second step. In this sense, a lot of entities covered in Wikipedia might be missing in the training data sets used by CL-NEA.

Further, we try an automatic processing by comparing the links to English Wikipedia detected by cross-lingual annotation of German/Spanish documents with the ones found by monolingual annotation of English documents. Since this processing was done on a collection of parallel documents, it is expected that the same annotations should be found in any language, which makes the detected links comparable.

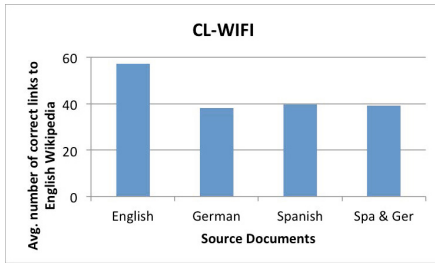
However, the number of the same links found by both cross-lingual and monolingual annotation indicates a low overlap between them. Fig. 4a shows the average number of overlapped CL-NEA links detected in both German/Spanish and English documents. The results of CL-WIFI and CL-ESA are illustrated in Fig. 4b and Fig. 4c, respectively. As shown in Fig. 4d, the average percentages of the overlapped links based on CL-NEA, CL-WIFI and CL-ESA are 9.6%, 18.1% and 4.4%, respectively. In general, we believe that the content imbalance and the



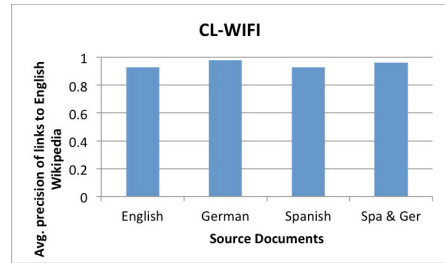
(a) Number of correct CL-NEA links.



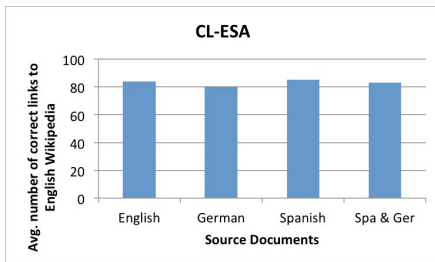
(b) Precision of CL-NEA links.



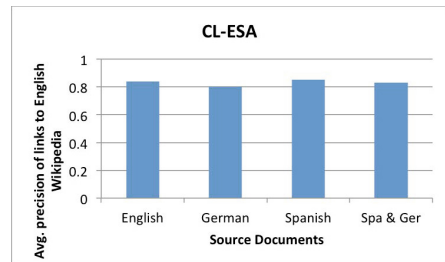
(c) Number of correct CL-WIFI links.



(d) Precision of CL-WIFI links.



(e) Number of correct CL-ESA links.



(f) Precision of CL-ESA links.

**Fig. 5.** Performance of different approaches

missing cross-language links in Wikipedia used by cross-lingual annotation is also the reason of such a low overlap for all approaches. Compared with CL-WIFI, the percentage achieved by CL-NEA is much lower. That is because CL-NEA is trained on the data sets that contain completely different named entities for each language while CL-WIFI is trained directly on Wikipedia in which there exists a larger overlap among the articles in different languages. It might seem less intuitive that CL-ESA which is also trained on Wikipedia even yields a lower percentage than CL-NEA. This is due to the fact that CL-ESA links the documents to the Wikipedia articles by topics based on the bag-of-words model. In such a coarse-grained manner, the specific contextual words in different languages increase the gap between cross-lingual and monolingual annotation in an unexpected way.

In addition to the automatic evaluation, we also investigate the performance of different approaches by a manual evaluation w.r.t. the number of correct links and the precision of detected links, i.e. the fraction of the correct ones. In this regard, the detected links to the English Wikipedia for each source language were manually evaluated by marking the correctness of them.

Figs. (5a+5c+5e) illustrate the number of correct links detected by each approach. Clearly, CL-ESA produces more correct links than CL-WIFI, which in turn finds more correct ones than CL-NEA. The average precision of links detected by CL-NEA is shown in Fig. 5b. The results of both cross-lingual and monolingual annotation are somewhat below our expectation. We believe the reason of less correct links and lower precision yielded by CL-NEA in comparison to the other approaches is still the distinction between its training data and Wikipedia. In contrast, the average precision obtained by CL-WIFI, as shown in Fig. 5d, exceeds 0.9 for all three languages. Fig. 5f shows the precision of CL-ESA links. Similar to CL-WIFI, CL-ESA trained on Wikipedia achieve much higher precision than CL-NEA. However, the more coarse-grained annotation of CL-ESA yields more correct links but slightly lower precision than CL-WIFI.

In summary, our experiments show that there are significant differences regarding the performance of the investigated approaches. As reasons we indicate the different training methods (using Wikipedia data or feature sets) and linking style (fine-grained or coarse-grained). Furthermore, the gap between cross-lingual and monolingual annotation is quite high – more than one would expect.

## 4 Conclusion

In this paper, we study the problem of cross-lingual text annotation. In particular, we investigate different approaches and propose a novel framework for evaluating them based on annotation of documents extracted from a parallel corpus to Wikipedia. According to the evaluation framework, we perform experiments for an empirical comparison of different approaches w.r.t. the performance of the annotation and analyze the reason of the variation of each approach. We are not aware of any previous evaluation framework and comparison of the investigated approaches w.r.t. cross-lingual text annotation tasks, so that our work represents an important contribution to the field and provides a step towards clarifying the differences between these approaches and demonstrating their advantages and disadvantages. Since the results clearly show a significant gap between cross-lingual and monolingual annotation, we consider narrowing such gap as our future work.

**Acknowledgments.** The authors acknowledge the support of the European Community’s Seventh Framework Programme FP7-ICT-2011-7 (XLike, Grant 288342) and of the German Federal Ministry of Education and Research (BMBF) under grant 02PJ1002 (SyncTech).

## References

1. Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a high-performance learning name-finder. In: Proceedings of the Fifth Conference on Applied Natural Language Processing, ANLC 1997, pp. 194–201. Association for Computational Linguistics, Stroudsburg (1997)
2. Sekine, S.: NYU: Description of the Japanese NE system used for MET-2. In: Proc. of the Seventh Message Understanding Conference, MUC-7 (1998)
3. Borthwick, A., Sterling, J., Agichtein, E., Grishman, R.: NYU: Description of the MENE Named Entity System as Used in MUC-7. In: Proceedings of the Message Understanding Conference, MUC-7 (1998)
4. Asahara, M., Matsumoto, Y.: Japanese Named Entity extraction with redundant morphological analysis. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL 2003, vol. 1, pp. 8–15. Association for Computational Linguistics, Stroudsburg (2003)
5. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, CONLL 2003, vol. 4, pp. 188–191. Association for Computational Linguistics, Stroudsburg (2003)
6. Carreras, X., Màrquez, L., Padró, L.: A simple named entity extractor using AdaBoost. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, CONLL 2003, vol. 4, pp. 152–155. Association for Computational Linguistics, Stroudsburg (2003)
7. Schapire, R.E., Singer, Y.: Improved Boosting Algorithms Using Confidence-rated Predictions. *Mach. Learn.* 37(3), 297–336 (1999)
8. Faruqui, M., Padó, S.: Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In: Proceedings of KONVENS 2010, Saarbrücken, Germany (2010)
9. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM 2007, pp. 233–242. ACM (2007)
10. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, pp. 509–518. ACM, New York (2008)
11. Cilibrasi, R.L., Vitanyi, P.M.: The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering* 19(3), 370–383 (2007)
12. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, vol. 6, p. 12 (2007)
13. Gabrilovich, E., Markovitch, S.: Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In: AAAI, pp. 1301–1306 (2006)
14. Sorg, P., Cimiano, P.: Cross-lingual Information Retrieval with Explicit Semantic Analysis. Working Notes of the Annual CLEF Meeting (2008)
15. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-Based Multilingual Retrieval Model. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 522–530. Springer, Heidelberg (2008)

# Mining Query Logs of USPTO Patent Examiners

Wolfgang Tannebaum and Andreas Rauber

Institute of Software Technology and Interactive Systems,  
Vienna University of Technology, Austria  
<http://www.ifs.tuwien.ac.at>  
{tannebaum,rauber}@ifs.tuwien.ac.at

**Abstract.** In this paper we analyze a highly professional search setting of patent examiners of the United Patent and Trademark Office (USPTO). We gain insight into the search behavior of USPTO patent examiners to explore ways for enhancing query generation in patent searching. We show that query generation is highly patent domain specific and patent examiners follow a strict scheme for generating text queries. Means to enhance query generation in patent search are to suggest synonyms and equivalents, co-occurring terms and keyword phrases to the searchable features of the invention. Further, we show that term networks including synonyms and equivalents can be learned from the query logs for automatic query expansion in patent searching.

**Keywords:** Patent Searching, Query Log Analysis.

## 1 Introduction

In preparing a patent application or judging the validity of an applied patent based on novelty and inventiveness, an essential task is searching patent databases for related patents that may invalidate the invention. Patent searching is usually performed by examiners in patent offices and patent searchers in private companies.

There is an increasing need to assist patent searchers in formulating queries, because query formulation is very time-intensive [1,5,6]. Yet, in the patent domain no sources, such as patent domain specific lexica or thesauri, are available. Actual queries being posed by patent experts could be valuable resources to explore the requirements for supporting patent searchers in query generation. The United Patent and Trademark Office (USPTO) has stored and published the query logs of the patent examiners. The goal of this paper is to analyze the query logs of the USPTO patent examiners to gain insights into the search behavior and characteristic of patent examiners queries. We first review state-of-the-art techniques for mining query logs. We then describe the nature of the query logs of USPTO patent examiners and analyze them. Following we present lexical term networks learned from the query logs. Finally, we provide conclusions and an outlook on future work.



## 2 Related Work

In several information retrieval applications query logs are being intensively studied. The purpose of all studies is to enhance either effectiveness or efficiency of searching by discovering patterns from query logs of search engines [2]. The main focus is on the analysis of web queries to enhance web searches [7]. Large-scale data sets of web queries, which have been made publicly available, such as AltaVista log or AOL log, have been studied [8]. Predominantly, basic statistics, such as query and term popularity, average query length, or co-occurring terms are used for characterizing the queries. Further specific analysis of the logs, such as distribution of the queries over time, variations of topics over time or distance between repetitions of queries over time, has been carried out. The classification of the queries, particularly through topic popularity, is a further task in mining query logs. The distribution of large-scale data sets across general topics enables to retrieve domain specific characteristics [7,8].

Finding query logs in the patent domain has been a difficult task [4]. Private companies and searchers are not interested in making their logs available as these may include terms revealing their current R&D activities. In earlier work we provided initial analyses of query logs of US Patent and Trademark Office (USPTO) patent examiners. We manually downloaded a limited set (346 log files) for one specific patent domain from the USPTO portal PAIR [10]. Initial results indicated that specialized term networks can be extracted directly from the query logs to complement resources for standard English [9]. In this paper we present a more in-depth analysis of this high professional search setting. We collect and analyzed the by now largest corpus of patent query logs to gain insight into query generation behavior as basis for automatic query expansion.

## 3 Query Logs of the USPTO

The query logs of USPTO patent examiners called “Examiner’s search strategy and results” are published for most patent applications since 2003 by the US Patent and Trademark Office Portal PAIR (Patent Application Information Retrieval) and can be downloaded from (<http://www.uspto.gov/>). The download is limited by the USPTO. For each patent application a verification code has to be entered. Google has begun crawling the USPTO’s public PAIR sites and provides free download of all patent applications published until now (<http://www.google.com/googlebooks/uspto-patents.html>). Google created single zip file for each patent application. Each file contains several folders including information on: Address and Attorney/Agent, Application Data, Continuity Data, Foreign Priority, Image File Wrapper, Patent Term Adjustments, Patent Term Extension History and Transaction History. The Image File Wrapper is of concern to us here. This folder can contain one or several query log files. Each query log of the USPTO is a PDF file consisting of a series of queries. Figure 1 shows an example, particularly an extract of four text queries of such a query log. Each query has several elements. We focus on the search query element showing the query formulated by the patent examiner. Further elements are reference, hits, database(s), default operator, plurals, and time stamp.

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
S1	1	mouth adj gaurd	US-PGPUB; USPAT; USOCR; FPFS; EPC; JPO	OR	ON	2011/07/18 09:53
S2	1	mouth with gaurd	US-PGPUB; USPAT; USOCR; FPFS; EPC; JPO	OR	ON	2011/07/18 09:54
S3	1	mouth near gaurd	US-PGPUB; USPAT; USOCR; FPFS; EPC; JPO	OR	ON	2011/07/18 09:54
S4	1151	mouth adj guard	US-PGPUB; USPAT; USOCR; FPFS; EPC; JPO	OR	ON	2011/07/18 09:54

**Fig. 1.** Example of a USPTO query log

There are several kinds of queries in the search query element. Text queries are used for querying whole documents (fulltext search) or only sections of patent documents, such as the title section (title search). Non-text queries are used for number search or classification search, for example “148/674.ccls.” for searching the class 148/674 for “Metal treatment”, specifically for “Cobalt or cobalt base alloy”. For query formulation text queries include search operators between the query terms. The types of search operators are (1) Boolean operators, such as “AND or OR” and (2) Proximity operators, like “SAME, ADJ(acent), NEAR, or WITH”. Furthermore, Truncation Limiters, such as “\$”, are used for query formulation. If the search operators are added manually, they are shown between the query terms in the text query element, else they are indicated by the default operator element. We are interested in the queries including the search operators.

## 4 Query Log Analysis

The USPTO published about 2.7 million patent applications, since 2003. The applications are classified into 473 US classes each including several subclasses. Hence, on average, about 6000 application documents are available for each US class. Because patent searchers use the classification system to narrow the search, we selected three collections of query logs each for a specific US class. We selected the US class 433 called “Dentistry”, the US class 128 called “Surgery” (a similar domain to the US class 433) and the US class 126 for “Stoves and Furnaces” (a domain very different from the US classes 433 and 128). For our query log analysis experiments we downloaded 2,721 files for the US class 126, 4,025 files for the US class 433 and 8,758 files for the US class 128. Through OCR conversion and segmentation of the 15,504 query log files we extracted the Boolean and Proximity Queries and the search operators between the query terms. We filtered all 3-grams in the form “X *b* Y”, where *b* is an Boolean or Proximity operator and X and Y are query terms.

### 4.1 Vocabulary Analysis

In this section we show for each US class some basic statistical properties of the vocabulary. At first we learn from the USPTO query logs how terms co-occur in

**Table 1.** Co-Occurring Terms based on Operator “OR”

<b>Stoves and Furnaces</b>	<b>Dentistry</b>	<b>Surgery</b>
tube pipe	tooth teeth	plurality plural
firewood fire	endodontic root	detection determination
hole opening	location position	motion movement
container pot	dental dentistry	stimulating stimulate
screen mesh	tube hose	hole opening

the query logs based on the Boolean and proximity operators. In Table 1 we present the five most frequently co-occurring terms for the three US classes based on the Boolean operator “OR”.

The majority out of the top-200 co-occurring terms are synonyms or equivalents at least for each specific domain. This show, that patent examiners use the Boolean operator “OR” to generate synonyms or equivalents. In Table 2 we show the top-five co-occurring terms based on the proximity operators “SAME”, “ADJ(cent)”, “NEAR” and WITH”. In all classes studied the majority of term pairs are keyword phrases. Hence, to narrow a search, particularly to limit a general query term, for example “mouth”, a keyword phrase is generated by the patent examiners, such as “mouth piece”.

**Table 2.** Co-Occurring Terms based on Proximity Operators

<b>Stoves and Furnaces</b>	<b>Dentistry</b>	<b>Surgery</b>
heat exchanger	teeth caries	blood vessel
liquid propane	dental implant	respiratory device
solar collector	dental bracket	intra vascular
fuel type	tooth brush	mouth piece
temperature sensor	wireless lan	tissue image

Further, we analyze the query terms of each class w.r.t. the part of speech using the CLAWS part of speech tagger [3], and if the query terms used by the patent examiners are domain specific (the terms appear only in one specific US class). We identified 37,097 unique query terms for class 126, 76,868 terms for class 433 and 80,208 terms for class 128. We find out, that in all classes about 70% of the terms are nouns followed by verbs (about 13%) and adjectives (about 10%). This can be useful for suggesting additional query terms from patent documents. The class 128 for “Surgery” and class 433 for “Dentistry” have the most common terms (3,673 terms) followed by the class 126 “Stoves” and US Class 433 “Dentistry” (having 3,483 common terms). Fewest common terms (1,751 terms) are shared between classes 126 and 128. Obvious, similar domains (classes 433 for “Dentistry” and 128 for “Surgery”) include more identical query terms than different classes. But we learn that patent searching is highly domain specific. Less than 5% of the query terms of the specific classes appear in the other classes, even across similar domains.

## 4.2 Search Operator Analysis

In this section we present for each class some basic statistical properties on the used search operators. First we analyze operator popularity for each domain based on the usage of the Boolean and proximity operators. Tab. 3 shows the relative spread of the used operators for formulating Boolean and proximity queries for each class.

**Table 3.** Search Operator Popularity

Search Operator	Stoves and Furnaces	Dentistry	Surgery
Boolean “OR”	57.65 %	46.92 %	48.24 %
Boolean “AND”	22.37 %	29.99 %	29.37 %
Proximity	19.98 %	23.09 %	22.39 %

In each domain about half of the queries are built using “OR”, nearly one third of the queries are generated using “AND” and the remaining queries are built by the proximity operators. The analysis shows, that the examiners’ behavior in formulating queries in the three domains is similar. For all domains they generate in the same proportions synonyms and equivalents, co-occurring terms and keyword phrases. Comparisons of the kinds of queries, particularly Boolean and proximity queries, show that two query terms can occur multiple times, but be connected by different operators. This would hint at conflicting usages, as two terms would be considered as synonyms and as phrases for more specific queries. The query terms “drill” and “bit” for example, appearing in the US class 433, are used in a Boolean and a proximity query. The proximity query serves to search the keyword phrase “drill bit”. The Boolean query is used to search for the synonyms or equivalents “drill” or “bit”.

## 5 Detecting Synonyms and Equivalents

In the patent domain significant efforts are invested to assist researchers in formulating better queries, preferably via automated query expansion. Currently, automatic query expansion in patent search is mostly limited on computing co-occurring terms. Learning synonyms and equivalents in the patent domain has been a difficult task. As we learned in Section 4 in patent searching the Boolean operator “OR” is used to expand a query term with an expansion term, which has the same meaning. We use that for automatically learning term networks from the query logs of USPTO patent examiners. Our approach resulted in 26,653 unique synonyms and 29,702 unique synonym relations for the three patent US classes as presented in Table 4 in detail.

**Table 4.** Learned Term Networks

US Class	unique relations	unique query terms
126	4,155	3,058
433	7,441	7,547
128	18,106	16,048
$\Sigma$	<b>29,702</b>	<b>26,653</b>

The learned lexical databases, particularly term networks, resemble thesauri of English terms for each specific patent domain. In each term network terms that have the same meaning are linked to each other. Finally, the learned term networks can be used in each specific US class for (semi-) automated query suggestion, particularly query expansion.

## 6 Conclusions and Future Work

In this paper we introduced and analyzed query logs of USPTO patent examiners. We show that query generation in patent searching is highly domain specific. Patent examiners follow a strict scheme for generating text queries. In each domain they use the Boolean operator “OR” to expand the queries and the operator “AND” for querying co-occurring features of the invention. The proximity operators are used to narrow the search, particularly to limit a general query term to a keyword phrase. Finally, means to enhance query generation in patent search are to suggest synonyms and equivalents, co-occurring terms and keyword phrases. Further we show, that specialized term networks including synonyms and equivalents can be extracted to complement resources for standard English. As shown in [9] this has positive effects on automated query expansion in patent searching. Currently, we are collecting and preprocessing a larger corpus of patent query logs to obtain a broader basis of USPTO classes. In future work we will focus on evaluating the performance of the learned term networks based on real query sessions done by the patent examiners. Further we want to use the proximity operators to learn term networks of keyword phrases, which we use for query limitation in patent searching.

## References

1. Azzopardi, L., Vanderbauwhede, W., Joho, H.: Search system requirements of patent analysts. In: *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, Geneva, Switzerland, pp. 775–776 (2010)
2. Clough, P., Berendt, B.: Report on the Treble CLEF query log analysis workshop 2009. *SIGIR Forum* 43, 71–77 (2009)
3. Garside, R., Smith, N.: A hybrid grammatical tagger: CLAWS4. In: Garside, R., Leech, G., McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pp. 102–121. Longman, London (1997)
4. Jürgens, J.J., Hansen, P., Womser-Hacker, C.: Going beyond CLEF-IP: The ‘Reality’ for Patent Searchers? In: Catarci, T., Forner, P., Hiemstra, D., Peñas, A., Santucci, G. (eds.) *CLEF 2012. LNCS*, vol. 7488, pp. 30–35. Springer, Heidelberg (2012)
5. Magdy, W., Jones, G.J.F.: A Study of Query Expansion Methods for Patent Retrieval. In: *Proceedings of PaIR 2011*, Glasgow, Scotland, pp. 19–24 (2011)
6. Piroi, F., Lupu, M., Hanbury, A.: Effects of Language and Topic Size in Patent IR: An Empirical Study. In: Catarci, T., Forner, P., Hiemstra, D., Peñas, A., Santucci, G. (eds.) *CLEF 2012. LNCS*, vol. 7488, pp. 54–66. Springer, Heidelberg (2012)

7. Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. *SIGIR Forum* 33, 6–12 (1999)
8. Silvestri, F.: Mining Query Logs: Turning Search Usage Data into Knowledge. *Foundations and Trends in Information Retrieval* 4(1-2), 1–174 (2010)
9. Tannebaum, W., Rauber, A.: Acquiring lexical knowledge from Query Logs for Query Expansion in Patent Searching. In: *The Proceedings of the 6th IEEE International Conference on Semantic Computing (IEEE ICSC 2012)*, Palermo, Italy (2012)
10. Tannebaum, W., Rauber, A.: Analyzing Query Logs of USPTO Examiners to Identify Useful Query Terms in Patent Documents for Query Expansion in Patent Searching: A Preliminary Study. In: *Salampasis, M., Larsen, B. (eds.) IRFC 2012. LNCS, vol. 7356*, pp. 127–136. Springer, Heidelberg (2012)

# Relevant Clouds: Leveraging Relevance Feedback to Build Tag Clouds for Image Search\*

Luis A. Leiva, Mauricio Villegas, and Roberto Paredes

ITI/DSIC, Universitat Politècnica de València  
Camí de Vera, s/n – CPI edif. 8G, 46022, Spain  
{luileito,mvillegas,rparedes}@{iti,dsic}.upv.es

**Abstract.** Previous work in the literature has been aimed at exploring tag clouds to improve image search and potentially increase retrieval performance. However, to date none has considered the idea of building tag clouds derived from relevance feedback. We propose a simple approach to such an idea, where the tag cloud gives more importance to the words from the relevant images than the non-relevant ones. A preliminary study with 164 queries inspected by 14 participants over a 30M dataset of automatically annotated images showed that 1) tag clouds derived this way are found to be informative: users considered roughly 20% of the presented tags to be relevant for any query at any time; and 2) the importance given to the tags correlates with user judgments: tags ranked in the first positions tended to be perceived more often as relevant to the topic that users had in mind.

**Keywords:** Image Search and Retrieval, Relevance Feedback, Tag Clouds.

## 1 Introduction

It is said that a picture is worth a thousand words, though the majority of commercial image search engines require the user to issue a textual query to retrieve images. This may be problematic because formulating the right query is difficult. This is especially true for users searching for uncommon topics or when users are unsure of how to express the query. In these cases, query autocompletion techniques might not be very helpful.

One possibility to improve search experience and increase retrieval performance of image search engines consists in assisting the user by suggesting tags that relate to the issued query. In this regard, tag clouds have been shown to be a useful approach [2,4,10]. For instance, Flickr features “tag clusters”<sup>1</sup> as tag clouds. Then, clicking on one tag within a tag cloud provides the user with semantic zoom, so that the initial image set is refined with images that were annotated with the clicked tag.

---

\* Prototype available at <http://risenet.iti.upv.es/rise/tc>

<sup>1</sup> E.g., <http://www.flickr.com/photos/tags/sky/clusters/>

Another option to increase the retrieval performance of image search engines is to capitalize on relevance feedback (RF) [8], i.e., presenting the user with a set of images according to the issued query, and letting the user select those images that are relevant and those that are non-relevant, possibly leaving some images unmarked. With this information, the retrieval engine can refine its results, leading to a hopefully better outcome after each RF iteration.

In an image search engine with RF, query suggestions can be derived by exploiting the relevance information given by the user [5]. The idea is that every time the user changes the image selection a new query is suggested, which the user can optionally follow to refine the initial search.

In this paper, we propose an alternative to the RF-based query suggestion approach. Our idea is presenting the user with a tag cloud that gets updated whenever the user selects/deselects images. This way, the tag cloud informs about the relevance of words for the images being selected; so that the most important tags would ideally be the ones that will help to retrieve more images of the kind the user has in mind. We also implemented a simple method and conducted a user study to support our idea.

## 2 Related Work

Tag clouds are seen as a “social” way to visualize information [10], and much work has been driven in this direction [4]. In the context of image retrieval, Callegari and Morreale [2] showed that less frequently used words in a tag cloud can significantly increase retrieval speed for the images associated with the tags. However, Zhang *et al.* [12] concluded that this has a mixed effect, as tags may lead the user to select irrelevant terms and introduce thus noise in the retrieval.

Typically, tag clouds are built either from keywords assigned by users [7,9] or from query logs [1,3]. This works well as long as the search engine has a very large user base and the query being searched is relatively popular among the users. Since these assumptions are not always fulfilled, other approaches to build tag clouds should be devised that are no so dependent on these factors.

In a different vein, Liu *et al.* [6] proposed an automatic image tag ranking method based on relevance labels. While this could be exploited to build RF-based tag clouds, unfortunately it is not always feasible nor scalable to have relevance labels for all of the crawled images. Moreover, their proposed method is computationally expensive for real-time applications. In the context of this paper, tag clouds should be *reactive*, in the sense that whenever the user indicates which images are considered relevant, a new tag cloud must be generated on the fly (see Figure 1).

## 3 RF-Based Tag Clouds

Our idea is to take advantage of the information obtained from RF to generate tag clouds. In other words, each time that the user modifies the set of relevant images, a tag cloud is updated according to this information. This behavior is





(a) Query "house", subset 1: related to the TV series.



(b) Query "house", subset 2: related to buildings.

**Fig. 1.** The tag cloud gets updated accordingly to inform the user about the topics that relate most to the selected images *and* less to the non-selected images

illustrated in Figure 1. The goal of our proposal is twofold. First, these tag clouds provide the user with a gist about the subjects that relate to a particular set of images, giving more importance to the relevant (selected) images than the non-relevant (non-selected) ones. Second, these tag clouds give another option to the user beyond traditional RF. Since tag clicking is optional, the user can alternate between traditional RF iterations and clicking on a tag to refine the presented image set.

This proposal has several lines of action that need to be explored in order to fulfill the underlying purpose, which is helping the user to retrieve the desired images with less effort. First, given a selection of images, words should be selected as candidates to be shown in the tag cloud. There are several resources from which the words can be obtained, e.g., text surrounding the image from web pages, automatic image annotation, image metadata, etc. Then, the candidate words need to be preprocessed with approaches that help to filter out unwanted tags, such as removing noise, stopwords and redundant terms. Once the candidate words are identified and preprocessed, they should be ranked in such a way that the highest scores are assigned to the words that would help the user to retrieve more relevant images. Finally, once the user clicks on a tag, it must be decided how this feedback information will be used for retrieving the next set of images. One example is to use the tag as a word that expands the original text query. Another possibility would be to use the tag to re-rank all the images that were retrieved with the original query. In this work, however, these alternatives are left as an opportunity for future work.

### 3.1 Proposed Approach

To make our approach scalable and applicable to any image on the Web, images are automatically annotated by using the text near the image from the web pages that contain such an image. These annotations are weighted depending on word distance to the image, term frequency, and the DOM elements.

Let  $\{w_1, \dots, w_n\}$  be the words of the vocabulary, i.e., all of the different words that appear in the associated text of the  $N$  images being shown to the user. We denote the set of relevant images as  $\mathcal{Q}^+$  and the set of non-relevant images as  $\mathcal{Q}^-$ . Let  $\mathcal{W}$  be the set of words  $w_i$  that appear in any of the relevant images. Each word  $w_i \in \mathcal{W}$  is scored as follows:

$$s(w_i) = \left[ \frac{\frac{1}{|\mathcal{Q}^+|} \sum_{j \in \mathcal{Q}^+} t_{ij}}{\frac{1}{|\mathcal{Q}^+|} \sum_{j \in \mathcal{Q}^+} t_{ij} + \frac{1}{|\mathcal{Q}^-|} \sum_{k \in \mathcal{Q}^-} t_{ik}} \right] + \left[ \frac{\delta(|\mathcal{Q}^-|)}{|\mathcal{Q}^+|} \sum_{j \in \mathcal{Q}^+} t_{ij} \right] + E, \quad (1)$$

where  $\delta()$  is the Kronecker delta function,  $t_{ij}, t_{ik}$  are the weights of the word  $w_i$  in the relevant image  $j$  and irrelevant image  $k$ , respectively,  $E = |\{\forall j \in \mathcal{Q}^+ : t_{ij} \neq 0\}|$  is the number of relevant images which contain  $w_i$ , and  $\sum_n t_{n\theta} = 1, \forall \theta$ .

## 4 User Study

To date, we have not found any suitable labeled dataset to perform an automatic evaluation of RF-based tag clouds. Generally, public image datasets have relevance labels but either no associated text (e.g., ImageNet<sup>2</sup>) or a fairly limited amount of text (e.g., Web Queries<sup>3</sup>) from which to generate meaningful tags. On the other hand, Flickr has human-generated tags, but this does not extrapolate to every image on the Web. Moreover, manually labeling an image dataset to perform a rigorous evaluation of our proposal is rather difficult. The labeling would imply, for a given query and a series of image subsets for that query, to have a ground truth list of tags for each particular image subset. Therefore, with the intention of shedding light on the value of RF-based tag clouds for image search, we performed a controlled lab study. For future work, we will evaluate the retrieval performance of our proposal.

**Materials:** We crawled 30 million of images by querying Google, Bing, and Yahoo! using the English dictionary, and for each image the surrounding text from the web page was extracted [11]. Then, we compiled a list of 164 queries by merging the two subtasks of ImageCLEF 2012<sup>4</sup>.

**Participants:** Fourteen subjects (3 females) in their thirties ( $M=31.42$ ,  $SD=5.3$ ) were recruited via email advertising to participate in the study. All participants were regular users of image retrieval engines. Each participant was assigned 12 queries to evaluate.

**Procedure:** For each query, participants were presented with a set of the top 10 ranked images according to that query. Then, participants had to select

<sup>2</sup> <http://www.image-net.org>

<sup>3</sup> <http://lear.inrialpes.fr/~krapac/webqueries/>

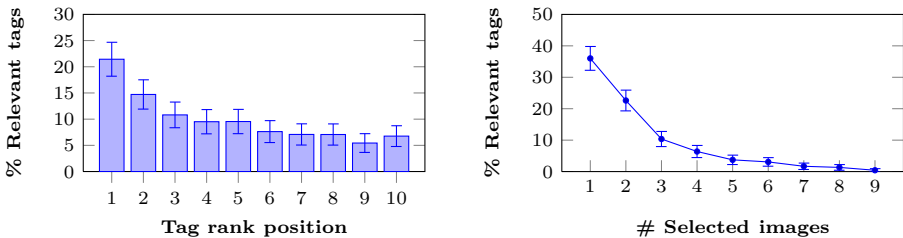
<sup>4</sup> <http://imageclef.org/2012/photo>

a subset of images for different subtopics from the presented image set. For instance, for the query "hot air balloon" one could select only clipart pictures, only photos showing two or more air balloons, or only photos taken from the inside of the balloon's basket. Then, whenever a relevant image was selected from the presented image set, a list of the top 10 scored RF-based tags was displayed. A check box attached to each tag allowed participants to indicate which tags were found to be most informative and/or most adequate to the different subtopics they had in mind for each presented image set. Participants had no restrictions on subtopic selection, i.e., no minimum or maximum subtopics per query were imposed, and a subtopic could have any number of relevant images associated.

## 5 Results and Discussion

In total, participants assessed the relevance of 928 tag lists. They reported that sometimes the tags shown were found to be really useful and beneficial for the current query, but also sometimes they were found to be meaningless. This fact may be explained in part by the noise due to our image indexing procedure, which was completely unsupervised and therefore the tag cloud may contain irrelevant terms for a particular query. The results of this experiment are shown in Figure 2a, where the bars represent the percentage of relevant tags (according to the participants) for each rank position as assigned by Eq. (1). As expected, tags with the highest scores tended to be perceived more often as relevant. Differences between the first ranked tag and the other tags are statistically significant at the 95% confidence level.

Figure 2b depicts the proportion of relevant tags according to the number of selected images. Differences between the number of tags presented when selecting 1 or 2 images with respect to the rest of selections were found to be statistically significant. A couple of observations were derived from this experiment: 1) as more images are selected, the topic overview the tag cloud provides about such a set of images tends to be more general; and 2) the perceived quality of the tags depends highly on the particular query.



**Fig. 2.** Evaluation results. Error bars denote 95% confidence intervals.

As observed, when a single image is selected, nearly half of the presented tags are considered to be relevant, since they are specifically tailored to such a single selection. We find it interesting for guiding users to nail down the concept of images they are looking for. On the contrary, selecting many images may be an indicator that the initial search is actually successful, so the associated tags are likely to be seen as less relevant. As a result, when many images are selected, a different strategy for generating RF-based tag clouds should be devised. Nonetheless, following our approach, 21.49% (SD=10) of all presented tags were considered as relevant at any time.

In general, participants liked the RF-based tag cloud idea. Some of them anecdotally commented that these tag clouds could be useful to decide which tags can lead to better retrieval results. All in all, this study indicates that our approach effectively informs the user about the relevance of the words for the images being selected. Furthermore, the tag cloud provides the user with more options to refine the image search results.

## 6 Conclusion

We have introduced the idea of generating RF-based tag clouds to improve image search, together with a simple approach that served as a proof of concept. The goal of these tag clouds is not only limited to providing the user with a gist about the underlying contents of the selected images. These tag clouds, in addition, give more options to the user beyond traditional RF. Then, a clicked tag can be used to disambiguate, filter, or re-rank the initial results and retrieve thus hopefully better images. We believe that this has an interesting potential and therefore deserves further research.

**Acknowledgements.** Work supported by EU FP7/2007-2013 under grant agreements 600707 (tranScriptorium) and 287576 (CasMaCat), and by the STraDA project (TIN2012-37475-C02-01).

## References

1. Begelman, G., Keller, P., Smadja, F.: Automated tag clustering: Improving search and exploration in the tag space. In: Collaborative Web Tagging (2006)
2. Callegari, J., Morreale, P.: Assessment of the utility of tag clouds for faster image retrieval. In: Proc. MIR (2010)
3. Ganchev, K., Hall, K., McDonald, R., Petrov, S.: Using search-logs to improve query tagging. In: Proc. ACL (2012)
4. Hassan-Montero, Y., Herrero-Solana, V.: Improving tag-clouds as visual information retrieval interfaces. In: Proc. InSciT (2006)
5. Leiva, L.A., Villegas, M., Paredes, R.: Query refinement suggestion in multimodal interactive image retrieval. In: Proc. ICMI (2011)
6. Liu, D., Hua, X.-S., Yang, L., Wang, M., Zhang, H.-J.: Tag ranking. In: Proc. WWW (2009)

7. Overell, S., Sigurbjörnsson, B., van Zwol, R.: Classifying tags using open content resources. In: Proc. WSDM (2009)
8. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: A power tool for interactive content-based image retrieval. *T. Circ. Syst. Vid.* 8(5) (1998)
9. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proc. WWW (2008)
10. Trattner, C., Lin, Y.-L., Parra, D., Yue, Z., Real, W., Brusilovsky, P.: Evaluating tag-based information access in image collections. In: Proc. HT (2012)
11. Villegas, M., Paredes, R.: Image-text dataset generation for image annotation and retrieval. In: Proc. CERI (2012)
12. Zhang, C., Chai, J.Y., Jin, R.: User term feedback in interactive text-based image retrieval. In: Proc. SIGIR (2005)

# Counting Co-occurrences in Citations to Identify Plagiarised Text Fragments

Solange de L. Pertile<sup>1</sup>, Paolo Rosso<sup>2</sup>, and Viviane P. Moreira<sup>1</sup>

<sup>1</sup> Instituto de Informática - UFRGS – Brazil  
{slpertile,viviane}@inf.ufrgs.br

<sup>2</sup> Natural Language Engineering Lab. - ELiRF  
Department of Information Systems and Computation  
Universitat Politècnica de València, Spain  
prossso@dsic.upv.es

**Abstract.** Research in external plagiarism detection is mainly concerned with the comparison of the textual contents of a suspicious document against the contents of a collection of original documents. More recently, methods that try to detect plagiarism based on citation patterns have been proposed. These methods are particularly useful for detecting plagiarism in scientific publications. In this work, we assess the value of identifying co-occurrences in citations by checking whether this method can identify cases of plagiarism in a dataset of scientific papers. Our results show that most the cases in which co-occurrences were found indeed correspond to plagiarised passages.

## 1 Introduction

Plagiarism is one of the most serious forms of academic misconduct. It is defined as the act of the appropriation of another person's ideas, words, or works without giving credit to the original source. With the growing popularity of the Internet, many documents are freely available enabling students and researchers to reuse words from other authors without crediting them.

According to a study by McCabe [10] 36% of undergraduate students and 24% of graduate students, admitted having copied or paraphrased sentences from the Internet without referencing them. More recently, the Journal of Zhejiang University-Science (China) [2] used the CrossCheck tool [1] to analyse the papers submitted to their revision process. They found that 22.8% (692 out of 2,233) of the papers presented unreasonable levels of copying or self-plagiarism [14]. High levels of text reuse have also been found by Gupta & Rosso [9] who analysed papers accepted by the ACL.

The interest in plagiarism detection has been rising in the last few years. The PAN benchmarks [3] has been running for four years with an increasing number of participants all over the world [12]. PAN's evaluations aim at detecting different forms of plagiarism providing a standardised evaluation framework.

Usually, plagiarism detection relies on content analysis. The idea is to identify text fragments in common between a suspicious document and possible sources.

Automatic detection techniques have been proposed to deal with the various forms of plagiarism. Content analysis is more difficult in the presence of paraphrasing [6] and even more so when more than one language is involved, *i.e.* in cross-language plagiarism [11].

More recently, Gipp et al. [8] propose methods for plagiarism detection based on citation analysis. In their work, two documents which cite the same references are considered as having a high degree of similarity. The ideas are interesting since citation-based plagiarism detection could potentially be used in cases which content-based retrieval is typically ineffective. However, experimental evidence of the effectiveness of citation-based methods is limited to the application of the method in a prominent case of plagiarism concerning the doctoral thesis of a German politician. In another study, Alzahrani et al. [5] use the citations within a scientific paper in a different way. Cases in which the original source has been properly referenced are ignored by the content analysis phase. Thus, citations are used as a filter and not as an evidence of similarity across papers.

In this paper, we aim to bridge the gap between these two aforementioned works. We compute citation co-occurrences on the dataset used in [5] and assess whether they are effective in pointing out cases of plagiarism.

## 2 Identifying Co-occurrences in Citations

Throughout this paper, we use the term *citation* to refer to the strings in the body of a scientific paper which point to where the original text was extracted from. The term *reference* is used to denote an entry in the Bibliography (or References) section of the paper.

Our aim is to compare the similarity of scientific papers based on the analysis of co-occurrences in citations. If two documents share at least a pair of citations within a text fragment, this is computed as an inter-document co-occurrence. Our assumption is that a high rate of inter-document co-occurrences is an indication of plagiarism. Given a pair of documents, these are represented as the co-occurrences of their citations. These intra-document co-occurrences are computed sliding a window of size  $s$  through the document. The inter-document co-occurrences are then calculated as the Jaccard similarity coefficient (or *overlap*) of these co-occurrences:  $sim(w_i, w_j) = \frac{w_i \cap w_j}{w_i \cup w_j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$  where  $w_i$  and  $w_j$  are the windows  $i$  and  $j$ , respectively,  $n_{i,j}$  is the number of shared co-occurrences between windows  $i$  and  $j$ ,  $n_i$  and  $n_j$  are the number of co-occurrences in windows  $i$  and  $j$ , respectively.

More specifically, the steps involved in our process are the following:

- **Pre-processing:** Identify citations within the text of the document and link them to their corresponding entry in the list of references.
- **Computing co-occurrences within a document:** Slide a window of size  $s$  through the document and compute co-occurrences within this window.
- **Computing co-occurrences across documents:** For each pair of co-occurrences between a window in a suspicious document and a window in a source document, check whether they match using approximate string matching.

**Table 1.** Results of Co-occurrence Analysis

	$s = 5$	$s = 15$	$s = 30$
Co-occurrences in Citations	90	160	161
Plagiarism with Co-occurrences	51	76	64
Precision	0.5667	0.4750	0.3975
Recall	0.0123	0.0183	0.0154
F1	0.0241	0.0353	0.0297

References with a similarity score higher than a given threshold  $t$  are considered as being referring to the same paper.

### 3 Experiments

The ideal dataset to analyse in our experiments would be a real collection of scientific papers with some cases of plagiarism. However, such a collection does not exist. Thus, we resorted to an artificial dataset originally described in [5] and available from [4], which is composed of scientific papers. There are 8,657 original and 6,755 suspicious papers containing verbatim and obfuscated cases of plagiarism. Annotation files revealing which fragments were plagiarised enable checking whether co-occurrences in citations are good indicators of plagiarism. At the moment, we can only handle papers which cite references using the numbered style. Thus, in our work, 4149 suspicious papers were compared against the 6035 source documents which adopt the numbered style.

In order to segment the references, we relied on the Ondux tool [7], which represents the state-of-the-art in information extraction by text segmentation. An extension of Levenshtein’s Edit distance, called Carla [13], which accounts for inversions of substrings, was used to compare references across documents. The similarity threshold used was  $t=0.86$ , based on empirical observations. Window sizes ( $s$ ) were 5, 15, and 30 through the documents.

The results are shown in Table 1. The smallest window (i.e., 5 lines) yielded the highest precision (56.67%), which means that in most cases in which co-occurrences were found, indeed correspond to plagiarism. The remaining cases with co-occurrences that were not considered plagiarism were due to three main reasons: (i) the suspicious document had cited the source from which text and references had been copied; (ii) two similar references were wrongfully treated as the same by our method, (iii) papers by the same authors and about the same topic had a high level of citation co-occurrence, but were not considered as plagiarism. In some cases, the paragraphs in the suspicious and source documents have identical contents and still were not annotated as plagiarism.

On the other hand, only a very small fraction of the cases of plagiarism have been identified. The main reason for the low recall is that, in most of the cases of plagiarism in this collection, the text fragment copied from the original did not include any references. Also, in some cases, the plagiarised fragment had been



extracted from an article from a totally different area (e.g. the source from a plagiarised text in economy was a paper on veterinary). In such cases, it is very unlikely that the source and suspicious paper would share any references.

## 4 Conclusion

This work presented a study on the validity of using co-occurrences in citations to detect plagiarism in scientific documents. We carried out experiments on a dataset of scientific papers with cases of plagiarism simulated artificially. Our results have shown that most of the cases with co-occurrences in citations correspond to plagiarism. Moreover, nearly all of these cases were within paraphrased text fragments. On the other hand, only a small fraction of plagiarism cases involved text fragments with citations. This suggests that a hybrid approach which combines content similarity and citation analysis can potentially yield better detection quality. As future work, we plan to test whether citation co-occurrences help identify portions of text reuse within a real collection of scientific papers by comparing with the results of [9] on the ACL corpus.

**Acknowledgements.** This work was partially funded by CNPq (478979/2012-6). Solange Pertile's 5-month internship at NLE Lab of Universitat Politècnica de València was funded by CAPES. P.Rosso's work was carried out in the framework of the the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and the European Commission WIQ-EI IRSES (no. 269180) and DIANA-APPLICATIONS-Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) research projects. We thank the authors of [5] for sharing their dataset with us and Enrique Flores for the preliminary brainstorming on how to identify co-occurrences in citations.

## References

1. CrossCheck, <http://www.crossref.org/crosscheck/>
2. Journal of Zhejiang University-Science, <http://www.zju.edu.cn/jzus/>
3. PAN, <http://www.pan.webis.de>
4. Plagiarism corpus, <http://www.c2learn.com/plagiarism/corpus/v1/>
5. Alzahrani, S., Palade, V., Salim, N., Abraham, A.: Using structural information and citation evidence to detect significant plagiarism cases in scientific publications. *JASIST* 63(2), 286–312 (2012)
6. Barrón-Cedeño, A., Vila, M., Marti, A., Rosso, P.: Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics* 39(4) (2013)
7. Cortez, E., da Silva, A.S., Gonçalves, M.A., de Moura, E.S.: Ondux: on-demand unsupervised learning for information extraction. In: *SIGMOD*, pp. 807–818 (2010)
8. Gipp, B., Meuschke, N.: Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence. In: *DocEng*, pp. 249–258 (2011)

9. Gupta, P., Rosso, P.: Text reuse with ACL (upward) trends. In: ACL 2012 Special Workshop on Rediscovering 50 Years of Discoveries, pp. 76–82 (2012)
10. McCabe, D.L.: Cheating among college and university students: A north american perspective. *International Journal for Educational Integrity* 1 (2005)
11. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-language plagiarism detection. *Language Resources and Evaluation* 45(1), 45–62 (2011)
12. Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Stamatatos, E., Rosso, P., Stein, B.: Overview of the 5th International Competition on Plagiarism Detection. In: CLEF 2013 - Working Notes (September 2013)
13. Ritt, M., Costa, A.M., Mergen, S., Orengo, V.M.: An integer linear programming approach for approximate string comparison. *European Journal of Operational Research* 198(3), 706–714 (2009)
14. Zhang, Y.: Crosscheck: an effective tool for detecting plagiarism. *Learned Publishing* 23, 9–14 (2010)

# The Impact of Belief Values on the Identification of Patient Cohorts

Travis Goodwin and Sanda M. Harabagiu

Human Language Technology Research Institute  
University of Texas at Dallas  
Richardson TX, 75080  
{travis,sanda}@hlt.utdallas.edu

**Abstract.** Retrieving relevant patient cohorts has the potential to accelerate clinical research. Recent evaluations have shown promising results, but also relevance measures that still need to be improved. To address the challenge of better modelling hospital visit relevance, we considered the impact of two forms of medical knowledge on the quality of patient cohorts. First, we automatically identified three types of medical concepts and, second, we asserted their belief values. This allowed us to perform experiments that capture the impact of incorporating knowledge of belief values within a retrieval system for identifying hospital visits corresponding to patient cohorts. We show that this approach generates a 149% increase for inferred average precision, a 36.5% increase of NDCG, and a 207% increase to the precision of the first ten returned documents.

## 1 Introduction

The advent of electronic medical records (EMRs) within the healthcare industry has immense potential for benefiting clinical research. By processing the narratives of EMRs, we can accurately and reliably identify a desired patient population and thus produce secondary usage of EMRs. To be able to evaluate the feasibility of using EMRs for patient cohort identification, the Text REtrieval Conference (TREC) launched a Medical Records Track in 2011 (TRECMed) [10]. This task, an information retrieval (IR) challenge, is pertinent to real-world clinical medicine because (1) it provides access to a large corpus of de-identified EMRs from the University of Pittsburgh Medical Center (<http://www.dbmi.pitt.edu/blulab>) and (2) it uses retrieval topics (queries) derived from an Institute of Medicine list providing conditions for comparative effectiveness research [4]. The topics which were processed as queries by the cohort identification system developed for the 2011 and 2012 challenges targeted specific hospital patient cohorts, characterised by various medical phenomena, as illustrated in Table 1. To be able to process topics similar to those listed in Table 1, cohort identification systems are provided access to a collection of medical records from the University of Pittsburgh BLU-Lab NLP Repository. This corpus constitutes 95,702 de-identified clinical records across multiple hospitals from 2007. Collected to aid NLP research, the documents

within the corpus are organized into 17,199 "hospital visits" wherein each hospital visit consists of all reports generated during a patient's hospital stay. These reports are composed of primarily free-text, and consist of medical histories, physical examinations, radiology reports, operative reports, and discharge summaries. Each report is lightly wrapped within eXtensible Markup Language (XML) containing the patient's admit diagnoses and discharge diagnoses as ICD-9 codes. Additionally, a mapping from individual clinical reports to their associate patient's hospital visit was provided.

**Table 1.** Examples of topics used in TREC Med 2011

#104. Patients diagnosed with localized prostate cancer and treated with robotic surgery.
#112. Female patients with breast cancer with mastectomies during admission.
#119. Adult patients who presented to the emergency room with with anion gap acidosis secondary to insulin dependent diabetes.

Cohorts are identified by a ranked list of hospital visits, in which the first hospital visit pertains to the patient deemed most relevant to the query's topic, while the following hospital visits correspond to patients from the same cohort in decreasing order of relevance. This constitutes a novel application of document retrieval wherein an incredible amount of medical knowledge must be processed in order to model the relevance of a given topic. Part of that knowledge consists of various medical concepts, such as medical problems, treatments, symptoms, and conditions. Another critical aspect of the knowledge encoded in a given topic constrains the gender or age of a patient. However, more importantly, in this paper, we claim that systems need to recognize the degree of beliefs associated with the medical concepts for (a) processing and expanding the topics and (b) processing the EMRs.

Mentions about clinical concepts are often qualified by the belief value asserted by their author, e.g., a symptom may be "present," "absent," a treatment may be "possible," "conditional," "hypothetical," "ongoing," "prescribed" or "suggested." In addition, the topics also present multiple cases of assertions that need to be identified in order to assess the relevance of hospital visits. For example, Topic #179, *patients taking atypical antipsychotics without a diagnosis [of] schizophrenia or bipolar depression*, asks for an assertion with the belief value ABSENT (for the medical problems of schizophrenia and bipolar depression). Since the belief values cast over medical concepts can influence the relevance of the retrieval criteria expressed by the topic, we asked ourselves whether a method of automatically identifying the belief values could improve the quality of patient cohort retrieval. To answer this research question, we used our system which was implemented for the TREC Med 2011 and 2012 evaluations and (1) processed both the topics and EMRs with the aim of identifying the medical concepts and their assertions; and (2) used a re-ranking of the retrieved hospital visits which accounts for knowledge about belief values associated with each medical concept. We also performed several other re-rankings and found that re-ranking based on assertions of medical concepts had the greatest impact on the

retrieval results. The remained of this paper is organized as follows. In Section 2, we present the architecture of a cohort identification system that has participated in the 2011 and 2012 TREC Med challenges. Section 3 details the topic and EMR analysis that enables the identification of medical concepts and their corresponding assertions. Section 4 describes the keyphrase expansion methodology which provides significantly improved recall. Section 5 provides details of the actual retrieval mechanism, including the re-ranking operations. Section 6 discusses the experimental results, and Section 7 summarizes the conclusions.

## 2 Patient Cohort Retrieval System Architecture

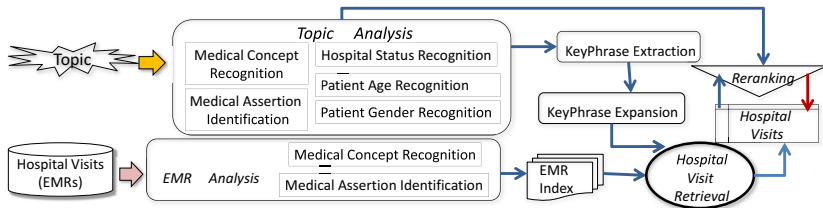


Fig. 1. Patient Cohort Retrieval System Architecture

The architecture of our system is illustrated in Figure 1. Both topics and the EMRs are analysed to identify medical concepts and their assertions. Because topics convey multiple semantic constraints, topic analysis aims to recognize additional semantic classes that are specific to patients, e.g. their age, gender and hospital status. Special submodules of the topic analysis distill the patient age (e.g. *elderly*, *children*), patient gender (e.g. *women*, *male patients*), hospital status (e.g. *presenting to the emergency room*, *discharged from the hospital*, *admitted with*), or medical assertion<sup>1</sup> status which captures the existence, absence or uncertainty of medical phenomena (e.g. *without a diagnosis of x*, *family history of x*, *recommended for possible x*). Because retrieval relies on topics that can be represented as medical keyphrases, a special module identifies such phrases and passes them to a keyphrase expansion module which employs several semantic relations such as synonyms, hyponyms, and meronyms to generate additional keyphrases. Because of the incredible diversity of medical knowledge expressed in the EMRs, each keyphrase is also expanded using the following knowledge sources: (1) the Unified Medical Language System (UMLS) Metathesaurus, (2) the English Wikipedia redirect database, (3) the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) 2011, and (4) co-occurrence information from PubMed Central. This form of semantically-enhanced query expansion attempts to improve the precision and recall of hospital visit retrieval. The ad-hoc retrieval uses an EMR index produced through the Apache Lucene 4.0

<sup>1</sup> A useful description of medical assertions is provided in [8].

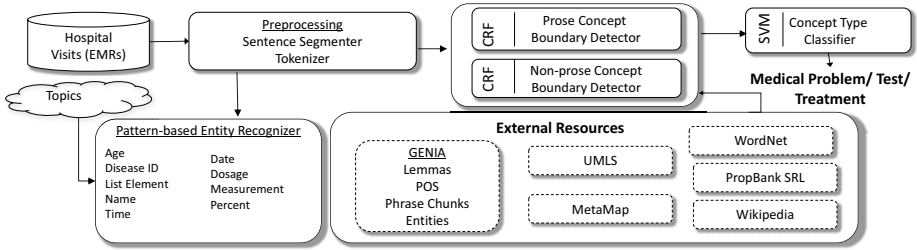
retrieval engine. To enhance the quality of retrieval, several forms of re-ranking were used, using the knowledge gleaned from topic analysis to yield the final ranking.

### 3 Topic and Electronic Medical Record Analysis

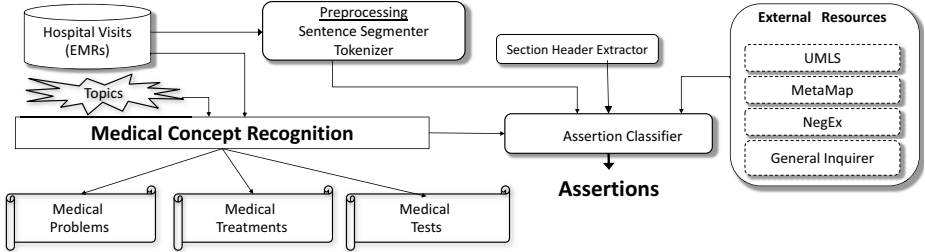
Retrieval of patient cohorts relevant to a topic query depends on the ability to automatically identify three types of medical concepts: (1) medical problems, (2) medical treatments, and (3) medical tests. Moreover, we claim, in this paper, that relevance for patient cohort retrieval is better modeled when assertions of these three types of concepts is also taken into account. Thus, the analysis of topics and of the EMRs consists of the automatic identification of medical concepts and their assertions. To do so, we have used the medical concept recognition techniques reported in [8] and extended the assertion identification available in that research to compute six additional values which we found important for our task. For this purpose, we used the framework that was created for the 2010 i2b2 challenge, which provided 22,846 medical concepts annotated by medical professionals. Concept extraction was cast as a classification task, in which two conditional random field (CRF) classifiers were used to detect medical concept boundaries within the narrative or the non-narrative parts of EMRs. A third, SVM-based, classifier was trained to distinguish between medical problems, treatments, and tests. The features we used have been reported in [8] and provide state-of-the-art results. The architecture of the medical concept recognition framework is illustrated in Figure 2a.

#### 3.1 Medical Assertion Detection

Unlike the 2010 i2b2 challenge, we considered assertions for all medical concepts, not merely medical problems. The belief status of a medical concept is determined by a single SVM classifier, as illustrated in Figure 2b. The SVM classifier uses several external resources, including the Unified Medical Language System (UMLS) [9] ontology, the Metamap [1] information extraction system, and the Negex [2] negation detection tool. Besides the extension of assertions to all medical concepts, we also added six additional belief values: CONDUCTED, HISTORICAL, ONGOING, ORDERED, PRESCRIBED, and SUGGESTED. The belief values which were tested in the i2b2 2010 challenge are: ABSENT, ASSOCIATED WITH ANOTHER, CONDITIONAL, HYPOTHETICAL, POSSIBLE, PRESENT. To be able to re-train our assertion identification, we manually annotated the assertion status of 2349 medical concepts (1183 problems, 614 tests, 552 treatments). We used a Support Vector Machine (SVM) to train on these annotations with the aim of classifying the status of each concept identified in a given topic. The resulting assertion detection technique follows the methods described in [8], modified to perform 12-way classification to support our new classes of assertions. Assertions were identified both within the EMRs and within the topics.



(a) Medical concept identification



(b) Medical assertion classification

Fig. 2. System architectures for medical concept and assertion identification

### 3.2 Discovering Hospital Status

We observed that there were three criteria that occurred frequently throughout the 2011 and NLM practice topics: where the report concerned the patient’s AD-MISSION, the patient’s DISCHARGE, or EMERGENCY ROOM. The desired hospital status was detected by comparing the lemmatized topic against a small set of simple patterns, described in the following examples.

*Example 1 (Hospital Admission).* Consider the topic “Patients admitted with a diagnosis of multiple sclerosis.” In order for a document to be considered relevant, it must fall within the context of patient admission. We detect this AD-MISSION state by checking for the following lexical patterns with a lemmatized topic: *admit for*, *admit to the hospital for*, or *present to the hospital*.

*Example 2 (Hospital Discharge).* Another significant state with a patient’s hos-pital stay is that of their discharge from the hospital. Their status during this state typically includes their final diagnosis and medications. Imagine the topic “Patients being discharged from the hospital on hemodialysis.” This clearly refers to patient cohorts wherein the patient is being dismissed from the hospital. We detect topics pertaining to the DISCHARGE state by checking if they contain the lemma *discharge*.

*Example 3 (Emergency Room).* Finally, suppose one is asked to retrieve infor-mation regarding “Patients with CAD who presented to the Emergency De-partment with Acute Coronary Syndrome and were given Plavix.” It is critical

that relevant documents contain information indicating that ACS (Acute Coronary Syndrome) was diagnosed within the Emergency Department rather than later during their stay. To that end, we classify such topics as pertaining to the EMERGENCY ROOM if they contain any of the following lemmas: *Emergency Department*, *ED course*, or *emergency room*.

### 3.3 Age and Gender Detection

Although somewhat rare, some topics targeted patients characterized by a specific age, or age range (such as topic 119 in table 1 which targets adult patients only). Patient age information is detected according to manually created grammar extrapolated from the sixty practice topics provided by the National Library of Medicine. Our grammar is described in detail in [6] captures topics of the form *patients younger than x*, *patients at most x years old*, as well as ranges such as *patients in their thirties to sixties*. We also detect common age ranges based on a lexicon of known phrases, such as *children*, *elderly*, *adult* have been manually mapped to their numerical ranges. Additionally, some topics target specific patient genders. We capture any gender constraints by detecting the presence of terms from a lexicon of common gender words (e.g. "male", "female").

### 3.4 Keyphrase Extraction

The topics presented in the TREC 2011 and 2012 medical record track target specific patient cohorts: groups of people constrained by specific medical problems, treatments, or tests. As such, we must detect these constraints – which we cast as keyphrases.

Because medical phenomena are often represented through multi-token, complex nominal phrases, our keyword extraction considers multi-word expressions that preserve the semantics encoded by the syntactic structure of the topic. Consider, for example, the major phenomena – keywords – extracted from the topics given in table 1: topic 104 contains *localized prostate cancer* and *treated with robotic surgery*. This requires determining which token sequences constitute a keyword, and which sequences should be decomposed into separate keywords.

To address this dilemma, we recursively consider all sub-sequences of tokens from each topic and check if that sequence corresponds to an article title in Wikipedia. This allows us to capture virtually any medical concept as well as common abbreviations, misspellings, short-hand, phrasal verbs, noun collocations and synonyms. However, many common phrases and stopwords exist as Wikipedia articles. To combat this, we ensure that any matched sequence occurs less than a threshold,  $\lambda^2$ , within the PubMed Central open access subset of biomedical text<sup>3</sup>.

<sup>2</sup> In our case,  $\lambda = 30,000$ . This was based on observed occurrences of keywords from the TREC 2011 topics.

<sup>3</sup> It is our belief that by using a biomedical corpus, we can more accurately target domain-specific keywords and filter domain-specific stopwords.



## 4 Keyphrase Expansion

Within natural language, particularly within medical records, the morphology of words varies extraordinarily both within and between medical texts. To mitigate this diversity of diction, we expand each keyword so as that it may match a variety of lexical forms encompassing synonymy, metonymy, and hyponymy as described in [6]. In order to ease slight variation in syntax, the following simple keyword expansions are produced: (1) a WordNet [5] lemmatized form, (2) an unabbreviated form based on an internal list of common medical abbreviations, (3 - 4) forms in which all hyphens are padded or replaced by spaces, and (5) a form in which all punctuation is removed.

Simple surface form variations are not enough to capture the range of terms doctors use to describe their patients' conditions. For example, consider the term *stroke*. This phrase may be referred to as *apoplexy*, *brain attack*, or *cerebrovascular accident*. In order to capture this degree of synonymy, we utilize the Unified Medical Language System (UMLS) Metathesaurus [9], which is a medical ontology aggregating knowledge from RxNorm, MeSH, SNOMED and other sources. We utilize this knowledge by expanding a given keyword so that it also matches all lexical forms which map to the same CONCEPT ID within the UMLS Metathesaurus database.

Despite the high precision achieved by incorporating knowledge from UMLS Metathesaurus, the recall was not sufficient for our needs. The terms used in the electronic medical records contained spelling variations and a wide variety of slang or less precise synonymy than UMLS encodes. To bridge this knowledge gap, we leveraged the English version of Wikipedia. We used a list of all redirect articles – pages that send the reader to a new article rather than containing information on their own. These redirect articles suite our needs because they typically correspond to alternate names, spellings, lexical forms, related words, or hyponyms. We use this information by expanding a given keyword such that it corresponds to any lexical forms used as article titles that redirect to the given keyword. For example, using Wikipedia redirects expansions allows us to expand the keyword *hearing loss* to *auditory impairment*, *deaf*, *deafness*, *hard of hearing*, *hearing damage*.

While synonymy and alternations are sufficient for many keyword matches, some questions are constrained by information that requires greater reasoning. Consider, for example, the keyword, *atypical antipsychotics*. Doctors will not use this phrase as-is in their records, but rather, will use hypernyms or meronyms – specific types of atypical antipsychotics in its place. In order to match this kind of variation, we incorporated the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), an ontology of clinical terms and, more importantly, the relationships between them. We incorporate this knowledge by expanding a given keyword so as to match any lexical form encoded in SNOMED CT that partakes in the child side of an IS\_A, PART\_OF, or COMPONENT relationship. By doing so, the keyword *atypical antipsychotics* may be expanded to include *abilify*, *aripiprazole*, *asenapine*, *clozapine*, *clozaril*.

While the previous keyword expansion techniques are sufficient for most scenarios, the text of electronic medical records is often terse, disjoint, and ungrammatical. Additionally, some keywords may require more domain knowledge than what we are able to simulate with mere keyword expansion. As a fall-back, to help mitigate this domain knowledge rift, we expand keywords so that they correspond to related terms. We calculate these related terms using co-occurrence information gleaned from the PubMed Central Open Access Subset (PMC), a collection of freely available biomedical texts. Related was determined by considering the normalized Google distance [3]. The normalized Google distance, or NGD, is defined below:

$$\text{NGD}(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

where  $M$  is the total number of documents in PMC;  $f(x)$  and  $f(y)$  are the number of documents containing terms  $x$  and  $y$ , respectively; and  $f(x, y)$  is the number of documents in which  $x$  and  $y$  co-occur.

We selected the top twenty expansions of sufficient similarity<sup>4</sup> as the expansions for each keyword. For example, *atypical antipsychotics* acquired *olanzapine*, *risperidone*, *quetiapine*, *clozapine*, and *antipsychotic drug* as expansions by using the NGD.

## 5 Hospital Visit Retrieval

After extracting and expanding the keywords that characterize a patient cohort, we must retrieve all relevant hospital visits that match the extracted keywords. This task is accomplished through the use of Apache Lucene 4.0 [7].

Prior to retrieval, we created an index over all hospital visits by merging all the electronic medical records associated with each hospital visit into a single document. The various fields encoded in each EMR were retained when indexed (admit diagnosis, chief complaint, etc.) so that per-field weights could be adjusted.

For retrieval, each topic is represented as an interpolation of its weighted expansions, and those of any subsumed keywords. For example, the keyword *chronic wound* would also include the weighted expansions for the keyword *wound*. More precisely, a topic's is converted to a "query" as follows:

$$\text{query}(k, \lambda) = \lambda [\alpha \text{UMLS}(k) + \beta \text{Wikipedia}(k) + \gamma \text{SNOMED}(k) + \delta \text{Co-Occurrence}(k)] + \sum_{s \in S} \text{query}(s, \mu \lambda)$$

where  $\lambda$  is the initial keyword score;  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$ , are the weights associated with the respective keyword expansion method;  $S$  is the set of keywords subsumed by  $k$ ; and  $\mu$  is the discounting factor such that  $0 < \mu < 1$ . In our experiments, we set  $\lambda = 16$ ,  $\alpha = 12$ ,  $\beta = 10$ ,  $\gamma = 8$ ,  $\delta = 1$ , and  $\mu = 0.5$ . These weighted expansions were then scored using the highly popular BM25 ranking function. This yields a ranked list of hospital visits, ordered by BM25's interpretation of our query representation.

<sup>4</sup> See [6] for more information regarding co-occurrence keyword expansion.

Although our initial Lucene retrieval performs reasonably well for the purpose of ranking documents strictly within respect to keyword relevancy, the topics presented in TRECmed are characterized by more complex constraints. We address these additional cohort constraints by an iterative re-ranking process: for each constraint identified by the TOPIC ANALYSIS module (patient age, patient gender, hospital status, medical assertion value), we heuristically re-rank all hospital visits for a given question. After each constraint has been considered, the final ranking of patient hospital visits is returned as the solution of our system. What follows is a description of each heuristic re-ranking sub-module.

### 5.1 Re-ranking According to Assertion Information

As described in section 3.1, each keyphrase in a given topic, if identified to be a medical concept, is automatically associated with its corresponding medical assertion. For the purposes of re-ranking, we used a scale of "negativity" indicators which we associated with each belief value (from the automatically generated assertions). These indicators' values were empirically assigned in the following way: ABSENT and ASSOCIATED WITH ANOTHER were given a negativity indicator value of 1.0 (the highest); HISTORICAL was indicated as 0.5; CONDITIONAL, HYPOTHETICAL, POSSIBLE, and SUGGESTED as 0.3; and all other belief values were indicated with a negativity score of 0.0 (suggesting that they do not indicate any negative belief). These negativity scores model our attempt to ascertain the degree to which mentions of a given medical concept correctly indicate an actually present medical condition, treatment, or test as opposed to an absent or unsure mention.

The negativity scores were then computed for all medical concepts identified throughout the EMR corpus: this entailed that a medical concept that was associated with different assertions received negativity indicators pertaining to each the assertion corresponding to each mention. This allowed us to compute all negativity scores from all EMRs associated with a given hospital visit and obtain a negativity score for the visit as the sum of all negativity scores inferred from that visit's EMRs. When the negativity score of the hospital visit had a value large than one third of the frequency of any of the medical concepts mentioned in the hospital visit, we adjusted the current relevance score of the visit (computed with the BM25 score illustrated in section 5) by subtracting the value 400 from the current relevance score.

### 5.2 Re-ranking Based on the Patient's Hospital Status

The goal of the hospital status re-ranker is to promote hospital visits wherein at least one EMR that matches a keyphrase from the topic also satisfies the requirements of the patient's hospital status detected in section 3.2. In order to achieve this, we consider the meta-data associated with each EMR (the *type* and *subtype* fields which indicate the type of each electronic medical report), as well as context for each keyword match: the previous section header, based on a simple section detection algorithm that looks for the last fully capitalized sentence

ending with a colon (e.g. *DISCHARGE SUMMARY:*), and the lemmatized sentence containing the given keyword. For example, when detecting hospital visits that satisfy patient admission criteria, we look for EMRs that have the *subtype* of *ADMISSION*, or keywords that fall within a section whose header contains *ADMISSION* or *ADMITTING* or whose lemmatized sentence contains *admit for*, *admit to the hospital for*, or *present to the hospital*. Likewise, the criteria for detecting patients discharged from the hospital in an EMR with the *type* of *DS* or *subtype* or *DISCHARGE* or any sentence containing the lemma *discharge* used as a verb. Finally, the requirements for asserting EMRs pertaining to the emergency room involves checking if the EMR’s *type* is *ER*, if any keyword’s section header contains *EMERGENCY DEPARTMENT* or *ED*, or if any keyword match lies within a lemmatized sentence containing *Emergency Department*, *ED course*, or *emergency room*. Visits wherein at least one EMR did not satisfy the requirements of any detected patient hospital status constraints have their score lowered by 50.

### 5.3 Re-ranking by the Patient’s Age and Gender

The current ranked list of hospital visits are re-ranked with respect to patient age by comparing the frequencies of de-identified patient age information within all the reports associated with each hospital visit. Any hospital visit wherein the number of de-identified age mentions falling outside the numerical range identified by the *TOPIC ANALYSIS* module (described in section 3.3) has its score lowered by 100 where a hospital visit’s score is based on the *BM25* score described in section 5; any hospital visit lacking any age information has its score lowered by 50 (so that hospital visits that match the desired criteria are elevated to the top).

When considering the patient’s gender, we utilize the same lexicons described in section 3.3 and compare the frequency of *MALE* to *FEMALE* words in all EMRs associated with a given hospital visit. Hospital visits for which there are more mentions of the opposite gender across all associated EMRs have their current score lowered by 100 (where score is the *BM25* score detailed in section 5).

**Table 2.** TRECmed evaluation results: infAP refers to the inferred average precision, NDCG refers to the normalized discounted cumulative gain, and P@10 refers to the precision of the first 10 retrieved EMRs

Approach	infAP	NDCG	P@10	Approach	infAP	NDCG	P@10	Expansion Assertions
NONE	0.302	0.467	0.468	NONE	0.254	0.518	0.447	
+UMLS	0.157	0.297	0.274	+UMLS	0.170	0.221	0.324	
+WIKI	0.320	0.499	0.485	+WIKI	0.283	0.350	0.440	✗ ✗ 0.112 0.398 0.145
+SNOMED	0.317	0.499	0.462	+SNOMED	0.270	0.340	0.428	✓ ✗ 0.204 0.425 0.445
+NGD	0.363	0.561	0.509	+NGD	0.266	0.334	0.428	✗ ✓ 0.280 0.535 0.445
								✓ ✓ 0.292 0.538 0.445

(a) Query expansion results for TRECmed 2011

(b) Query expansion results for TRECmed 2012

(c) Query expansion and assertion re-ranking results for TRECmed 2012.

## 6 Performance Evaluation

We participated in the 2011 and 2012 Text REtrieval Conference (TREC) medical records task, TREC<sub>Med</sub>. The task for this evaluation was to return a ranked list of hospital visits corresponding to a given topic (a patient cohort). A hospital visit is defined as the set of electronic medical records (EMRs) generated during a patient's single visit to the hospital. To be able to produce a ranked list of hospital visits, we merged all EMRs corresponding to a visit into a single virtual document.

In order to evaluate the impact of medical knowledge on our patient cohort retrieval system, we first analyzed the impact of each of the keyphrase expansion methods that we have considered. Table 2 presents the impact of each adding each component to a baseline system of no key-phrase expansion nor any re-ranking techniques. It is clear that query expansion yields improved accuracy for overall retrieval. That said, it is interesting to note that some techniques actually hindered our overall performance (e.g. UMLS and SNOMED) in this evaluation. This is likely due to the nature of these ontologies which were not designed with the primary goal of aiding natural language processing. Tuning these ontologies, particularly UMLS, by removing highly ambiguous terms would likely benefit future work. Regardless, the value of Wikipedia redirect expansion and PubMed central co-occurrence information (NGD) is obvious.

Keyphrase expansion was not the only strategy for improving the quality of patient cohort retrieval that we pursued. The re-ranking methods were also designed for the reason. But, more interestingly, we wanted to evaluate the quality of retrieval when both strategies were combined. Table 2c shows the effect of combining query expansion with hospital visit re-ranking based on medical assertions. These evaluations are from the testing data used for TREC<sub>Med</sub> 2012. It is to be noted that the other re-ranking techniques were available to the original system. It is clear, from this table, that both query expansion and incorporating assertion knowledge significantly improve the relevancy of returned documents. By incorporating assertion information alone (-EXPANSION +ASSERTIONS), our inferred average precision improved by 149.1%, NDCG improved by 36.5%, and our precision within the first ten documents improved by 207.3%. However, the impact of including both components yields diminishing effects. This is likely due to the recall-oriented nature of query expansion introducing too much noise for our heuristic-based assertion re-ranking model to handle. Additionally, it should be noted that the effects of patient gender, age, and hospital status were negligible due to the absence of these kinds of constraints in the test topics.

## 7 Conclusion

In this paper we have described a model for ranking electronic medical records (EMRs) for the purpose of patient cohort retrieval. By incorporating the physician's belief values corresponding to medical concepts – the degree to which something is present, uncertain, or absent – we are able to better model the

relevancy of an EMR. Additionally, we presented four methods for performing keyphrase expansion useful for the task of retrieving EMRs relevant to patient cohorts. We showed that, when applied to the 2011 and 2012 Text REtrieval Conference (TREC) medical records task (TREC Med), our model yields significantly improved performance to a baseline Lucene BM25 retrieval model. We approached this task by extracting the constraints encoded by a given cohort (patient's age, patient's gender, patient's hospitalization status, and keyword assertion status) and the keywords that encode any medical phenomena found in the topic. These keywords were then expanded using knowledge from UMLS, SNOMED, and Wikipedia, as well as PubMed Central co-occurrence information. We then perform retrieval to achieve an initial ranking of hospital visits (based on a BM25 relevance model). Next, the EMRs are re-ranked to account for (1) any constraints on the patients age, gender or hospital status, and (2) to ensure that the belief value corresponding to each keyphrase mention supports the belief values present in the topic.

The incorporation of belief value of a medical concepts helps to better capture the semantics encoded within the narrative of an EMR. This type of knowledge could be of significant value to future language processing applications within the medical domain, as it allows one to move beyond the presence of a keyphrase and ascertain the actual motivation behind its usage in the text.

## References

1. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metemap program. In: Proceedings of the AMIA Symposium, p. 17. American Medical Informatics Association (2001)
2. Chapman, W., Bridewell, W., Hanbury, P., Cooper, G., Buchanan, B.: A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics* 34(5), 301–310 (2001)
3. Cilibrasi, R., Vitanyi, P.: The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 370–383 (2007)
4. Committee on Comparative Effectiveness Research Prioritization, Institute of Medicine (US): Initial national priorities for comparative effectiveness research. National Academies Press (2009)
5. Fellbaum, C.: *WordNet: An Electronic Lexical Database*. The MIT press (1998)
6. Goodwin, T., Rink, B., Roberts, K., Harabagiu, S.: Cohort shepherd: Discovering cohort traits from hospital visits. In: The Twentieth Text REtrieval Conference Proceedings, TREC 2011 (2011)
7. Hatcher, E., Gospodnetic, O.: *Lucene in Action*. Manning Publications (2005)
8. Roberts, K., Harabagiu, S.: A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association* 18(5), 568–573 (2011)
9. Schuyler, P., Hole, W., Tuttle, M., Sherertz, D.: The umls metathesaurus: Representing different views of biomedical concepts. *Bulletin of the Medical Library Association* 81(2), 217 (1993)
10. Voorhees, E., Tong, R.: Overview of the trec 2011 medical records track. In: The Twentieth Text REtrieval Conference Proceedings (TREC 2011). National Institute for Standards and Technology, Gaithersburg (2011)

# Semantic Discovery of Resources in Cloud-Based PACS/RIS Systems

Rafael Berlanga<sup>1</sup>, María Pérez<sup>1</sup>, Lledó Museros<sup>1</sup>, and Rafael Forcada<sup>2</sup>

<sup>1</sup> Universitat Jaume I, Castellón, Spain

{berlanga,mcatalan}@lsi.uji.es, museros@icc.uji.es

<sup>2</sup> ActualMed, Castellón, Spain

rafael.forcada@actualmed.com

**Abstract.** PACS/RIS systems store a huge volume of clinical data that are mostly accessed by the patient identifier. However, clinicians would like to retrieve information about similar clinical cases. In this paper, we claim that the semantics-based technology could improve the discovery and integration of information in this type of systems. We propose a semantic approach that semantically annotates the clinical information and retrieves the resources relevant to the clinician's query, independently of their language and format. Moreover, cloud-based systems allow the integration with external resources. In this paper, we present preliminary results that show that current semantic technologies can produce good enough results to perform classification and retrieval tasks.

## 1 Introduction

The irruption of cloud-based architectures in e-Health is challenging the current technology of PACS (Picture Archiving and Communication Systems)/RIS (Radiology Information Systems) with new issues related to massiveness, multi-modality and multi-linguality of the data they should support [8]. PACS/RIS cloud-based systems store hundred of thousands of medical reports and images with different formats. This scenario opens new challenges and opportunities to effectively exploit all these data. For example, clinicians would like to retrieve similar clinical cases from the cloud independently of their language and format. It is worth noticing that reports and image metadata are usually written in the local language of the clinicians, and that scientific publications are massively written in English.

As described by [8], the main requirements of radiologists about these systems are: *(i)* more precise user's requirements specification, e.g., query by text and image, or by more specific fields such as pathology and modality, *(ii)* multilingual and multimodal retrieval, *(iii)* automatic classification of cases, *(iv)* semantic retrieval, and *(v)* linking of results to external resources.

In this paper, we claim that semantics-based technology can greatly help to address these requirements by developing new retrieval mechanisms for the new cloud-based systems. Specifically, we assume that current e-Health knowledge

resources (e.g., MeSH, UMLS, etc.) are comprehensive enough to perform the automatic semantic annotation of all the items stored in a cloud-based PACS/RIS. Semantic annotation allows us to summarize the knowledge stored in clinical reports as well as images in terms of a reference ontology which is independent of the language. This allows the system to perform searches independently of the language, as well as to present the results in the target language of the users. Moreover, external queries can be forwarded to other information systems such as PubMed, which also rely on similar knowledge resources (e.g., MeSH) to find related literature to each clinical case.

For this purpose, this paper presents a joint project developed by the Temporal Knowledge Bases Group (TKBG) and the ActualMed corporation to evaluate the effectiveness and usefulness of semantic retrieval in a cloud-based PACS/RIS. In next sections we describe the infrastructure, the goals, and the preliminary results we achieved.

## 2 Cloud-Based PACS/RIS

The use of digital radiographic images has been extended and accepted by the radiology community. The definition of the standard DICOM (Digital Imaging and Communications in Medicine), a uniform and robust standard for the exchange and storage of medical digital images, has improved the integration and processing of medical images in PACS/RIS in healthcare systems. PACS enables image communications between individual components such as archive systems, diagnostic workstations, postprocessing workstations, and image distribution workplaces. And, typically, a RIS comprises a series of software modules supporting radiology workflow such as creation of orders, scheduling, reading, reporting, medical coding, recording of services, and interfaces to a billing system. Figure 1 shows a typical PACS/RIS configuration. PACS and RIS represent the main information technology (IT) of a radiology department. Both systems are jointly deployed and closely integrated.

Because the use of images is expanding, the technology is ever evolving, and interdisciplinary collaboration over the Internet is in demand, as hospitals are starting to replace their initial PACS system to keep ahead of demand. First the images and reports were moved over networks, outside the radiology department and directly into the hands of the clinical staff. Then, within a local firewall, images were exposed on local workstations through PACS client applications or Web viewers, and, in the last few years, VPN (Virtual Private Network) solutions have emerged which give the clinicians the possibility to access the images outside the local firewall. This is a good solution for physicians who are members of a system, but not for those who are part of a separate organization. Therefore, nowadays, there is a new trend of using cloud computing for medical images, which will allow the delivery of better, more secure and less expensive medical imaging services. Cloud computing refers to a provisioning model for virtualized processing and storage capacity [4]. The physical processors and storage systems are housed in large data centers, usually widely distributed, and managed by



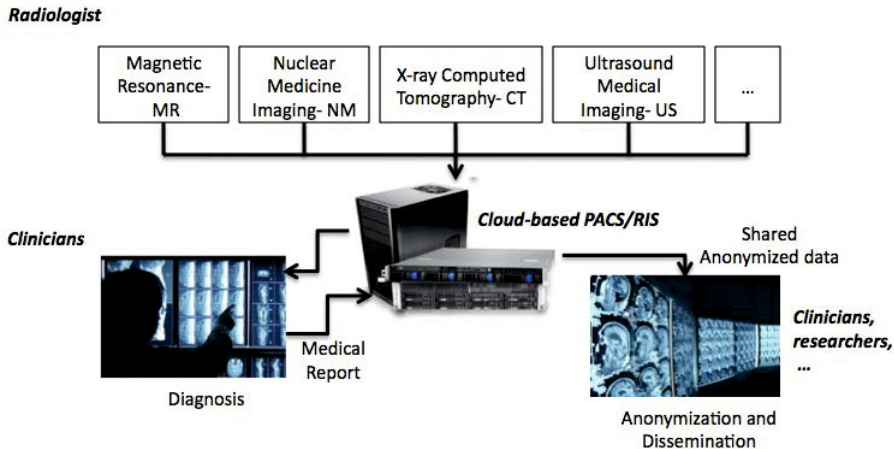


Fig. 1. Typical PACS/RIS configuration

professional IT organizations. All functionality is provided by web services accessed over the Internet. Several recent works introducing cloud computing in medical applications have been developed [24][20][9]. [24] proposed a new framework based on cloud computing for cancer imaging research; [20] proposed cloud method for gathering patient information, allowing also information distribution and remote access by medical staff; and [9] designed a complete HIS based on cloud computing. An analysis of the advantages and disadvantages of using cloud computing for biomedical applications can be found in [21], and specifically for PACS/RIS in [16]. In summary, this shows that the cloud computing model is clearly attractive for medical imaging, and specifically for PACS/RIS.

Our proposal consists of the extension of the PACS/RIS system *ActualMed PACS*<sup>1</sup> with a set of new services for the semantic retrieval and integration of resources stored in the system and external resources relevant for clinicians, like PubMed. These services annotate semantically the resources and perform semantic-based retrieval, whatever their type (images, clinical reports, articles and so on) and their language. The architecture of our proposal is shown in Figure 2.

### 3 Automatic Semantic Annotation

Clinical images such as DICOM images have metadata that describe technical features of the image as well as information about the patient, the part of the body the image is about, and technician’s comments among others. Most clinical reports derived from these images are written in free-text with just a few structured fields such as the date, patient identifier, and so on. Lately, there has been

<sup>1</sup> [http://www.actualmed.com/es/actualmed\\_pacs.html](http://www.actualmed.com/es/actualmed_pacs.html)

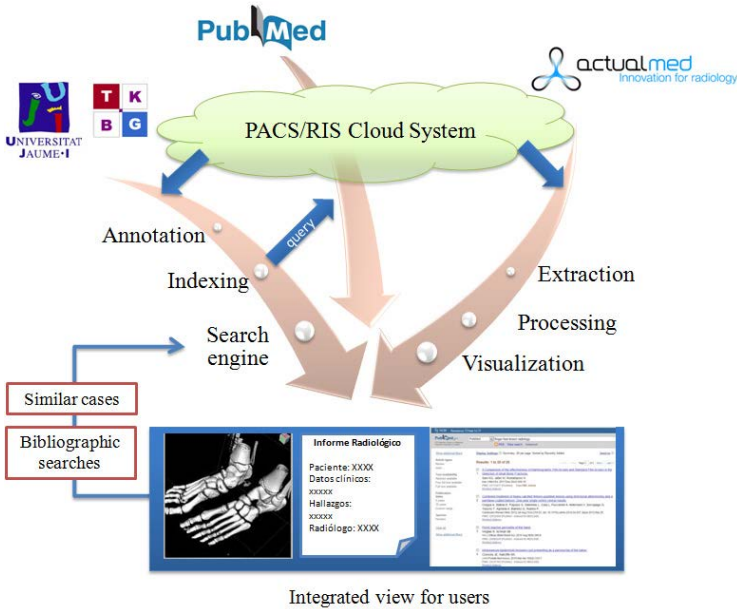


Fig. 2. Proposed architecture for the semantic-based PACS/RIS in the cloud

a tendency to normalize these reports using DICOM Structured Reporting [5]. However, the fixed metadata structure limits the users in the specification of the information, and frequently relevant metadata are not provided. Current PACS/RIS systems only provide searches by the patient identifier and simple keyword searches over some data fields. Therefore, these systems makes the multi-lingual retrieval of data infeasible.

In this paper, we propose the use of automatic semantic annotation to process and integrate all the data stored in a cloud-based PACS/RIS system. Semantic annotation consists of finding out mappings between text chunks identified in a text and concepts described in a knowledge resource (KR). A KR is usually expressed as either an ontology or thesaurus. Nowadays there exist several widely-accepted KRs in the biomedical domain, like MeSH<sup>2</sup> and UMLS<sup>3</sup>, as well as some specific KRs for radiology, like RadLex<sup>4</sup>. In this project, we are specially interested in KRs that provides lexica for different languages, e.g., UMLS and RadLex.

Currently, there are several tools to semantically annotate biomedical texts, e.g., BioPortal [13], MetaMap [1], and Whatizit [17]. However, these annotators do not cover simultaneously all the terminologies used in a PACS/RIS system and, moreover, they only provide annotations for English. With respect to the

<sup>2</sup> <http://www.nlm.nih.gov/mesh/>

<sup>3</sup> <http://www.nlm.nih.gov/research/umls/>

<sup>4</sup> <http://www.radlex.org/>

semantic annotation of images, there are tools to retrieve and annotate DICOM images, e.g., RadSem [10] or [23,14], which use the ontology RadLex.

In this work, we use an automatic and unsupervised semantic annotator [3] that can support multiple languages, and it can use several knowledge resources in order to increase the vocabulary coverage. This annotation tool was tested within CALBC competition over a collection of around one million PubMed abstracts about immunology [18] using UMLS as KR. It is based on concept retrieval, that is, it finds the most relevant concepts w.r.t. the text words and, then, selects those that best cover the underlying text semantics.

The semantic annotation of a text chunk  $T$  consists of the concepts that best cover its semantics. The information coverage of  $T$  with the concept  $c$  is estimated as follows:

$$sim(c, T) = \max_{S \in lex(c)} \frac{info(S \cap T) - info(S - T)}{info(S)} \quad (1)$$

The function  $lex(c)$  returns the lexical strings associated with the concept  $c$  in the KR. The function  $info(S)$  measures the information coverage of the string  $S$  with an estimation of the words entropy in a background corpus  $\mathcal{G}$  (e.g., Wikipedia).

$$info(S) = - \sum_{w \in S} \log(P(w|\mathcal{G})) \quad (2)$$

All these definitions are inspired by the information-theoretic matching function presented in [11] and the word content evidence defined in [6].

The set of annotations associated to each text chunk  $T$  are those concepts that maximize both  $sim(c, T)$  and the word coverage of  $T$ . That is, the system selects the top ranked concepts whose lexical variants best cover the text chunk  $T$ . In order to avoid spurious and incomplete annotations, a minimum threshold for  $sim(c, T)$  is required (usually above 0.7).

For example, given the text chunk “Conclusión: Discreto derrame articular y focos de edema óseo”, its semantic annotation is:

“Conclusión: Discreto  $\langle_{C125396}$  derrame articular  $\rangle$  y  $\langle_{C2609134}$  focos de edema óseo  $\rangle$ ”

The semantic annotation will be applied to all the data in the PACS/RIS system suitable for querying, i.e., the images metadata and clinical reports, and also external resources that can be relevant to the clinicians, e.g., scientific publications. Moreover, thanks to the concept relationships provided by the KR (e.g., “is a”), it is possible to explore the repository contents by navigating through the resulting taxonomies. These methods of semantic data access are completely novel in this kind of systems.

## 4 Semantic Retrieval in the PACS/RIS

Data retrieval in a PACS/RIS system could be performed with traditional keyword-based search engines. Some extensions of mono-lingual IR have been

proposed to support cross-lingual queries, also called cross-lingual information retrieval (CLIR). Broadly speaking, these extensions are based on either the automatic translation of the queries to the target languages [7], or on the existence of parallel corpora [19]. These systems usually focus on pair of languages (English and another) and do not consider a scenario with multiple languages. More recently, a query translation proposal based on a multi-lingual KR has been proposed in [22]. As in this proposal, we also claim that the canonical language must be the conceptual space provided by the KR, so that all language expressions must be mapped into it. In [22] a categorization based on MeSH is proposed, whereas we propose to use unsupervised semantic annotation based on any available KR.

Apart from the semantic annotation, semantic retrieval must take into account relevance issues. Not all the annotations have the same relevance for the different user tasks. For example, annotations related to anatomic parts are more relevant for technicians looking for similar images, whereas annotations related to diseases are more relevant for clinicians looking for similar cases. In this project we propose to evaluate different relevance schemes proposed in the literature. More specifically, we will mainly focus on the *tf-idf* scheme, and topic-based language models [15]. The latter one have been shown quite effective in retrieving resources related to different Bioinformatics tasks.

In this proposal, the similarity between two resources  $r_1$  and  $r_2$  (e.g., query-resource, resource-resource) is given by the similarity of their semantic representations  $r_1^{KR}$ ,  $r_2^{KR}$ .

$$\text{sim}(r_1, r_2) \propto \text{sim}(r_1^{KR}, r_2^{KR}) \quad (3)$$

The first relevance scheme we propose is the cosine coefficient in which the similarity of two resources is calculated as follows:

$$\text{sim}(r_1^{KR}, r_2^{KR}) = \frac{r_1^{KR} \cdot r_2^{KR}}{\|r_1^{KR}\| \|r_2^{KR}\|} \quad (4)$$

The second proposed relevance scheme is the topic-based model. In a PAC-S/RIS system, the main topics are about diagnosis, treatment, surgery, and image analysis among others. Given a specific query or a specific resource  $r_1$ , the relevance of a resource  $r_2$  to  $r_1$  is calculated considering the different topics  $T$  defined in the system as follows:

$$\text{sim}(r_1^{KR}, r_2^{KR}) = \prod_{c_i \in r_1^{KR}} \sum_{t_k \in T} p(c_i|t_k) \cdot p(t_k|r_2^{KR}) \quad (5)$$

Both  $p(c|t)$  and  $p(t|r^{KR})$  are given by the topic-based model. More information about how this model is built is given in [15].

Once defined the relevance scheme, the semantic index can be directly built from the semantic annotations generated for the data. This semantic index will be implemented with existing technology based on inverted files [2]. User requests for related data and documents are directly performed over this semantic index independently from the local language of each user. Of course, user free-text

queries should be previously annotated to perform the search over the semantic search.

Additionally, the semantic index can be used to perform external queries to bibliographic resources (e.g., PubMed). This kind of queries are usually performed by clinicians that look up scientific publications to get additional information about diagnosis, medical treatment and surgery. In this case, relevant concepts of the clinical report are expressed in English, and then transformed into a proper query for performing the bibliographic search. In case the KR includes MeSH terms, the generation of a query for PubMed is straightforward since scientific documents are already indexed under this vocabulary.

## 5 Preliminary Results

In order to test the viability of the project, we conducted an experiment with a subset of the reports and images stored in *ActualMed PACS*. We extracted 8088 reports associated with medical images (730 Doppler images, 4320 ecographies, and 4145 MRN (Magnetic Resonance Neurography)) from more than 50000 stored resources (as of April 2013). Moreover, we also extracted the metadata of 5893 DICOM images. All the resources have been previously anonymized in order to preserve the patients identity. Table 1 shows the statistics of these metadata. As it can be noticed, not much information is provided in DICOM files, and the StudyDescription field, which is supposed to be the most descriptive field, usually only specifies the anatomic part that is being described. As KR, we used a subset of the UMLS Metathesaurus (version 2012AB), which contains the lexicon in English and Spanish for a restricted set of entities (e.g., anatomical entities, disease and disorders, and so on).

**Table 1.** Statistics of DICOM files

DICOM field	Frequency
StudyDescription	4458
AnatomicStructure	0
AnatomicRegion	0
BodyPartExamined	1092
TherapyType	0
TherapyDescription	0
InterventionDescription	0
Type of Patient	0
PatientGroupLength	0
Allergies	0
PatientBirthDate	5893
PatientSex	5893
PatientWeight	1852
Total DICOM files	5893

**Table 2.** Main characteristics of the annotations generated for the selected datasets

Report set	Annotations	Annotations Avg. size	Ambiguity	Semantic Vectors	Anatomy	Disorders	Phys. features
Doppler	14991	1.7	0.9%	673	673	673	673
Ecographies	60598	1.5	18.6%	4320	4317	4276	1356
MRN	65358	1.6	10.0%	3094	3093	3094	1157
All	140947	1.6	12.9%	8087	8083	8043	3186

In Table 2 some statistics over the annotated reports are shown. In this table we show the main features of the generated annotations, namely: the number of annotations, the average number of matched words per identified concept (Avg. size), and the percentage of annotations that are ambiguous (i.e., annotations that have more than one entity type associated (vocabulary used in the PACS/RIS system). The average size of the annotations shows that around 30% of the annotations have more than one word, which indicates the precision of the annotations (see examples in Table 3). The proportion of ambiguous annotations is relatively low, and it can be further alleviated by filtering out single-word annotations with low IDF.

In order to measure the coverage of the semantic annotations at document level, we have created a semantic vector for each image report regarding the main facets clinicians use to retrieve them, namely: anatomy, disorders, and physical features (e.g., clinical attributes). Table 2 shows for each kind of image report the number of generated semantic vectors, and their characterization across facets. As it can be noticed, almost all of the vectors have concepts in both the anatomy and disorders facets. Depending on the type of report, we can find more or less concepts about physical attributes. Finally, Table 3 shows the most frequent concepts associated with the selected facets for the characterization of the semantic vectors.

The semantic vectors allow us to analyze the data and to extract relevant information. For example, we can find out the most frequent associations between anatomic parts and disorders. Table 4 shows the most frequent clusters of anatomy and disorders facets in the MRNs of the left knee.

## 6 Issues and Challenges

Preliminary results show that automatic semantic annotation can produce good enough results to perform classification and retrieval tasks over the resulting semantic vectors. In particular, we plan to add the following functionalities to the PACS/RIS according to this new semantic component, namely:

- **Task 1.** To perform the clustering of all the clinical reports according to the semantic vectors. This clustering will be useful to identify groups of similar cases as well as to identify groups of similar images with similar contexts for posterior image analysis.

**Table 3.** Top ranked concepts for semantic vector facets

Anatomy	Disorders	Physical features
body 6791	injuries 3906	sex 1348
lien 1877	malign neoplasm T1 2027	bone densities 1348
spleen 1877	effusion into joint 1746	projection 726
bone 1784	rupture 1618	fluid pressure 655
kidney 1721	abnormal degeneration 1518	liver size 335
biliary tract 1686	abnormal dilation 1244	age 317
bile tract 1686	normal size breast 1148	kidney feature 311
liver 1569	changes nail 1012	kidney size 151
collum femoris 1442	degenerated intervertebral disc 829	body height 48
tendon 1398	bulging 816	normal muscle function 46
abdominal aorta 1365	hernia nucleus pulposus 774	acoustic shadowing 26
lumbar vertebra 1250	metal foreign body in hip 661	filling of bladder 16
lumbar spine 1248	calculoses 657	appearance of anterior chamber 15
conus medullaris 1160	abnormal narrowing 647	edema grade 10
bladder 1119	hepatic steatosis 601	uterus feature 7
hip left 1118	malign neoplasm T2 598	hepatic function 7
Total=1442	Total=1256	Total=95

**Table 4.** Most frequent clusters in left knee MRN

Anatomy	Disorder	Number of Reports
Anterior horn	Abnormal degeneration	65
Entire medial meniscus	Abnormal degeneration	60
Entire lateral meniscus	Rupture	35
Anterior horn	Laceration	31
Anterior horn	Rupture	26
Region of bone	Effusion into joint	26
Bursa	Augmentation of size	21
Bursa	Benign cystic mucinous tumor	20
Entire medial meniscus	Cartilage tear in knee	20
Soft tissues	Dropsy	19
Anterior horn	Cartilage tear in knee	19
Bursa	Effusion into joint	16
Entire lateral meniscus	Abnormal degeneration	16
Articular	Effusion	13
Ligament	Sprain	13
Condyle of femur	Bone edema	13
Ligament	Dropsy	11
Bone structure of tibia	Bone edema	11
Internal oblique	Cartilage tear in knee	10
Soft tissues	Sprain	10

- **Task 2.** To extract interesting patterns from images and the associate meta-data. This task is usually performed by applying some automatic classification method, which need a predefined set of negative and positive training examples. In our case, as clusters are intended to represent classes (e.g., disorders), training examples can be picked up from these clusters.
- **Task 3.** To perform semantic retrieval of cases stored in the PACS/RIS given either a free-text query or a selected case.
- **Task 4.** To perform semantic retrieval outside the PACS/RIS, that is, to fetch queries to external on-line resources such as PubMed, Wikipedia or WikiRadiography<sup>5</sup>.

The implementation of these tasks present several challenges. Firstly, for Tasks 1 and 2 we need to produce more precise annotations that take into account the right sense of the annotation. For example, the sentence “do not present  $\langle_c injuries \rangle$ ” gives a negative sense to the annotation “ $\langle_c injuries \rangle$ ”. Capturing the relationship between semantic annotations can be also relevant for Tasks 3 and 4 as they provide the right intention of the information request. For example, the query “retrieve images related to injuries in the tendon” will require that relevant reports have “injuries” and “tendon” directly related in the text. In other words, semantic retrieval requires to go beyond the classical bag-of-words vision of IR.

## 7 Conclusions

This paper shows preliminary results on how semantic annotation and semantic-based retrieval can improve the effectiveness and usefulness of a cloud-based PACS/RIS. In fact, we have described how *ActualMed PACS* can be extended by using automatic semantic annotation and semantic retrieval for processing and integrating the data stored in a cloud-based system.

The use of semantics allows to support multiple languages and, therefore, user’s requests can be done independently from the local language. Moreover, the search and integration of scientific publications and other external resources about diagnosis, medical treatment and surgery are also possible by semantically annotating those resources. It is also important to remark that with this semantic infrastructure, the repository contents can be explored by navigating through the constructed taxonomies, which is a new method of semantic data access.

Considering the experiments of this paper as a proof of concept of the use of semantics in the PACS/RIS systems, as future work we aim to extend the cloud-based *ActualMed PACS* system with the following set of new functionalities:

- Semantic annotation of reports and radiographic images.
- Semantic indexing of resources in the PACS/RIS.
- Semantic search of radiologic resources similar to a given one.

---

<sup>5</sup> <http://www.wikiradiography.com/>



- Bibliographic search related with a radiology resource.
- Integration and visualization of all the resources.

With respect to the validation of our techniques, we aim to evaluate them with the datasets provided by ImageClef<sup>6</sup>, and make a comparison with the results of other techniques presented in [12].

## References

1. Aronson, A.R., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17(3), 229–236 (2010)
2. Badue, C., Baeza-yates, R., Ribeiro-neto, B., Ziviani, N.: Distributed query processing using partitioned inverted files. In: *Proceedings of 9th String Processing and Information Retrieval Symposium (SPIRE)*, pp. 10–20. IEEE CS Press (2001)
3. Berlanga, R., Nebot, V., Jimenez, E.: Semantic annotation of biomedical texts through concept retrieval. In: *Workshop on Language Technology Applied to Biomedical and Health Documents, BioSEPLN* (2010)
4. Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Brandic, I.: Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems* 25(6), 599–616 (2009), <http://dx.doi.org/10.1016/j.future.2008.12.001>
5. Clunie, D.A.: *DICOM Structured Reporting*. PixelMed Publishing (2000)
6. Couto, F.M., Silva, M.J., Coutinho, P.: Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics* 6(S-1) (2005)
7. Davis, M.W., Ogden, W.C.: Free Resources and Advanced Alignment For Cross-Language Text Retrieval. In: *Text Retrieval Conference (TREC)*, pp. 385–395 (1997)
8. Markonis, D., Holzer, M., Dungs, S., Vargas, A., Langs, G., Kriewel, S., Müller, H.: A survey on visual information search behavior and requirements of radiologists. *Methods of Information in Medicine* 51(6), 539–548 (2012)
9. He, C., Jin, X., Zhao, Z., Xiang, T.: A cloud computing solution for hospital information system. In: *IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS)*, vol. 2, pp. 517–520 (2010)
10. Möller, M., Regel, S., Sintek, M.: RadSem: Semantic Annotation and Retrieval for Medical Images. In: Aroyo, L., et al. (eds.) *ESWC 2009*. LNCS, vol. 5554, pp. 21–35. Springer, Heidelberg (2009)
11. Mottaz, A., Yip, Y.L., Ruch, P., Veuthey, A.L.: Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics* 9(S-5) (2008)
12. Müller, H., de Herrera, A.G.S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., Eggel, I.: Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: *CLEF 2012 Working Notes* (2012)
13. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research* (2009), <http://nar.oxfordjournals.org/content/early/2009/05/29/nar.gkp440.abstract>

---

<sup>6</sup> <http://www.imageclef.org/2012>

14. Pathak, S., Criminisi, A., Shotton, J., White, S.: D., Robertson, Sparks, B., Munasinghe, I., Siddiqui, K.: Validating automatic semantic annotation of anatomy in DICOM CT images. In: *Progress in Biomedical Optics and Imaging- Proceedings of SPIE* (2011)
15. Pérez, M., Berlanga, R., Sanz, I., Aramburu, M.J.: A semantic approach for the requirement-driven discovery of web resources in the Life Science. *Knowledge and Information Systems* 34(3), 671–690 (2013)
16. Philbin, J., Prior, F., Nagy, P.: Will the Next Generation of PACS Be Sitting on a Cloud? *Journal of Digital Imaging* 24(2), 179–183 (2011)
17. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., Jimeno, A.: Text processing through Web services: calling Whatizit. *Bioinformatics* 24(2), 296–298 (2008)
18. Rebholz-Schuhmann, D., Yepes, A.J.J., Van Mulligen, E.M., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E., Hahn, U.: CALBC Silver Standard Corpus. *Journal of Bioinformatics and Computational Biology* 8(1), 163–179 (2010)
19. Rehder, B., Littman, M.L., Dumais, S.T., Landauer, T.K.: Automatic 3-language cross-language information retrieval with latent semantic indexing. In: *Text Retrieval Conference (TREC)*, pp. 233–239 (1997)
20. Rolim, C.O., Koch, F.L., Westphall, C.B., Werner, J., Fracalossi, A., Salvador, G.S.: A Cloud Computing Solution for Patient’s Data Collection in Health Care Institutions. In: *Proceedings of the 2010 Second International Conference on eHealth, Telemedicine, and Social Medicine, ETELEMED 2010*, pp. 95–99. IEEE Computer Society, Washington, DC (2010)
21. Rosenthal, A., Mork, P., Li, M.H., Stanford, J., Koester, D., Reynolds, P.: Methodological Review: Cloud computing: A new business paradigm for biomedical information sharing. *Journal of Biomedical Informatics* 43(2), 342–353 (2010), <http://dx.doi.org/10.1016/j.jbi.2009.08.014>
22. Ruch, P.: Query translation by text categorization. In: *Proceedings of the 20th International Conference on Computational Linguistics*, p. 686. Association for Computational Linguistics (2004)
23. Seifert, S., Kelm, M., Moeller, M., Mukherjee, S., Cavallaro, A., Huber, M., Comaniciu, D.: Semantic annotation of medical images. In: *Proceedings of SPIE Medical Imaging*, pp. 762808–762808–8 (2010), <http://dx.doi.org/10.1117/12.844207>
24. Tsumoto, S., Hirano, S.: Data mining in hospital information system for hospital management. In: *ICME International Conference on Complex Medical Engineering (ICME)*, pp. 1–5 (2009)

# Subtopic Mining Based on Head-Modifier Relation and Co-occurrence of Intents Using Web Documents\*

Se-Jong Kim and Jong-Hyeok Lee

Department of Computer Science and Engineering  
Pohang University of Science and Technology (POSTECH)  
{sejong, jhlee}@postech.ac.kr

**Abstract.** This paper proposes a method that mines subtopics using the head-modifier relation and co-occurrence of users' intents from web documents in Japanese. We extracted subtopics using the simple patterns based on the head-modifier relation between the query and its adjacent words, and returned the ranked list of subtopics by the proposed score equation. We re-ranked subtopics according to the intent co-occurrence measure. Our method achieved good performance than the baseline methods and suggested queries from the major web search engine. The results of our method will be useful in various search scenarios, such as query suggestion and result diversification.

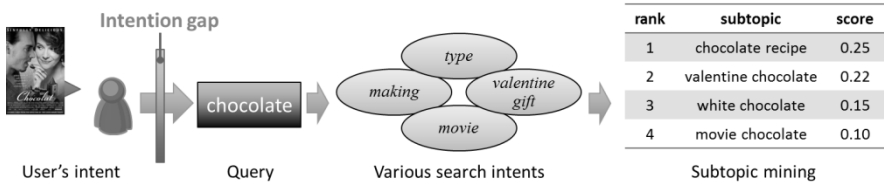
**Keywords:** search intent, subtopic mining, diversity, pattern, head-modifier.

## 1 Introduction

Many web queries are unclear and short because it is not easy for users to explicitly express their search intents through keywords. Some users do not choose appropriate words for a web search, and others omit specific terms needed to clarify search intents. This intention gap between users' search intents and queries results in queries which are ambiguous and broad. For ambiguous queries, users may get results quite different from their intents; for broad queries, results may not be as specific as users expect. As a solution for these problems, subtopic mining is proposed, which finds possible subtopics for a given query and returns a ranked list of them in terms of the relevance to the query, popularity and diversity of subtopics using resources such as query logs and web documents (Fig. 1). According to the NTCIR-9 subtopic mining task [1], a subtopic of a given query is a query that specifies and disambiguates the search intent of the original query. For evaluation, the collected subtopics with similar search intents are manually merged into *intents* as cluster names. For example, if a query is "chocolate," its specific hyponyms "white chocolate" and "dark chocolate" can be subtopics. These subtopics will be merged into one *intent* "chocolate type."

---

\* This work was supported by the Korea Ministry of Knowledge Economy (MKE) under Grant No.10041807, in part by the National Korea Science and Engineering Foundation (KOSEF) (NRF-2010-0012662).



**Fig. 1.** Flow from a user's intent to subtopic mining

Subtopic mining can be used to improve the results of various search scenarios, such as query suggestion (autocomplete and related queries) and result diversification.

The NTCIR-9 subtopic mining task motivated various methods based on query suggestion [2-8], term selection [9-12], and term disambiguation [13], [14] for the Chinese and Japanese languages. To achieve high-level performance, [15-18] used suggested queries from major web search engines (Baidu, Bing, Google, and Yahoo), and [12], [19], [20] used top-ranked documents obtained from search engines. In addition to the resources provided, in [1], [15], [17], [19], query logs, web documents, or online encyclopedias were used, and [1], [20] utilized anchor texts and URLs. However, the methods with high performance depended on external resources, as mentioned earlier, rather than the given resources. Moreover, most of the query logs were proprietary resources, and the methods which depended on query logs have data sparseness to find subtopics for rare queries because they are few or non-existent in query logs. Furthermore, these methods were non-linguistic, and focused on relevance and popularity.

This paper proposes a method that mines subtopics using the head-modifier relation between the query and its adjacent words, ranks them by the proposed score equation which measures relevance, popularity and diversity, and re-ranks the subtopics according to the co-occurrence of users' intents for the Japanese language. Our contributions are as follows:

- Our method does not use external resources and query logs. We only use the document collection provided (ClueWeb09-JA).
- Our method is a linguistic approach based on the head-modifier relation. We define the structure of a subtopic as “query's *modifier* (*sense-hyponym-modifier*) + query (*modifier* + *head*) + query's *head* (*information-head*),” and find or generate various subtopics using the simple patterns reflecting this structure.
- Our method demonstrates the usefulness of queries' *modifiers* and relevant documents in subtopic mining.
- Our method pays attention to the co-occurrence of users' intents in web documents for improving diversity.

In Section 2, we review previous work and present evaluation methods. A description of our proposed method is given in Section 3. Our results are presented in Section 4. Section 5 discusses the results, and we conclude this paper in Section 6.

## 2 Previous Work

### 2.1 Query Suggestion, Term Selection and Disambiguation

Subtopic mining methods are related to query suggestion, term selection, and term disambiguation, because the concept of a subtopic is similar to the results obtained by these approaches. Suggested queries help users choose appropriate words, while keywords of documents and senses of ambiguous queries can clarify the intents of users. Typically, the query suggestion approach uses query logs. In [3], query pairs that co-occurred frequently in same search sessions were used. Similarly, [2], [4] used click-through data to find similar queries, which share a large number of clicked URLs. [5] segmented queries into phrases and generated suggested queries using phrases in an English query log. For example, the query “Britney Spears mp3s” is segmented into the phrases “Britney Spears” and “mp3s,” which are used to generate suggested queries like “Britney Spears lyrics” and “music mp3s.” [6], [7] divided a query into the topic word (the first part of the query) and the facet word (the second part of the query, separated by a space from the first part), and found facet attributes in a Japanese query log. Meanwhile, [21] utilized only the co-occurrence of terms in a document corpus, and suggested queries with higher quality. This method found phrases containing the last query word from all  $n$ -grams ( $n \leq 3$ ) in the corpus, and calculated the phrase selection probability and phrase-query correlation.

The approach of term selection uses web documents. [9], [10] found keywords from top-ranked web documents retrieved by a query, and [11] selected keywords that maximized the divergence between a language model defined by top-ranked web documents and that defined by entire web documents. [12] changed a clustering problem into a term selection problem. This method extracted all  $n$ -grams from titles and snippets of top-ranked web documents, and used the learned linear regression of the phrase length and several properties to rank the  $n$ -grams (candidate cluster names). The simple method of term disambiguation is to utilize online encyclopedias. [13], [14] extracted term lists from Wikipedia disambiguation pages, and constructed a test collection for ambiguous queries.

### 2.2 Japanese Subtopic Mining

The NTCIR-9 subtopic mining task provided the web document collections for Chinese (SogouT) and Japanese (ClueWeb09-JA), but provided only the log for Chinese queries (SogouQ). However, the Japanese subtopic mining task achieved its best performance when only external web documents were used [1], and performed well using suggested queries from major web search engines [16].

ORG-S-J-1 [1] used anchor texts and URLs extracted from external web documents, and did not depend on any other resources. To gather web documents, this method used Microsoft’s internal web search platform, and achieved first place in the task. This method assumed that “if there are various domain names for a subtopic, then the popularity of the subtopic increases.” The process of this method was:

1. Retrieve all anchor texts containing the query.
2. Merge duplicate texts by performing word segmentation on the anchor texts.
3. Rank anchor texts (subtopics)  $sts$  by:

$$Imp(st) = \log(1 + |URL(st)|) \cdot \sum_{dm \in DM(st)} \left(1 + \log(freq(st, dm))\right) \quad (1)$$

where  $URL(st)$  is a set of web documents pointed to by  $st$ ;  $DM(st)$  is a set of domain names of web documents pointed by  $st$ ;  $freq(st, dm)$  is the number of anchor texts that include  $st$  and point to web documents with the domain name  $dm$ .

ORG-S-J-2 [1] placed second in the task by applying the process described in [12], except for the phrase independence property. This method used titles and snippets of top-ranked external web documents retrieved by Microsoft's internal web search platform. The process of this method was as follows:

1. Extract all  $n$ -grams ( $n \leq 3$ ) from titles and snippets of top-ranked web documents.
2. Calculate the phrase frequency and inverted document frequency ( $FreqIDF$ ), cluster entropy ( $CE$ ), and several properties for the  $n$ -gram (subtopic)  $w$  as:

$$FreqIDF(w) = freq(w) \cdot \log \frac{N}{|D(w)|} \quad (2)$$

$$CE(w) = - \sum_{w' \in W, w' \neq w} \frac{|D(w) \cap D(w')|}{|D(w)|} \cdot \log \frac{|D(w) \cap D(w')|}{|D(w)|} \quad (3)$$

where  $freq(w)$  is the frequency of  $w$ ;  $N$  is the total number of top-ranked web documents for the query;  $D(w)$  is a set of IDs assigned to the web documents extracting  $w$ ;  $W$  is the set of all extracted  $n$ -grams (subtopics) for the query.

3. Rank  $ws$  using the linear regression of the phrase length and calculated properties in the step 2.

The evaluation methods of the NTCIR-9 subtopic mining task were I-rec, D-nDCG, and D#-nDCG [22]. I-rec (*intent* coverage) measures diversity, D-nDCG measures overall relevance and popularity across search intents, and D#-nDCG is an average of I-rec and D-nDCG. The assessors manually clustered the collected subtopics with similar search intents, and labeled cluster names. These cluster names were called *intents*. Each subtopic could belong to only one of the *intents*. Non-relevant or non-understandable subtopics were given relevance level 0. The probabilities of *intents* were estimated by a popularity voting process involving ten assessors. However, the decisions of a few assessors cannot accurately reflect the popularity of *intents*. This evaluation issue is left as future work.

### 3 Method

Our method consisted of three parts. The first part was to find or generate subtopics using the head-modifier relation between the query and its adjacent words. We created simple patterns based on the head-modifier relation in Japanese, extracted

subtopics using the patterns, and measured subtopics by the proposed score equation. The second part was to rank the subtopics by applying several weights to the score equation considering relevant documents, domains and URLs. The third part was to re-rank the subtopics using the co-occurrence of users' intents estimated from web documents. We identified subtopics with high values for the intent co-occurrence measure, and re-scored these subtopics.

### 3.1 Extracting Subtopics Using the Head-Modifier Relation

We can specify words using other words. In the head-modifier relation, specified words and specifying words are called *heads* and *modifiers* respectively. In Japanese (also Chinese, Korean), *heads* of noun phrases appeared after *modifiers*. We assumed that “a subtopic of a given query is a noun phrase consisting of the original query and its adjacent words with head-modifier relations” because a subtopic is a phrase with the specific meaning (sense, hyponym, or information) of a query. Based on this assumption, for the query “office,” we can find the specific senses “software office” and “workplace office” using queries' *modifiers* “software” and “workplace.” For the query “chocolate,” we can also find the specific hyponyms “valentine chocolate” and “white chocolate” using queries' *modifiers* “valentine” and “white.” Moreover, we can find noun phrases as the specific information of each query such as “office update” and “chocolate recipe” using queries' *heads* “update” and “recipe” (Fig. 2).

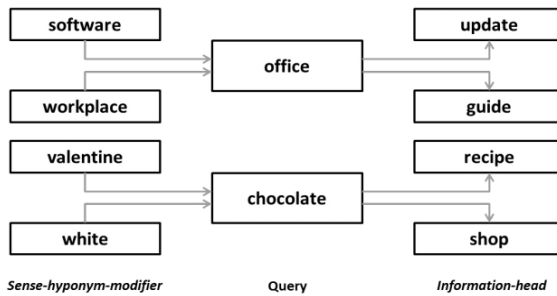


Fig. 2. Head-modifier relations of subtopics for the queries “office” and “chocolate”

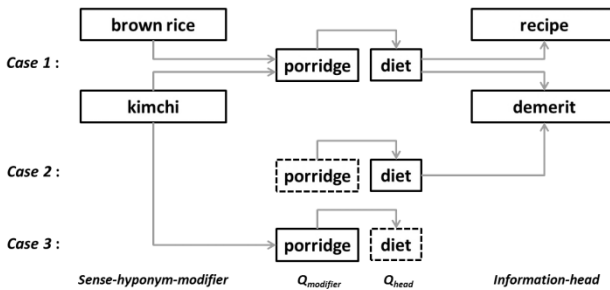


Fig. 3. Head-modifier relations of subtopics for the query “porridge diet”

**Table 1.** Patterns and examples of extracted subtopics for *Cases 1-3*

Case	Pattern	Extracted noun phrase			Subtopic
		Front nouns	Query, $Q_{head}$ or $Q_{modifier}$	Back nouns	
1	(noun)+(の)?(query) (の)?(noun)+	玄米 (brown rice) modifier	お粥ダイエツト (porridge diet) modifier, query	レシピ (recipe) head	玄米お粥ダイエツトレ シピ (recipe of brown rice porridge diet)
2	(noun)+(の)?( $Q_{head}$ ) (の)?(noun)+	any nouns modifier	ダイエツト (diet) modifier, $Q_{head}$	デメリツト (demerit) head	お粥ダイエツトデメリ ツト (porridge diet demerit)
3	(noun)+(の)?( $Q_{modifier}$ ) (の)?(noun)+	キムチ (kimchi) modifier	お粥 (porridge) modifier, $Q_{modifier}$	any nouns head	キムチお粥ダイエツト (kimchi porridge diet)

However, if a query consists of more than two keywords, because the number of noun phrases that fully match the query decreases, we cannot thoroughly extract various subtopics from web documents. To overcome this limitation, we divided the query into a *modifier* part  $Q_{modifier}$  and a *head* part  $Q_{head}$ , and found noun phrases that matched  $Q_{modifier}$  or  $Q_{head}$ . If  $Q_{modifier}$  or  $Q_{head}$  involve significant meanings for the query, then the partially matched noun phrases contain useful words that can be used to generate subtopics. As shown in Fig. 3, there were three cases for noun phrases for subtopic mining. *Case 1* showed a fully matched noun phrase that contained a *modifier* and a *head* for the query. *Case 2* was a partially matched noun phrase that contained a *modifier* (any nouns, including  $Q_{modifier}$ ) and a *head* for  $Q_{head}$ . *Case 3* was a partially matched noun phrase that contained a *modifier* and a *head* (any nouns, including  $Q_{head}$ ) for  $Q_{modifier}$ . In *Case 1*, we simply extracted these noun phrases as subtopics. If the query consisted of one keyword, we also extracted “query + head” and “modifier + query”. In *Case 2* and *Case 3*, we extracted these noun phrases except those in *Case 1*, then replaced “modifier (any nouns) +  $Q_{head}$ ” and “ $Q_{modifier}$  + head (any nouns)” in the noun phrases with the query. Noun phrases generated by this replacement were considered as subtopics.

From these three cases, we could define the structure of a subtopic as “query’s *modifier* (*sense-hyponym-modifier*) + query (*modifier* + *head*) + query’s *head* (*information-head*).” Since it is not easy to parse a mass corpus of web documents using a dependency parser, to extract subtopics for each case in our method, we created simple patterns that reflected the Japanese characteristic and this structure using a POS tagger, which is shown in the second column of Table 1. In the patterns, the + operator indicates there are one or more preceding elements, the ? operator indicates there is zero or one preceding element, and the meaning of “ $A$ の $B$ ” is “ $B$  of  $A$ .”

Our method found and generated various subtopics using the patterns (Table 1), and measured the scores of subtopics. For the convenience of implementation, we used the last noun and the remained noun phrase of the query as  $Q_{head}$  and  $Q_{modifier}$  respectively. The process of our method was as follows:



1. Retrieve all web documents that have titles or anchor texts with the query; titles and anchor texts are important features in document retrieval.
2. From the retrieved web documents, extract noun phrases satisfying the patterns, and find or generate subtopics according to the cases mentioned before.
3. Evaluate the score (*Score*) of each subtopic  $st$  by combining two equations applying *FreqIDF* or *CE* in Section 2 as:

$$AvgFreqIDF(st) = \frac{\sum_{st \in ST} FreqIDF(st)}{|ST| \cdot avg(AvgFreqIDF)} \quad (4)$$

$$CE_{sh}(md_{st}) = \frac{CE(md_{st})}{avg(CE_{sh})} \quad (5)$$

$$Score(st) = (1 - \lambda)AvgFreqIDF(st) + \lambda CE_{sh}(md_{st}) \quad (6)$$

where  $AvgFreqIDF(st)$  is used to measure the relevance and popularity of  $st$ ;  $ST$  is the set of extracted noun phrases for  $st$ ;  $avg(AvgFreqIDF)$  is the average of all  $AvgFreqIDFs$  for normalization;  $CE_{sh}(md_{st})$  is used to measure the cluster entropy of the query's modifier  $md_{st}$  (*sense-hyponym-modifier*) in  $st$  (if  $st$  does not have  $md_{st}$ ,  $md_{st}$  is the unique tag "NON");  $avg(CE_{sh})$  is the average of all  $CE_{sh}$ s; and  $1 - \lambda$  and  $\lambda$  are weights.

In  $CE_{sh}$ , we used  $md_{st}$  and the set of all the query's modifiers (including "NON") instead of  $w$  and  $W$  of  $CE$  to improve diversity because a web document generally relates to one sense of some ambiguous query or one hyponym of some broad query, and  $md_{st}$  clarified senses and generated hyponyms of the query (Fig. 2).

### 3.2 Ranking Subtopics Applying Weights

We implemented various ranking methods applying four types of  $\lambda$  for *Score* (Equation 6), and two types of weights for the appearance of subtopics and web document IDs. We assumed that "if the number of domains related to the query is large, the number of senses and hyponyms of the query is large," and "if the number of URLs related to the query is large, the number of subtopics for the query is large." Under these assumptions, we defined the weight for the equation as:

$$AnchorWeight = \frac{|Distinct\ domains\ of\ anchor\ texts\ with\ the\ query|}{|Distinct\ URLs\ of\ anchor\ texts\ with\ the\ query|} \quad (7)$$

The first type of equation weight set  $\lambda = AnchorWeight$  in *Score*. The name of this proposed method was PROP-A. The purpose of this method was to decide automatically the equation weight considering diversity of subtopics for each query. The second, third, and fourth types of equation weights set  $\lambda$  in *Score* equal to the constant values of 0.3, 0.5, and 0.7 respectively. The names of these proposed methods were PROP-1, PROP-2, and PROP-3. If the value of  $\lambda$  was larger, the corresponding method focused more on diversity of subtopics.

For the appearance weight, the first type was to set the frequency unit of subtopics and web document IDs to 1.0, and apply the unit to  $AvgFreqIDF$  and  $CE_{sh}$ . To distinguish methods, we attached "U" after the name of the proposed methods, such as

PROP-AU. The second type of appearance weight was to set the frequency unit of subtopics extracted from web documents that had titles containing the query and the IDs of this web documents to 1.0, set the frequency unit of the others to 0.9 by subtracting the relevance-penalty 0.1, and apply the units to  $AvgFreqIDF$  and  $CE_{sh}$ . The attach tag was “W,” and the aim of this method was to consider document relevance, because if the title of a web document contains the query, then it is more relevant to the query.

### 3.3 Re-ranking Subtopics Using the Co-occurrence of Intents

To maintain consistency, authors of web documents describe one or more contents of a topic in each document. Based on this characteristic, we assumed that “a web document contains various intents of the author (user) for a specific topic (query).” For example, Fig. 4 depicts web documents related to “Mozart.” In each web document, the author describes three contents for “Mozart,” which reflect the intents of the author, such as “Mozart music,” “Mozart symphony” and “Mozart concerto.” Because authors are part of users, various intents of authors are applicable to that of users.

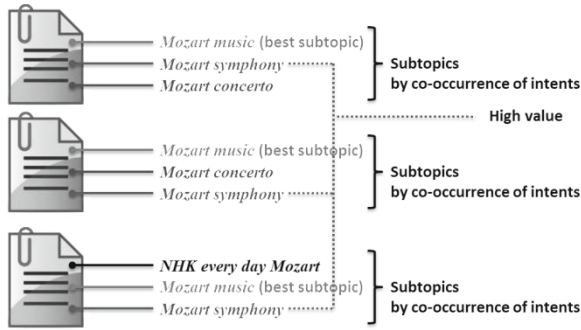


Fig. 4. Illustration of the concept for the co-occurrence of users' intents

To improve popularity and diversity of subtopics, our method focused on the co-occurrence of users' intents. We estimated the intent co-occurrence measure from various intents of authors of the web documents, and re-ranked the extracted subtopics using the measure. If the subtopic “Mozart music” was assigned a high value using *Score*, and this subtopic and “Mozart symphony” co-occurred frequently in the same documents, then the intent co-occurrence measure for “Mozart symphony” was high, and others such as “NHK every day Mozart” were low (Fig. 4). Our re-ranking process was as follows:

1. In documents containing the best subtopic  $st_{best}$ , find other subtopics that satisfy the query's *modifier* (including “NON”) in  $st_{best}$ .
2. Apply *Score* to sort these subtopics by the intent co-occurrence measure ( $IC_{intent}$ ):

$$IC_{intent}(st) = \frac{|D(st) \cap D(st_{best})|}{|D(st_{best})|} \times Score(st) \tag{8}$$

3. Re-rank the top subtopic  $st_{top}$  by:

$$Score_{re-rank}(st_{top}) = \frac{Score(st_{top}) + Score(st_{best})}{2} \quad (9)$$

4. Except for  $st_{best}$  and  $st_{top}$ , repeat steps 1 to 4.

Methods performing the re-ranking process were marked by attaching “R” after the names of the proposed methods, such as PROP-AWR.

## 4 Experiments

We mined subtopics for the 20 Japanese queries (topic IDs 0101-0120) of the NTCIR-9 subtopic mining task. The average number of *intents* for these queries was 15.1. We used only the Japanese document collection Clue-Web09-JA that consisted of 67,000,000 web documents. To perform word segmentation and identify nouns, we used the morphological analyzer MeCab tagger<sup>1</sup>.

We implemented various proposed methods (PROP-\*) applying our method. To equally compare our method against previous work, as baseline methods, we implemented BASE-QS, BASE-SM1 and BASE-SM2 using only the collection of documents provided. BASE-QS and BASE-SM1 implemented exactly the process of document-centric approach in query suggestion [21] and ORG-S-J-1 [1] respectively, while BASE-SM2 implemented the modified process of ORG-S-J-2 [1]. In BASE-SM2, we extracted the text consisting of 50 words that appeared around the query in the document, and considered this text as the snippet. For the linear regression, we used 0.25 for each weight. In addition, to compare with suggested queries from the major web search engine, we reported Bing related queries (EXT-QS) [16] which had the best performance among suggested queries in the NTCIR-9 subtopic mining task. We evaluated and compared the results using I-rec (diversity measure), D-nDCG (relevance and popularity measure), and D#-nDCG (representative measure) [22]. The number of top ranked subtopics we evaluated was  $l = 10, 20, \text{ and } 30$ .

In Tables 2-4, the “underlined bold” values represent the best performances among all methods. For  $l = 10$ , our best method was PROP-2WR. Compared to BASE-QS (the best baseline method), its mean I-rec@10, mean D-nDCG@10, and mean D#-nDCG@10 were improved by 0.0277, 0.0914, and 0.0595 respectively (Table 2).

**Table 2.** Results of baseline and proposed methods for  $l = 10$

Method name	Mean I-rec@10	Mean D-nDCG@10	Mean D#-nDCG@10
BASE-QS	0.2801	0.3263	0.3032
BASE-SM1	0.2345	0.3398	0.2872
BASE-SM2	0.1417	0.1758	0.1587
PROP-AU / AUR	0.2960 / 0.2901	0.4058 / 0.4109	0.3509 / 0.3505
PROP-AW / AWR	0.2927 / 0.2884	0.4055 / 0.4116	0.3491 / 0.3500
PROP-IU / IUR	0.2859 / 0.2819	0.4011 / 0.3987	0.3435 / 0.3403

<sup>1</sup> <http://mecab.sourceforge.net>

**Table 2.** (Continued)

PROP-1W / 1WR	0.2859 / 0.2847	0.4010 / 0.3993	0.3434 / 0.3420
PROP-2U / 2UR	0.3010 / 0.3017	0.4094 / 0.4124	0.3552 / 0.3571
PROP-2W / 2WR	0.3010 / 0.3078	0.4149 / 0.4177	0.3579 / <b>0.3627</b>
PROP-3U / 3UR	0.2968 / 0.2901	0.4197 / 0.4226	0.3583 / 0.3564
PROP-3W / 3WR	0.3012 / 0.2929	0.4199 / <b>0.4237</b>	0.3606 / 0.3583
EXT-QS	<b>0.3322</b>	0.3871	0.3597

For  $l = 20$ , our best method was PROP-3W, and compared to BASE-QS (the best baseline method), its mean I-rec@20, mean D-nDCG@20, and mean D#-nDCG@20 were improved by 0.0848, 0.1152, and 0.1000 respectively (Table 3).

**Table 3.** Results of baseline and proposed methods for  $l = 20$ 

Method name	Mean I-rec@20	Mean D-nDCG@20	Mean D#-nDCG@20
BASE-QS	0.3577	0.2992	0.3284
BASE-SM1	0.3239	0.3210	0.3224
BASE-SM2	0.2455	0.1799	0.2127
PROP-AU / AUR	0.4287 / 0.4249	0.4021 / 0.4056	0.4154 / 0.4153
PROP-AW / AWR	0.4321 / 0.4299	0.4001 / 0.4050	0.4161 / 0.4175
PROP-1U / 1UR	0.4344 / 0.4275	0.3884 / 0.3956	0.4114 / 0.4116
PROP-1W / 1WR	0.4344 / 0.4275	0.3885 / 0.3974	0.4115 / 0.4124
PROP-2U / 2UR	0.4310 / 0.4364	0.3992 / 0.4052	0.4151 / 0.4208
PROP-2W / 2WR	0.4384 / 0.4339	0.4000 / 0.4045	0.4192 / 0.4192
PROP-3U / 3UR	0.4401 / 0.4375	0.4107 / 0.4141	0.4254 / 0.4258
PROP-3W / 3WR	<b>0.4425</b> / 0.4399	0.4144 / <b>0.4167</b>	<b>0.4284</b> / 0.4283
EXT-QS	0.3322	0.2897	0.3110

For  $l = 30$ , our best method was PROP-AW. Compared to BASE-SM1 (the best baseline method), its mean I-rec@30, mean D-nDCG@30, and mean D#-nDCG@30 were improved by 0.1308, 0.0967, and 0.1138 respectively (Table 4).

**Table 4.** Results of baseline and proposed methods for  $l = 30$ 

Method name	Mean I-rec@30	Mean D-nDCG@30	Mean D#-nDCG@30
BASE-QS	0.4099	0.2797	0.3448
BASE-SM1	0.3857	0.3151	0.3504
BASE-SM2	0.2877	0.1934	0.2405
PROP-AU / AUR	0.5143 / <b>0.5184</b>	0.4108 / 0.4084	0.4625 / 0.4634
PROP-AW / AWR	0.5165 / 0.5107	0.4118 / 0.4115	<b>0.4642</b> / 0.4611
PROP-1U / 1UR	0.5079 / 0.5010	0.3926 / 0.3939	0.4502 / 0.4475
PROP-1W / 1WR	0.5079 / 0.5017	0.3917 / 0.3941	0.4498 / 0.4479
PROP-2U / 2UR	0.5150 / 0.5104	0.4043 / 0.4075	0.4596 / 0.4590
PROP-2W / 2WR	0.5150 / 0.5071	0.4062 / 0.4078	0.4606 / 0.4574
PROP-3U / 3UR	0.5021 / 0.5116	0.4145 / 0.4133	0.4583 / 0.4624
PROP-3W / 3WR	0.4966 / 0.5137	<b>0.4156</b> / 0.4136	0.4561 / 0.4637
EXT-QS	0.3322	0.2629	0.2976

## 5 Discussion

Our method achieved good performance for  $l = 10, 20,$  and  $30$  using the limited resource (only the document collection provided). The mean  $D\#-nDCG@10$  of PROP-2WR was  $0.3627$ , the mean  $D\#-nDCG@20$  of PROP-3W was  $0.4284$ , and the mean  $D\#-nDCG@30$  of PROP-AW was  $0.4642$ . Our methods with  $\lambda$  set to  $0.5$  or  $0.7$  in *Score* produced mostly better results than the other methods, and this means that  $CE_{sh}$  considering queries' *modifiers* was more useful than *AvgFreqIDF*. Actually, queries' *modifiers* specified senses and hyponyms of queries so that we could find or generate various subtopics using the *modifiers*. On the other hand, our methods with  $\lambda$  in *Score* set to *AnchorWeight* obtained good performance for  $l = 30$  only because the number of domains or URLs of anchor texts was not large enough to derive appropriate weights for the equations.

The names of our best methods were PROP-\*W\*, and this indicates that the weight corresponding to the relevance of a document for the query was more useful than other weights. Because subtopics in relevant documents could be more related to the query, document relevance was the important factor for subtopic mining. In addition, one of the reasons for the low performance of BASE-SM2 was that we did not extract subtopics from top-ranked (more relevant) web documents using a web search engine.

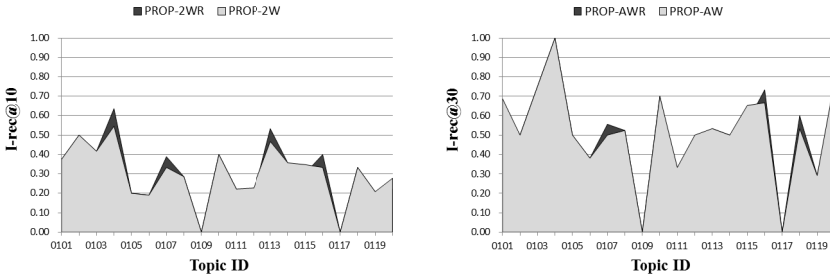


Fig. 5. I-rec for  $l = 10$  and  $30$  for each topic ID

In Fig. 5, we found that the intent co-occurrence measure was partially useful in improving I-rec (diversity) for each query (topic ID). Compared to our methods that did not re-rank, the methods that were re-ranked by applying the intent co-occurrence measure were improved for several parts.

However, we applied our method only to Japanese subtopic mining, and could not mine subtopics for two queries (topic IDs 0109: “Kim So Youn” and 0117: “the origin of the Zodiac”) due to data sparseness for relevant documents and partial matching. Therefore, to overcome these problems, we will construct new patterns, modify our methodology to be applicable to other languages such as English, combine our method with open resource based approaches, and apply it to various queries.

## 6 Conclusion

This paper proposed a method that mined subtopics using the head-modifier relation between the query and its adjacent words, and re-ranked them considering the co-occurrence of users' intents using only the provided collection of web documents for Japanese. The proposed method achieved a mean  $D\#-nDCG@10$  of 0.3627 (PROP-2WR), a mean  $D\#-nDCG@20$  of 0.4284 (PROP-3W), and a mean  $D\#-nDCG@30$  of 0.4642 (PROP-AW). Compared to the best baseline methods, these results were improved by 0.0595, 0.1000, and 0.1138 respectively. Our best methods also outperformed the previous method that used suggested queries from the major web search engine. These results will be useful in other subtopic mining tasks, or various search scenarios such as query suggestion and result diversification.

## References

1. Song, R., Zhang, M., Sakai, T., Kato, M.P., Liu, Y., Sugimoto, M., Wang, Q., Orii, N.: Overview of the NTCIR-9 INTENT Task. In: NTCIR-9 Workshop Meeting, pp. 82–105 (2011)
2. Beeferman, D., Berger, A.: Agglomerative clustering of a search engine query log. In: The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 407–416 (2000)
3. Huang, C.-K., Chien, L.-F., Oyang, Y.-J.: Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs. *Journal of the American Society for Information Science and Technology* 54, 638–649 (2003)
4. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query Recommendation Using Query Logs in Search Engines. In: Lindner, W., Fischer, F., Türker, C., Tzitzikas, Y., Vakali, A.I. (eds.) EDBT 2004. LNCS, vol. 3268, pp. 588–596. Springer, Heidelberg (2004)
5. Jones, R., Rey, B., Madani, O., Greiner, W.: Generating Query Substitutions. In: The 15th International Conference on World Wide Web, pp. 387–396 (2006)
6. Fujita, S., Machinaga, K., Dupret, G.: Click-graph Modeling for Facet Attribute Estimation of Web Search Queries. In: RIAO 2010 Adaptivity, Personalization and Fusion of Heterogeneous Information, pp. 190–197 (2010)
7. Fujita, S., Uchiyama, T., Dupret, G., Baeza-Yates, R.: Search Facet Creation from Click Logs. In: SIGIR 2010 Workshop on Query Representation and Understanding, pp. 25–28 (2010)
8. Dang, V., Croft, W.B.: Query Reformulation Using Anchor Text. In: The 3rd ACM International Conference on Web Search and Data Mining, pp. 41–50 (2010)
9. Xu, J., Croft, W.B.: Query Expansion Using Local and Global Document Analysis. In: The 19th Annual International ACM SIGIR Conference, pp. 4–11 (1996)
10. Lam-Adesina, A.M., Jones, G.J.F.: Applying Summarization Techniques for Term Selection in Relevance Feedback. In: The 24th Annual International ACM SIGIR Conference, pp. 1–9 (2001)
11. Carpineto, C., de Mori, R., Romano, G., Bigi, B.: An Information-Theoretic Approach to Automatic Query Expansion. *ACM Transactions on Information Systems* 19, 1–27 (2001)
12. Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., Ma, J.: Learning to Cluster Web Search Results. In: The 27th Annual International ACM SIGIR Conference, pp. 210–217 (2004)

13. Sanderson, M.: Ambiguous Queries: Test Collections Need More Sense. In: The 31st Annual International ACM SIGIR Conference, pp. 499–506 (2008)
14. Song, R., Qi, D., Liu, H., Sakai, T., Nie, J.-Y., Hon, H.-W., Yu, Y.: Constructing a Test Collection with Multi-Intent Queries. In: The 3rd International Workshop on Evaluating Information Access, pp. 51–59 (2010)
15. Zhang, S., Lu, K., Wang, B.: ICTIR Subtopic Mining System at NTCIR-9 INTENT Task. In: NTCIR-9 Workshop Meeting, pp. 106–110 (2011)
16. Santos, R.L.T., Macdonald, C., Ounis, I.: Exploiting Query Reformulations for Web Search Result Diversification. In: The 19th International Conference on World Wide Web, pp. 881–890 (2010)
17. Xue, Y., Chen, F., Zhu, T., Wang, C., Li, Z., Liu, Y., Zhang, M., Jin, Y., Ma, S.: THUIR at NTCIR-9 INTENT Task. In: NTCIR-9 Workshop Meeting, pp. 123–128 (2011)
18. Santos, R.L.T., Macdonald, C., Ounis, I.: University of Glasgow at the NTCIR-9 Intent task. In: NTCIR-9 Workshop Meeting, pp. 111–115 (2011)
19. Jiang, X., Han, X., Sun, L.: ISCAS at Subtopic Mining Task in NTCIR9. In: NTCIR-9 Workshop Meeting, pp. 168–171 (2011)
20. Han, J., Wang, Q., Orii, N., Dou, Z., Sakai, T., Song, R.: Microsoft Research Asia at the NTCIR-9 Intent Task. In: NTCIR-9 Workshop Meeting, pp. 116–122 (2011)
21. Bhatia, S., Majumdar, D., Mitra, P.: Query suggestions in the absence of query logs. In: The 34th Annual International ACM SIGIR Conference, pp. 795–804 (2011)
22. Sakai, T.: NTCIREVAL: A Generic Toolkit for Information Access Evaluation. In: The Forum on Information Technology 2011, vol. 2, pp. 23–30 (2011)

# Cultural Heritage in CLEF (CHiC) 2013

Vivien Petras<sup>1</sup>, Toine Bogers<sup>2</sup>, Elaine Toms<sup>3</sup>, Mark Hall<sup>3</sup>, Jacques Savoy<sup>4</sup>,  
Piotr Malak<sup>4</sup>, Adam Pawłowski<sup>5</sup>, Nicola Ferro<sup>6</sup>, and Ivano Masiero<sup>6</sup>

<sup>1</sup> Berlin School of Library and Information Science, Humboldt-Universität zu Berlin,  
Dorotheenstr. 26, 10117 Berlin, Germany  
vivien.petras@ibi.hu-berlin.de

<sup>2</sup> Royal School of Library and Information Science, Copenhagen University, Birketinget 6,  
2300 Copenhagen S, Denmark  
mvs872@iva.ku.dk

<sup>3</sup> The Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield,  
S1 4DP, UK  
{e.toms,m.hall}@sheffield.ac.uk

<sup>4</sup> Department of Computer Science, University of Neuchatel, rue Emile Argand 11,  
2000 Neuchatel, Switzerland  
{jacques.savoy,piotr.malak}@unine.ch

<sup>5</sup> Institute of Library and Information Science, University of Wrocław,  
pl. Uniwersytecki 9/13, 50-137 Wrocław, Poland  
apawlow@uni.wroc.pl

<sup>6</sup> Department of Information Engineering, University of Padova, Via Gradenigo 6/B,  
35131Padova, Italy  
{ferro,masieroi}@dei.unipd.it

**Abstract.** The Cultural Heritage in CLEF 2013 lab comprised three tasks: multilingual ad-hoc retrieval and semantic enrichment in 13 languages (Dutch, English, German, Greek, Finnish, French, Hungarian, Italian, Norwegian, Polish, Slovenian, Spanish, and Swedish), Polish ad-hoc retrieval and the interactive task, which studied user behavior via log analysis and questionnaires. For the multilingual and Polish sub-tasks, more than 170,000 documents were assessed for relevance on a tertiary scale. The multilingual task had 7 participants submitting 30 multilingual and 41 monolingual runs. The Polish task comprised 3 participating groups submitting manual and automatic runs. The interactive task had 4 participating research groups and 208 user participants in the study. For the multilingual task, results show that more participants are necessary in order to provide comparative analyses. The interactive task created a rich data set comprising of questionnaire of log data. Further analysis of the data is planned in the future.

**Keywords:** cultural heritage, Europeana, ad-hoc retrieval, semantic enrichment, multilingual retrieval, Polish, interactive, user behavior.

## 1 Introduction

Cultural heritage collections – preserved by archives, libraries, museums and other institutions – consist of “sites and monuments relating to natural history, ethnography,



archaeology, historic monuments, as well as collections of fine and applied arts" [8]. Cultural heritage content is often multilingual and multimedia (e.g. text, photographs, images, audio recordings, and videos), usually described with metadata in multiple formats and of different levels of complexity. Cultural heritage institutions have different approaches to managing information and serve diverse user communities, often with specialized needs. The targeted audience of the CHiC lab and its tasks are developers of cultural heritage information systems, information retrieval researchers specializing in domain-specific (cultural heritage) and / or structured information retrieval on sparse text (metadata) and semantic web researchers specializing in semantic enrichment with LOD data. Evaluation approaches (particularly system-oriented evaluation) in this domain have been fragmentary and often non-standardized. CHiC aims at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems.

After a pilot lab in 2012, where a standard ad-hoc information retrieval scenario was tested together with two use-case-based scenarios (diversity task and semantic enrichment task), the 2013 lab diversifies and becomes more realistic in its task organization. The pilot lab has shown that cultural heritage is a truly multilingual area, where information systems contain objects in many different languages. Cultural heritage information systems also differ from other information systems in that ad-hoc searching might not be the prevalent form of access to this type of content. The 2013 CHiC lab therefore focuses on multilinguality in the retrieval tasks and adds an interactive task, where different usage scenarios for cultural heritage information systems were tested. The multilingual task required multilingual retrieval in up to 13 languages, making CHiC the most multilingual CLEF lab ever. The Polish task concentrated on a rarely tested language in detail. Combining ad-hoc information retrieval and interactive information retrieval test scenarios in one lab provided an environment where both methodologies could overlap and benefit from each other.

CHiC has teamed up with Europeana<sup>1</sup>, Europe's largest digital library, museum and archive for cultural heritage objects to provide a realistic environment for experiments. Europeana provided the document collection (digital representations of cultural heritage objects) and queries from their query logs. The interactive task also provided a topic clustering algorithm and a customized browsable portal based on Europeana data.

The paper is structured as follows: Chapter 2 introduces the Europeana document collection, which is used in all 3 tasks. Chapters 3-5 describe the tasks in detail, their requirements, participants and results. The conclusion provides an outlook on the future of CHiC and the potential synergies of combining ad-hoc and interactive information retrieval evaluation.

## 2 The Europeana Collection

The Europeana information retrieval document collection was prepared for the CHiC pilot lab in 2012 (Petras et al., 2012). It consists of the complete Europeana metadata

---

<sup>1</sup> <http://www.europeana.eu>

index as downloaded from the production system in March 2012. It contains 23,300,932 documents. With the move of Europeana to an open data license in the summer of 2012 and the subsequent changes in content, this test document collection represents a snapshot of Europeana data from a particular time. However, the overlap to the current content is about 80%.

The collection consists of metadata records describing cultural heritage objects, e.g. the scanned version of a manuscript, an image of a painting or sculpture or an audio or video recording. Roughly, 62% of the metadata records describe images, 35% describe text, 2% describe audio and 1% video recordings.

The collection was divided into 14 sub-collections according to the language of the content provider of the record (which usually indicates the language of the metadata record). A threshold was set: all languages with less than 100,000 documents were grouped together under the name “Others”. The 13 language collections included Dutch, English, German, Greek, Finnish, French, Hungarian, Italian; Norwegian, Polish, Slovenian, Spanish, Swedish. For the CHiC 2013 experiments, all sub-collections except the “Others” were used, totaling roughly 20 million documents.

The XML metadata contains title and description data, media type and chronological data as well as provider information. For ca. 30% of the records, content-related enrichment keywords were added automatically by Europeana based on a mapping between metadata terms and terms from controlled lists like DBpedia names. In the Europeana portal, object records commonly also contain thumbnails of the object if it is an image and links to related records. These were not included with the test collection, but relevance assessors were able to look at them at the original source.

### 3 The CHiC Multilingual Task

This task is a continuation of the 2012 CHiC lab, using similar task scenarios, but requiring multilingual retrieval and results. Two sub-tasks were defined: multilingual ad-hoc retrieval and multilingual semantic enrichment.

The traditional ad-hoc retrieval task measures information retrieval effectiveness with respect to user input in the form of queries. The 13 language sub-collections form the multilingual collection (ca. 20 million documents) against which experiments were run. Participants were asked to submit ad-hoc information retrieval runs based on 50 topics (provided in all 13 languages) and including at least 2 and at most all 13 collection languages. For pooling purposes, participants were also asked to submit monolingual runs choosing any of the collection languages. Because the topics were provided in all collection languages, the focus of the task was not on topic translation, but on multilingual retrieval across different collection languages.

The multilingual semantic enrichment task requires systems to present a ranked list of related concepts for query expansion. Related concepts can be extracted from Europeana data or other external resources (e.g. Wikipedia or other resources from the Linked Open Data cloud). Participants were asked to submit up to 10 query expansion terms or phrases per topic. This task included 25 topics in all 13 languages. Participants could choose to experiment on monolingual or multilingual semantic

enrichments. The suggested concepts were assessed with respect to their relatedness to the original query terms or query category.

### 3.1 Topic Creation

A set of 50 topics was created for the 2013 edition of CHiC, where topic selection was determined partially by the potential for retrieving a sufficient number of relevant documents in each of the collection languages. CHiC 2012 used topics from the Europeana query logs alone, which resulted in zero results for some of the 3 languages [13]. The problem of having zero relevant results is aggravated when collection languages are varied, especially in the cultural heritage area. Many topics are relevant for only a few languages or cultures. For 2013, more focus was put on testing all topics in all languages for retrieving relevant documents, which resulted in fewer zero relevant result topics. The topic creation process started with creating a pool of candidate topics, which derived from four different sources:

- 15 topics that showed promising retrieval performance were re-used from the 2012 topic set (only in 3 languages) to test their performance in 13 languages.
- Another 19 topics that were not specific to only a handful of languages were taken from an annotated snapshot of the Europeana query log (the same procedure was used for the 2012 topics).
- The Polish task also suggested topics, 17 of which were not considered to be relevant only in Polish and input in the candidate pool.
- Finally, two of the track organizers generated another 21 test queries covering a wide range of topics contained in Europeana's collections that would span all collection languages.

These 73 candidate topics were then translated into all 13 languages by volunteers. The translated candidate topics were run against the 13 language collections using Indri 5.2 with default settings<sup>2</sup>. We retained the 50 topics that returned the highest number of relevant documents for all thirteen languages. Another factor that affected the final selection of the 2013 topics was the abundance of named-entity queries (around 60%) in the 2012 topic set. While named-entity queries are a common type of query for Europeana [18], they are less challenging than non-entity queries that describe a more complex information need. For this we wished to down-sample the proportion of named-entity queries to around 20%.

The final topics set covers a wide range of topics and consisted of 12 topics from the 2012 topic set, 13 log-based topics, 13 topics from the Polish subtask, and 12 intellectually derived queries. In form and type, the different query types are indistinguishable and usually include 1-3 query terms (e.g. "silent film", "ship wrecks", and "last supper"). For later relevance assessment, descriptions of the underlying information needs were added, but were not admissible for information retrieval. The underlying information need for a query can be ambiguous if the intention of the query is

---

<sup>2</sup> Jelinek-Mercer smoothing with  $\lambda$  set to 0.4 and no stemming or stopword filtering.

not clear. In this case, the track organizers discussed the query and agreed on the most likely information need.

### 3.2 Pooling and Relevance Assessments

This year, we produced 13 pools, one for each target language using different depths depending on the language and the available number of documents. The pools were created using all the submitted runs. A 14th pool, for the multilingual task, is the union of the 13 pools described above. We used graded relevance, i.e. highly relevant, partially relevant, and not relevant. To compute the standard performance measures reported in Section 3.3, we used binary relevance and conflated highly relevant and partially relevant to just relevant. The DIRECT system [1] has been used to collect runs, perform relevance assessment, and compute performances.

For all languages except English, native language speakers performed the relevance assessments. Fifteen assessors took 2 weeks to assess the ca. 140,000 documents. The assessors received detailed instructions on how to use the assessor interface and guidelines, how the relevance assessments were to be approached. Constant communication via a common mailing list ensured that assessors across languages treated topics from the same perspective.

### 3.3 Participants and Results

#### Multilingual Ad-hoc

Seven different teams participated in the 2013 edition of the ad-hoc track. Out of the 71 runs submitted, 30 were multilingual runs using at least 2 collection languages; 10 runs used all available languages for topics and documents. All languages were also represented in the monolingual runs (41 total). English (10 runs), German (6), French (6) and Italian (8) were the popular languages for the monolingual runs, all other languages had only 1 or 2 runs. Table 1 shows the best runs by participating group ordered by MAP showing the collection languages that were used for retrieval. Note that only the best run is selected for each group, even if the group may have more than one top run.

**Table 1.** Best Experiments per Group (in MAP)

Participant	Experiment Identifier	Collection Languages	MAP
Chemnitz	TUC_ALL_LA	All	23.38%
CEA List	MULTILINGUALNOEXPANSION	All except EL, HU, SL	18.78%
Neuchatel	UNINEMULTIRUN5	All	15.45%
RSLIS	RSLIS_MULTI_FUSION_COMBSUM	All	8.37%
MRIM	MRIM_AR_2	EN	6.43%
Westminster	R005	EN,IT	6.30%
UC Berkeley	BERKMONODE03	DE	4.14%

It is difficult to interpret these figures as all runs regardless of the language sub-collections used were measured against the multilingual pool. Monolingual runs or runs using fewer languages could not have reached better numbers. The working notes paper includes a more detailed analysis for the different run types [14]. Table 2 below lists the participating groups and briefly summarizes their approaches to the ad-hoc track.

**Table 2.** Participating groups and their approaches to the multilingual ad-hoc track

Group	Description of approach
RSLIS, University of Copenhagen & Aalborg University (Denmark)	Language modeling with Jelinek-Mercer smoothing and no stopword filtering or stemming. One run each for English, French, and German where these topic languages are run against a multilingual index. Two fusion runs using the CombSUM and CombMNZ methods combining these three monolingual runs against the multilingual index [17].
University of Neuchâtel (Switzerland)	Probabilistic IR using Okapi model with stopword filtering and light stemming. Collection fusion on the results lists from 13 different monolingual indexes using z-score normalization merging [2].
MRIM/LIG, University of Grenoble (France)	Language modeling approach using Dirichlet smoothing that uses Wikipedia as an external document collection to estimate the word probabilities in case of sparsity of the original term-document matrix [20].
CEA LIST (France)	Query expansion of a Vector Space model with tf-idf weighting by using related concepts extracted from Wikipedia using Explicit Semantic Analysis [15].
Technical University of Chemnitz (Germany)	Apache Solr with special focus on comparing different types of stemmers (generic, rule-based, dictionary-based) [22].
School of Information, UC Berkeley (USA)	Probabilistic text retrieval model based on logistic regression together with pseudo-relevance feedback for all of the runs. Runs with English, French, and German topic sets and sub-collections, as well translations generated by Google Translate [9].
University of Westminster (Great Britain)	Divergence from randomness algorithm using Terrier on the English and Italian collections [21].

### Multilingual Semantic Enrichment

Only 2 groups participated in the semantic enrichment task, making a comparison more difficult. Participants could choose between monolingual and multilingual runs. Almost all experiments contained only English concepts.

MRIM/LIG (Univ. of Grenoble) used Wikipedia as a knowledge base and the query terms in order to identify related Wikipedia articles for enrichment candidates.

Both in-links and out-links to and from these related articles (particularly their titles) were then used to extract terms for enrichment.

CEA List used Explicit Semantic Analysis (documents are mapped to a semantic structure) also with Wikipedia as a knowledge base. Whereas MRIM/LIG used the title of Wikipedia articles and their in- and out-links for concept expansion, CEA List concentrated on the categories and the first 150 characters within a Wikipedia article. When Wikipedia category terms overlapped with query terms, these concepts were boosted for expansion. In ad-hoc retrieval, the topic and expanded concepts were matched against the collection and the results were then matched again to a consolidated version of the topics (favoring more frequent concept phrases) before outputting the result. For multilingual query expansion, the interlingual links to parallel language versions of a Wikipedia article were used in a fusion model. For most expansion experiments, only concepts were considered that appear in at least 3 Wikipedia language versions, allowing for multilingual expansions.

The semantic enrichments were evaluated using a tertiary relevance assessment (definitely relevant, maybe relevant, not relevant) and P@1, P@3 and P@10 measurements. Table 3 shows the results for the best 2 runs for each participants using either the strict relevance measurement (just definitely relevant) or the relaxed relevance measurement (definitely relevant and maybe relevant).

**Table 3.** Semantic enrichment results

Run name	P@1	P@3	P@10
	Strict relevance		
MRIM_SE13_EN_WM	0.0400	0.0533	0.0422
MRIM_SE13_EN_WM_1	0.0800	0.0667	0.0522
ceaListEnglishMonolingual	0.5200	0.5467	0.4680
ceaListEnglishRankMultilingual	0.4800	0.4533	0.3400
	Relaxed relevance		
MRIM_SE13_EN_WM	0.2800	0.1333	0.1448
MRIM_SE13_EN_WM_1	0.2800	0.1467	0.1598
ceaListEnglishMonolingual	0.6800	0.7067	0.6600
ceaListEnglishRankMultilingual	0.6800	0.7200	0.5600

## 4 The CHiC Polish Task

The main objective of the Polish task was to obtain a better understanding of information retrieval problems for complex languages such as Polish [19] when facing short text descriptions. We know that the complex morphology of the Polish language may have an impact on both retrieval effectiveness and its relevance. Can this aspect be ignored under the assumption that the morphological complexity will not or have only a small impact on the retrieval performance? If not, can we evaluate the extent of the retrieval effectiveness variations when having a poorer or a better understanding of the Polish morphology? With a related language like Czech, previous studies indicate

that the stemming phase might improve the overall retrieval effectiveness of around 44% over an approach ignoring this word normalization procedure [4]. Can we achieve similar findings with relatively short description of CH objects?

To answer these questions we have organized a Polish task as a standard ad-hoc retrieval task, measuring the information retrieval effectiveness with respect to user input in the form of queries. The resulting ranked list of retrieved items is produced without any prior knowledge about either the user needs or the context.

The Polish collection is a part of the CHiC 2013 multilingual collection and each descriptor contains on average 35 terms. For this task, we have offered both an automatic and manual submission mode. In both cases, the participants are free to use the logical tags they want for indexing the various CH objects. Regarding those titles or the CH objects descriptions, participants are free to manually or automatically enrich the corresponding queries and/or document surrogates (e.g., using specific thesauri, dedicated ontologies or the web in general). Moreover, automatic blind feedback or query expansion mechanisms are allowed to hopefully improve the proposed ranking.

#### 4.1 Topic Creation

Based on the Europeana query logs, we have generated a set of 50 topics consisting of a mixture of topical and named-entity queries. The 50 short topics in title-format only (e.g., “królowie polscy w 18 wieku” – “Polish kings in 18 century”) tend to reflect information needs as expressed by real Europeana users. To provide an overview of the topic meaning, we manually translated them into the English language. For each topic, an additional description was provided to give the relevance assessor an idea of what subjects were intended to be retrieved. This last field cannot be used during the search process. When inspecting the number of search keywords in the title section only, we can count 10 titles composed only by a single word, and 11 titles with two terms. On average, the topic contains 2.82 search keywords.

As this year Poland has celebrated the 150<sup>th</sup> anniversary of the January uprising, we have added topics related to Polish territories and history within the 18th and 19th centuries. There are also 8 topics on certain historical periods (e.g., “chłopi w 18 lub 19 wieku” – “peasants in 18 or 19 century”) as well as 8 on temporary issues concerning Poland. 12 topics contain also personal names (e.g., “obrazy Jana Matejki” – “Jan Matejko's paintings”), but we also have 6 topics with geographical names (e.g., “kościół w Toruniu” – “churches in Torun”) or five with historical names (e.g., “Powstanie Styczniowe” – “January Uprising”). Finally, we can find 5 topics about religion or beliefs (e.g., “Matka Boża w sztuce” – “Our Lady in art”), and 7 on social groups or functions (e.g., “ruch robotniczy” – “workers movement”).

#### 4.2 Pooling and Relevance Assessments

Relevance assessments were done manually first by collaboratively generating an assumed information need for the topic and then describing it. The pooled documents (with a pool depth = 100, resulting in 32,144 judged documents) were then assessed

for their relevance according to the topic and the information need. This assumption is built around the perspective of an average user. We assumed that the majority of users typing that particular query would like to obtain that particular piece of information. Two experts have done the relevance assessments.

For this task, we have selected a three graded relevance value, with “fully relevant,” “partially relevant,” and “irrelevant”. By default, we will opt for a strict interpretation assuming that only items judged “fully relevant” are judged relevant. The assessors have found 8,530 fully relevant CH objects. On the other hand, 4,758 CH objects have been judged as partially relevant to the corresponding query.

Fully relevant items can be found for every topic, with a minimum of 5 relevant CH objects for Topic #17 (“Czesław Miłosz”), and a maximum of 562 pertinent items for Topic#20 (“PRL” People’s Republic of Poland). On average, we can find 170.6 relevant objects per topic (median: 125; stdev: 139.6).

Under the lenient option, we will consider as pertinent items judged fully or partially relevant. Under this condition, all topics have at least 22 relevant CH objects. This minimum value of 22 can be found for Topic#43 (“II Wojna Światowa” – “2nd World War”) and the maximum of 562 pertinent items for Topic#3 (“medycyna w 19 wieku” – “medicine in 19 century”). On average, we can find 265.8 relevant objects per topic (median: 263; stdev: 132.2).

### 4.3 Participants and Results

From the 7 teams having expressed an interest in this task, we only obtained runs from 3 groups, namely 1 in the automatic mode, and 2 in the manual mode. We have also received request for information from 2 other teams in Poland but they were not able to send their runs in time. Table 4 shows the list of active participants.

**Table 4.** Polish Task 2013 Participating Groups and Country

Institute of Information Science and Book Studies, Nicolaus Copernicus University	Poland
Institute of Library and Information Science Institute, University of Wrocław	Poland
Computer Science Dept., University of Neuchâtel	Switzerland

When analyzing their results, we have considered mainly mean average precision (MAP), an evaluation measure corresponding to a user who wants to retrieve all pertinent CH objects. As a second measure, we have also reported P@10, a measure reflecting the result given by the Europeana search engine in its first result screen.

#### Automatic Runs

In this mode, our intent was to explore the best search strategy to automatically search within a morphologically rich language. As a general overview of the automatic runs, Table 5 depicts the main results together with their descriptions, ordered by MAP. The third row (PLWR0Base) corresponds to an automatic run submitted by the To-run’s team [10] and used as a baseline for comparison for their manually enrichment query modifications. The University of Neuchatel (UniNE) sent the other runs [2]. To test for significant improvements, we applied a paired *t*-test. In our analysis,



statistically significant differences were detected by a two-sided test ( $\alpha=5\%$ ) and are denoted by “†”. There is no statistically significant difference between the first three runs.

**Table 5.** Strict Evaluation of Official Runs of the Automatic Mode

Rank	Name	Parameter Setting	MAP	P@10
1	UniNEFusion	Data fusion (Okapi: no stem, light stem, trunc-5)	0.3433	0.614
2	UniNEDFR	DFR-I( $n_c$ )B2, light stemming, with stopword	0.3308	0.568
3	PLWR0Base	Okapi, no stemming, with stopword	0.3140	0.552
4	UniNEPRF	Data fusion, PRF (Rocchio, 5 docs, 10 terms)	0.2578 †	0.494
5	UniNEBaseline	<i>tf idf</i> (cosinus), no stemming, with stopword	0.2566 †	0.492
6	UniNE-	Data fusion, 5-gram, PRF	0.2203 †	0.472

From the runs depicted in Rank#2, #3, and #5, we can see the performance differences achieved mainly when using the classical *tf\*idf* IR model [11], the Okapi model [16] and 1 implementation of the DFR probabilistic paradigm [3]. The MAP of the DFR-I( $n_c$ )B2 without stemming is 0.3028. Comparing the Okapi with the classical *tf\*idf* model, we notice a relative improvement of +22.4% (from 0.2566 to 0.3140).

Additional runs presented by UniNE [2] indicate that indexing the CH objects with isolated words tends to perform better than either the  $n$ -gram or trunc- $n$  indexing approaches. For example, the DFR-I( $n_c$ )B2 based on the trunc-6 indexing scheme achieves a MAP of 0.3078 (or a MAP of 0.2641 for the 6-gram scheme). Using the same IR model with a light stemming (word-based), we can obtain a MAP of 0.3308 (see UniNEDFR in Table 5). Of course, in the CH domain where names can be an important source of evidence to discriminate between relevant and irrelevant objects, taking into account the short sequences of terms (e.g., “Jaroslaw city” instead of only “Jaroslaw” because this might also be a personal name) may hopefully improve these retrieval performances. The use of a stopword list also seems a good practice. Based on additional runs described in [2], the Okapi model with stemming and without a stopword list achieves a MAP of 0.3258. When applying a stopword list (composed of 304 terms), the MAP increases to 0.3433 (a relative improvement of +5.3%). Indexing the CH objects with the Europeana automatically enrichment tags (indicated by the prefix *europæana*;) does not have any impact of the retrieval effectiveness because only a few enrichment tags have been added in the Polish corpus.

An interesting question is to analyze the retrieval performance comparing the performance difference between different stemming strategies as well as the use of a lemmatizer. Based on the submitted runs, only a partial answer can be provided. The UniNE group has compared the use of a light stemmer (removing only the inflectional suffixes related to the gender, number and grammatical cases) with approaches ignoring this word normalization procedure. Based on the *tf\*idf*, Okapi and DFR-I( $n_c$ )B2 models, the mean relative improvement of applying a light stemmer is 5.3%.

The run “UniNEFusion” indicates the retrieval effectiveness when combining 2 word-based Okapi models (with and without a light stemming procedure) with an Okapi model based on trunc-5 indexing scheme (only the first 5 letters of each word are considered). This data fusion strategy does not seem to be really effective because

we have another run based only on the Okapi model that already obtains a MPA of 0.3433. The runs “UniNEPRF” and “UniNEGramPRF” were also based on a data fusion between runs using pseudo-relevant feedback. According to unofficial runs described in [2], this automatic query expansion does not result in better retrieval effectiveness. For example, adding 5 terms extracted from the first 5 top-ranked retrieved items (Rocchio’s approach [11]) with the DFR-I(n<sub>c</sub>)B2 changes the MAP from 0.3028 before the query expansion to 0.2189 (after a relative decrease of -27.7%).

### Manual Runs

Within the manual mode, the participants are free to use any source of knowledge, tools, or strategies to modify and enrich the topics. No further user-system interaction is assumed after the first set of results is retrieved (but automatic blind feedback or query expansion mechanisms are allowed, although not used by the participants).

In Table 6, we have regrouped the evaluation of the official runs submitted in the manual mode, ordered by MAP. The run prefixed by the string “PLWR” comes from the Wroclaw University group [12] while those with the prefix “PLTO” are from the Torun group [10]. In both cases, the searchers have added a text description to semantically enrich the topic title. These additional terms were added under an “<enrich>” tag in the topic formulation. As depicted in Table 6, there is no statistically significant difference between the runs submitted by the Torun group. However, the retrieval performance differences are statistically significant between the best run (PLTO1EduLS) and all runs provided by the Wroclaw’s group.

**Table 6.** Strict Evaluation of Official Runs of the Manual Mode

Rank	Name	Enrichment (Parameter Setting)	MAP	P@10
1	PLTO1EduLS	Educated, light stemmer	<b>0.2774</b>	0.454
2	PLTO1EduNO	Educated, no stemmer	0.2724	<b>0.460</b>
3	PLTO2HighLS	High, light stemmer	0.2709	0.528
4	PLTO2HighNO	High, no stemmer	0.2690	0.528
5	PLWR2Exp	Experts (Okapi, no stemming)	0.1795 †	0.378
6	PLWR1Edu	Educated (Okapi, no stemming)	0.1529 †	0.350
7	PLWR3Stu	Students (Okapi, no stemming)	0.1279 †	0.268
	PLWR0Base	Basic (Okapi, no stemming)	0.3140	0.552

The Torun group wants to compare the difference in retrieval performance that can be achieved when comparing “educated” users vs. “specialists”. In the first case, the educated users have considered spelling variations, added other spellings for the same location or name or enriched the title by considering alternative formulations. With the specialists, the enrichment was based mainly on encyclopedias and a deeper elaboration of the main topic by including narrower terms (e.g., a list of writer names for a topic about “stories”). The educated users have added, on average, 3.3 terms, letting the mean length of the queries increase from 2.8 terms to 6.1 search keywords. With the specialists, this manual enrichment increases the mean topic length from 2.8 to 9.8 search terms.

As depicted in Table 5, these different forms of manual query enrichments do not improve the MAP over a simple search strategy using the title of the topic (run PLWR0Base). A first overview shows that mainly broad terms were added by the different user types and therefore the search system was not able to improve the ranking of the pertinent items. A query-by-query analysis reveals that the manual enrichment (PLTO1EduLS) improves the average precision (AP) for 22 queries over 50 compared to the automatic run (PLWR0Base). The largest improvement was obtained with the Topic #29 (“Warszawa w 19 wieku w sztuce” – “Warsaw in 19 century in art”). In this case, the AP increases from 0.001 (automatic run) to 0.3463, mainly by adding the terms “*architektura*” (architecture) and “*dzielnica*” (district). The specialists have also obtained a better retrieval performance for 20 topics over 50. The largest improvement was achieved with the Topic #32 (“kobiety w powstaniach w wojsku” – “uprising or military and women”) for which the MAP increases from 0.004 to 0.2825.

Moreover, the retrieval effectiveness of the various runs presented in Table 5 seems to indicate that applying a light stemming approach produces mixed results (see the performance difference between runs “*nnnLS*” and “*nnnNO*”).

When analyzing Wrocław’s run, we can use the same search strategy (Okapi in this case) and baseline performance as with Torun. The manual query enrichment done by experts (run “PLWR2Exp”) produces the best overall performance within this group. The performance difference with run “PLWR1Edu” is however not statistically significant (based on a paired *t*-test, two-sided,  $\alpha=5\%$ ). With the students’ run (run “PLWR3Stu”), the performance difference is larger (0.1795 vs. 0.1279, a relative difference of -28.7%), close to a statistically significant one (*p-value* = 0.0706).

As unofficial runs, the Torun team suggests that we can apply a Boolean search model [10]. In this approach, all keywords appearing in the title of the topic must be present in the retrieved items. With this model, they can achieve an MAP of 0.3484, the highest retrieval performance for this task. Of course this search strategy will not provide the best answer for all queries. An interesting example is Topic#24 (“Fryderyk Szopen” – “Fryderyk Chopin”) that achieves an AP of 0.113 when using the Okapi search engine (PLWR0Base) but an AP of 0.996 (+881%) when using a Boolean search model. Clearly having both terms in the retrieved documents implies higher chance to be pertinent. However, such a Boolean strategy does not perform well in all cases. For example, with Topic #41 (“barok”) the ranking provided by the Okapi model was better (AP: 6162) than that proposed by the Boolean model (AP: 0.004) based though on a single search keyword.

## 5 The CHiC Interactive Task

The intent of the CHiC*i* task was to collect a large enough data set that represented user interactivity with the Europeana collection so as to a) model user search/browse behaviour initially, and b) build a collection of user-centred data that might be augmented and used in future for testing various types of hypotheses about the process, the context and the nature of the interactivity. With that broad objective, the research

task focused on one user task: one with an implicit goal that reflects the exploratory nature of the interaction with culture and heritage information objects, particularly when the user is not an expert in the topic. As such it was designed to encourage interactivity and immersion in a culture and heritage environment, and the research design enabled multiple questions: what do people do when exposed to such an environment? How does the search process change over the course of that immersion? How do people interact with the images and their associated metadata? What can we learn from a user “session”? For this task, one common experimental system, one set of content and one interface was deployed and used by all teams [6].

### 5.1 Research Protocol, i.e., the Lab Task

The ‘task’ thus was a multi-part protocol that extracted multiple types of data from participants and observed participants virtually in their interactivity with the system. The protocol followed the pattern outlined in Fig. 1. All teams used the same protocol, which could be accessed remotely over the internet.

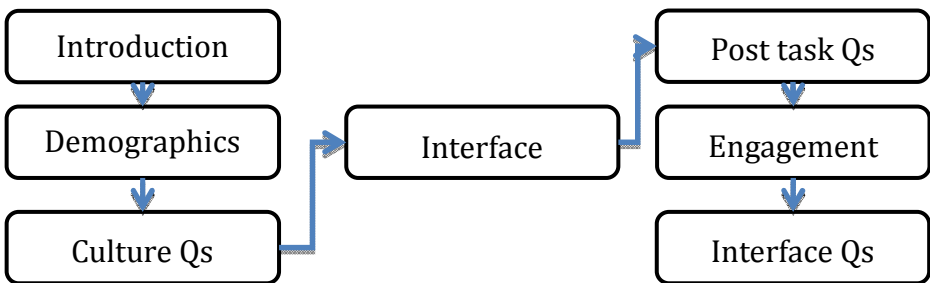


Fig. 1. CHiCi Research Protocol

An information sheet and informed consent (required by the University of Sheffield’s research ethics review process) was first presented to participants, followed by sets of questions about:

- basic demographic questions to create a profile of participant group;
- country of birth and residence, mother tongue, and language used to speak at home or search the web, to understand the potential impact of an individual’s culture;
- museum visits, familiarity and interest in European culture and heritage and experience with the European Digital Library, to address whether the participant was ‘of convenience’ or interested in the topic matter.

All of these may have influenced the level and intensity of their interaction with this resource. While participants were engaged in the assigned experimental task, the system logged and time stamped the entire set of user actions and events including:

queries, category selection, items examined, added to the bookbag, and so on. After the assigned task (see section 5.3), participants:

- responded to a 31-item User Engagement Scale to assess the overall experience;
- provided a narrative explanation of why they included the objects in the bookbag, and their level of satisfaction with what they found;
- assessed the usefulness of each object on the interface;
- assessed the usefulness of each piece of metadata in assisting with assessing an item.

## 5.2 IR System and Interface

The content contained 1,107,176 million records from the English-language collections of the Europeana Digital Library. The IR system was based on Apache Solr<sup>3</sup>, which provides the text search, spelling checker, and the “more like this” suggestions. The default settings were used for all components and all fields specified in the source records were loaded without any pre-processing.

Access to the IR system was provided using a novel Cultural and Heritage Explorer (see Fig. 2); it offered three key ways of accessing the content and additional features intended to support the assigned task.

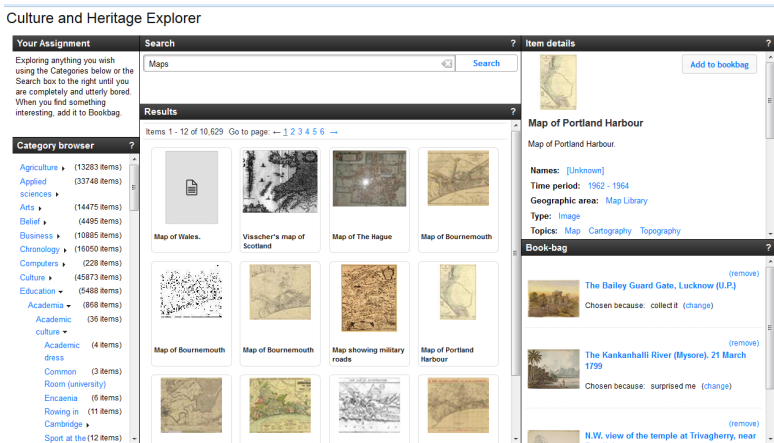


Fig. 2. CHiC Cultural and Heritage Explorer

In addition, a hierarchical category browser was added, based on the work of [5]. This process resulted in a set of 24 top-level categories, with between 3 and 14 sub-levels (median 5). The individual levels in the category hierarchy had between 1 and 384 sub-categories (median 3). A total of 267,768 items were automatically mapped into the category hierarchy. When the item – category mappings were loaded into

<sup>3</sup> <http://lucene.apache.org/solr/>

Solr, each item was linked both to the category the pre-processing had linked it to, and to all of that category's ancestors. When the user selected a category from the category browser, the Solr index was searched for all items that were mapped to this category. Because of the way the items were linked not only to their category, but also to the category's ancestors, this query would also return all items that were linked to the selected category's descendants.

In addition to the task assignment in the upper left corner, the interface contained:

- 1) Category hierarchy: The hierarchy was navigated using the right arrow located to the right of each category, which expanded the level within the space.
- 2) Search box: a conventional implementation of query entry that accepted keywords. After submitting a query, the results display below the box was updated.
- 3) Results display: displayed 16 thumbnails and titles of the thumbnails in a 3x4 grid layout that also enabled navigation within the list. When an item in this display was clicked, it appeared in the item display to the extreme upper right.
- 4) Item display: contained the thumbnail and metadata fields associated with the item; unfortunately, only the thumbnail is present in the data collection. The metadata use the Dublin core standard, but some Dublin core labels used expert jargon and were modified for a naïve participant. At this point, an item could be added to the bookbag using the button in the upper right corner. At the bottom of each item, the "more like this" was displayed using thumbnail images.
- 5) Bookbag: used for collected images that were deemed useful. Items in the bookbag could be redisplayed or removed. The display included the item and the rationale for including the item as well.

On startup, no query was inserted, but the results grid was populated with randomly selected images to serve as a stimulus for starting the task. At that point, a participant could enter a query, scan the categories, examine the results or an individual item, or select from "more like this." At the item display, a participant could search by any of the metadata contents, or add an item to the bookbag. Once the "add to Bookbag" was selected, a popup box asked why the object was selected with the following options:

- I wanted to show someone
- I wanted to use the image in something
- I wanted to collect for a future purpose
- It surprised me!
- I simply liked it! No particular reason.

### 5.3 Experimental Task

The implicit task (which remained stationary in the upper left corner of the Explorer) was: "Your Assignment: exploring anything you wish using the Categories below or the Search box to the right until you are completely and utterly bored. When you find something interesting, add it to the Bookbag." Prior to being assigned the task,

participants were presented with a situation to set the stage for the task: “Imagine you are waiting to meet a friend in a coffee shop or pub or the airport or your office. While waiting, you come across this website and explore it looking at anything that you find interesting, or engaging, or relevant...” No further guidance was given, and participants were free to explore the resource; a mouse click on a ‘Next Page’ button disengaged the participant from the activity.

#### 5.4 Research Teams

Four teams participated in this task, which required each team to process 30 participants via the web and 10 in a fixed observable lab-based location; not all participants met this objective as illustrated in Table 7. The language of operation was English, and all protocols and systems were expressed only in that language.

**Table 7.** Participating Research Teams

	<b>Web</b>	<b>Lab</b>	<b>Total</b>
Humboldt Universität	18	8	26
Royal School of Library and Information Science	12	19	31
Stockholm University	9	0	9
University of Sheffield	117	20	137
Other	4	1	5
<b>Total</b>	160	48	208

#### 5.5 Participants

The participant group (208) contained a well-educated group of about 1/3 male (f=136, m=72), about 2/3 were under 35, and about half had undergraduate degrees, and all were currently enrolled in a programme of study. Participants came from 16 countries but more than half are residents in the UK, but originated, i.e., by birth, in 35 countries. 20 languages are spoken today, but they speak 26 languages at home. However, the predominant language is English, both as a mother tongue and as the current language spoken.

On a scale of 1 to 5 (from not familiar to very familiar), participants rated familiarity with European culture and heritage at 2.2, and their interest in the topic in the middle of the scale at 2.5. Of the participants, 78% indicated that they have never visited Europeana and 81% visited museums and galleries on the web or in person less than monthly. Thus, participants were dominated by well-educated, English-speaking and origin, females under 35 who were relatively non-expert in European culture and heritage and neither particularly interested or uninterested in the topic, and who primarily had never visited Europeana.

## 5.6 Results

From both user responses and the log files, we aggregated selected measures by participant. See Table 8 for that summary. Because data was collected in two types of locations: via the Web and in the Lab, we present data by location as it became apparent in preliminary analyses that there may be differences. But, because of the variation in size of the two location groups we are hesitant to say that these differences are statistically significant, and thus report the result and identify what looks suggestive (identified with an asterisk \*).

**Table 8.** Summary Results across all participants

Measure	Definition	Web		Lab		Mean	
		#	<i>SD</i>	#	<i>SD</i>	#	<i>SD</i>
Queries	# of queries	3.5	8.6	5.3	6.6	3.9	8.2
Categories*	# of categories selected (hierarchy)	9.3	11.3	19.6	22.8	11.7	15.3
Metadata facets*	# of metadata facets examined	0.7	2.1	2.4	6.4	1.1	3.6
Query Time	Time (sec) spent querying	187.5	600.4	234.3	253.1	198.1	541.2
Category time*	Time (sec) spent using categories	239.2	299.8	493.0	362.1	296.8	331.7
Metadata time*	Time (secs) spent using metadata	22.8	78.1	65.7	179.4	32.5	110.5
Objects*	# of objects viewed	12.9	16.7	22.9	18.4	15.1	17.6
Objects (query)	# of objects viewed from query	5.4	11.0	7.7	9.78	5.92	10.8
Objects (categories)*	# of objects viewed from categories	5.7	9.1	13.2	12.8	7.4	10.5
Objects (metadata)	# of objects viewed from metadata	1.1	5.4	1.8	6.0	1.2	5.5
Interaction*	# of events/actions with system	57.1	63.4	97.1	67.6	66.2	66.4
Results page used*	# of results pages viewed	24.7	36.2	42.4	41.8	28.7	38.2
Bookbag	# of objects	6.0	8.3	4.5	4.2	5.7	7.6
Bookbag (category)	# of objects saved after category	2.9	4.7	2.5	3.1	2.8	4.4
Bookbag (metadata)	# of objects saved after metadata	0.3	1.3	0.3	0.7	0.3	1.2
Bookbag (query)	# of items in Bookbag after query	2.5	5.8	1.6	2.5	2.3	5.2



**Table 8.** (Continued)

Expected	Scale of 1-5, degree to which objects were as expected	1.54	0.977	1.94	1.099	1.63	1.017
Satisfied	Scale of 1-5, degree to which objects were as expected	1.74	1.119	1.92	1.145	1.78	1.125

As illustrated, participants issued on average approximately 4 queries, examined almost 12 categories, and about one of the metadata items associated with each object. They examined on average about 15 of the objects, with about 6 of those resulting from queries to the system and seven emerging from using the category explorer. Of these objects approximately 6 (50%) were deemed interesting enough to add to the Bookbag. On average they clicked on something on the interface 66 times, and clicked through the results pages 28 times. Overall, they were dissatisfied with what they found, and found the objects they examined not to be what they would have expected of Europeana.

In addition to understanding the effect of the interface, we also asked about the usefulness of each of the objects in the Explorer, but all were rated on the negative side on a five-point scale. Similarly, each object had a set of metadata associated with it, and of the set the Title, Description, and Thumbnail were considered to be useful in helping to assess the object with the title rated 2.8. Thus, in general neither the interface nor the details associated with each object were considered useful in exploring the content. There may be many reasons for this including the limited amount of information associated with an object and the very limited thumbnail associated with the original object.

Of all of the potential differences between their use in the Lab versus on the Web, most notable is no difference in terms of interesting objects saved. The differences appear at the level of interactivity – both in aggregate and in use of the Category Explorer, suggesting that being overseen in the lab may have changed their behavior, or doing the test off the web similarly gave them the anonymity that ensured participation without commitment. The individual lab studies in which people came into the lab should illuminate this issue.

The results presented here are descriptive and summary. What resulted from the work is a rich data set that contains both user response and log data. Unlike other tracks and/or tasks in which each lab uses the same data set to test multiple algorithms, this track *jointly* collected a data set using a common procedure and system which has resulted in a large data set that may now be used for multiple types of studies.

## 6 Conclusion and Outlook

The results of this year's CHiC lab show that multilingual information retrieval experiments are challenging not only because of the number of languages that need to be processed but also because of the number of participants necessary in order to produce comparable results. As the number of possible language variations increases

(CHiC had 13 source languages and 13 target languages), very few experiments across participants can be compared. While this year's results have shown that searching in several languages increases the overall performance (an obvious result), we could not show which languages contributed more to retrieval results. Future research in the multilingual task needs to focus on more narrowly defined tasks (e.g. particular source languages against the whole collection) or define a GRID experiment where a particular information retrieval system performs all possible run variations to arrive at better answers.

The interactive study collected a rich data set of questionnaire and log data for further use. Because the task was designed for easy entrance (predetermined system and research protocol, this is somewhat different from the traditional lab and is planned to follow a 2-year cycle (assuming the lab's continuation). In year two, the data gathered this year should be released to the community in aggregate form having been assessed by the user interaction community with the goal of identifying a set of objects that need to be developed. The intention of this second cycle is that the interactive experiment results of year one should inform system designers about which features are desirable for cultural heritage access and thus make it easier to focus development efforts into systems and interfaces. In a second year, any such developed system and interface features could be evaluated in more controlled interactive experiments. The ad-hoc retrieval tasks can benefit from the interactive task as well by re-using the real queries in ad-hoc retrieval test scenarios – effectively merging both evaluation methods.

**Acknowledgements.** This work was supported by PROMISE (Participative Research Laboratory for Multimedia and Multilingual Information Systems Evaluation), Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191. This research was supported in part by the Sciex-NMS under Grant POL 11.219. We would like to thank Europeana for providing the data for collection and topic preparation and providing valuable feedback on task refinement. We would like to thank Maria Gäde, Preben Hansen, Anni Järvelin, Birger Larsen, Simone Peruzzo, Juliane Stiller, Theodora Tsikrika and Ariane Zambiras for their invaluable help in translating the topics. We would also like to thank our relevance assessors Tom Bekers, Veronica Estrada Galinanes, Vanessa Girth, Ingvild Johansen, Georgios Katsimpras, Michael Kleineberg, Kristoffer Liljedahl, Giuliano Migliori, Christophe Onambélé, Tímea Peter, Oliver Pohl, Siri Soberg, Tanja Špec, Emma Ylitalo. Last but not least, we would like to thank all participants (either in the lab or online) in the interactive study.

## References

1. Agosti, M., Ferro, N.: Towards an Evaluation Infrastructure for DL Performance Evaluation. In: Tsakonias, G., Papatheodorou, C. (eds.) *Evaluation of Digital Libraries: An Insight to Useful Applications and Methods*, pp. 93–120. Chandos Publishing, Oxford (2009)
2. Akasereh, M., Naji, N., Savoy, J.: UniNE at CLEF – CHIC 2013. In: *Proceedings CLEF 2013, Working Notes* (2013)

3. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems* 20, 357–389 (2002)
4. Dolamic, L., Savoy, J.: Indexing and Stemming Approaches for the Czech Language. *Information Processing & Management* 45, 714–720 (2009)
5. Fernando, S., Hall, M.M., Agirre, E., Soroa, A., Clough, P., Stevenson, M.: Comparing taxonomies for organising collections of documents. In: *Proceedings of COLING 2012: Technical Papers*, pp. 879–894 (2012)
6. Hall, M.M., Toms, E.: Building a common framework for IIR evaluation. In: *CLEF 2013. LNCS*, vol. 8138, pp. 17–28. Springer, Heidelberg (2013)
7. Hall, M., Villa, R., Rutter, S., Bell, D., Clough, P., Toms, E.: Sheffield Submission to the CHiC Interactive Task: Exploring Digital Cultural Heritage. In: *Proceedings CLEF 2013, Working Notes* (2013)
8. International Council of Museums, Scope Definition of the CIDOC Conceptual Reference Model (2003), <http://www.cidoc-crm.org/scope.html>
9. Larson, R.: Pseudo-Relevance Feedback for CLEF-CHiC Adhoc. In: *Proceedings CLEF 2013, Working Notes* (2013)
10. Malak, P.: The Polish Task within Cultural Heritage in CLEF (CHiC) 2013. Torun runs. In: *Proceedings CLEF 2013, Working Notes* (2013)
11. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008)
12. Pawlowski, A.: Polish Monolingual Task within Cultural Heritage in CLEF (CHiC) 2013. Wroclaw Runs. In: *Proceedings CLEF 2013, Working Notes* (2013)
13. Petras, V., Ferro, N., Gäde, M., Isaac, A., Kleineberg, M., Masiero, I., Nicchio, M., Stiller, J.: Cultural Heritage in CLEF (CHiC) Overview 2012. In: *Proceedings CLEF-2012, Working Paper* (2012)
14. Petras, V., Bogers, T., Ferro, N., Masiero, I.: CHiC Multilingual Task Overview and Analysis. In: *Proceedings CLEF 2013, Working Notes* (2013)
15. Popescu, A.: CEA LIST’s participation at the CLEF CHiC 2013. In: *Proceedings CLEF 2013, Working Notes* (2013)
16. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. *Information Processing & Management* 36, 95–108 (2000)
17. Skov, M., Bogers, T., Lund, H., Jensen, M., Wistrup, E., Larsen, B.: RSLIS/AAU at CHiC 2013. In: *Proceedings CLEF 2013, Working Notes* (2013)
18. Stiller, J., Gäde, M., Petras, V.: Ambiguity of Queries and the Challenges for Query Language Detection. In: *CLEF 2010 LABs and Workshops* (2010), [http://clef2010.org/resources/proceedings/clef2010labs\\_submission\\_41.pdf](http://clef2010.org/resources/proceedings/clef2010labs_submission_41.pdf) (retrieved)
19. Swan Oscar, E.: Polish Grammar in a Nutshell, <http://polish.slavic.pitt.edu/firstyear/nutshell.pdf>
20. Tan, K., Almasri, M., Chevallet, J., Mulhem, P., Berrut, C.: Multimedia Information Modeling and Retrieval(MRIM)/Laboratoire d’Informatique de Grenoble (LIG) at CHiC 2013. In: *Proceedings CLEF 2013, Working Notes* (2013)
21. Tanase, D.: Using the Divergence Framework for Randomness: CHiC 2013 Lab Report. In: *Proceedings CLEF 2013, Working Notes* (2013)
22. Wilhelm-Stein, T., Schürer, B., Eibl, M.: Identifying the most suitable stemmer for the CHiC multilingual ad-hoc task. In: *Proceedings CLEF 2013, Working Notes* (2013)

# Overview of the ShARe/CLEF eHealth Evaluation Lab 2013

Hanna Suominen<sup>1,\*</sup>, Sanna Salanterä<sup>2</sup>, Sumithra Velupillai<sup>3</sup>,  
Wendy W. Chapman<sup>4</sup>, Guergana Savova<sup>5</sup>, Noemie Elhadad<sup>6</sup>,  
Sameer Pradhan<sup>5</sup>, Brett R. South<sup>7</sup>, Danielle L. Mowery<sup>8</sup>, Gareth J.F. Jones<sup>9</sup>,  
Johannes Leveling<sup>9</sup>, Liadh Kelly<sup>9</sup>, Lorraine Goeuriot<sup>9</sup>,  
David Martinez<sup>10</sup>, and Guido Zuccon<sup>11,\*\*</sup>

<sup>1</sup> NICTA and The Australian National University, ACT, Australia  
Hanna.Suominene@nicta.com.au

<sup>2</sup> University of Turku, Finland  
sansala@utu.fi

<sup>3</sup> DSV Stockholm University, Sweden  
sumithra@dsv.su.se

<sup>4</sup> University of California, San Diego, CA, USA  
wwchapman@ucsd.edu

<sup>5</sup> Harvard University, MA, USA  
Firstname.Lastname@childrens.harvard.edu

<sup>6</sup> Columbia University, NY, USA  
noemie@dbmi.columbia.edu

<sup>7</sup> University of Utah, UT, USA  
brett.south@hsc.utah.edu

<sup>8</sup> University of Pittsburgh, PA, USA  
d1m31@pitt.edu

<sup>9</sup> Dublin City University, Ireland  
Firstname.Lastname@computing.dcu.ie

<sup>10</sup> NICTA and The University of Melbourne, VIC, Australia  
David.Martinez@nicta.com.au

<sup>11</sup> The Australian e-Health Research Centre, CSIRO, QLD, Australia  
Guido.Zuccon@csiro.au

**Abstract.** Discharge summaries and other free-text reports in health-care transfer information between working shifts and geographic locations. Patients are likely to have difficulties in understanding their content, because of their medical jargon, non-standard abbreviations, and ward-specific idioms. This paper reports on an evaluation lab with an aim to support the continuum of care by developing methods and resources that make clinical reports in English easier to understand for patients, and which helps them in finding information related to their condition. This ShARe/CLEFeHealth2013 lab offered student mentoring and shared tasks: identification and normalisation of disorders (1a and 1b) and normalisation of abbreviations and acronyms (2) in clinical

---

\* Corresponding author.

\*\* In alphabetical order, HS, SS & SV co-chaired the lab; WWC led Tasks 1 & 2, GS, NE & SP as the leaders of Task 1 and BRS & DLM as the leader of Task 2; GJFJ, JL, LK & LG led Task 3; and DM & GZ were the leaders of result evaluations.

reports with respect to terminology standards in healthcare as well as information retrieval (3) to address questions patients may have when reading clinical reports. The focus on patients' information needs as opposed to the specialised information needs of physicians and other healthcare workers was the main feature of the lab distinguishing it from previous shared tasks. De-identified clinical reports for the three tasks were from US intensive care and originated from the MIMIC II database. Other text documents for Task 3 were from the Internet and originated from the Khresmoi project. Task 1 annotations originated from the ShARe annotations. For Tasks 2 and 3, new annotations, queries, and relevance assessments were created. 64, 56, and 55 people registered their interest in Tasks 1, 2, and 3, respectively. 34 unique teams (3 members per team on average) participated with 22, 17, 5, and 9 teams in Tasks 1a, 1b, 2 and 3, respectively. The teams were from Australia, China, France, India, Ireland, Republic of Korea, Spain, UK, and USA. Some teams developed and used additional annotations, but this strategy contributed to the system performance only in Task 2. The best systems had the F1 score of 0.75 in Task 1a; Accuracies of 0.59 and 0.72 in Tasks 1b and 2; and Precision at 10 of 0.52 in Task 3. The results demonstrate the substantial community interest and capabilities of these systems in making clinical reports easier to understand for patients. The organisers have made data and tools available for future research and development.

**Keywords:** Information Retrieval, Evaluation, Medical Informatics, Test-set Generation, Text Classification, Text Segmentation.

## 1 Introduction

Discharge summaries transfer information between working shifts and geographical locations. They are written or dictated by a physician, nurse, therapist, specialist, or other clinician responsible for patient care to describe the course of treatment, the status at release, and care plans. Their primary purpose is to support the care continuum as a handover note between clinicians, but they also serve legal, financial, and administrative purposes. In several countries these documents are regulated by law. For example, in Sweden, the Patient Data Law 255/2008 and in Finland, the Statute 298/2009 on Patient Documents state that in order to ensure good care, clinical documents must cover all necessary information and adequately detail the patient's conditions, care, and recovery. This legislation also stipulates that the documents must be explicit, comprehensive, and include only generally well-known, accepted concepts and abbreviations.

However, the law and practice differ substantially [1, 2]. The patient and her next of kin are likely to have difficulties in understanding this simple example sentence from a US discharge: “*AP: 72 yo f w/ ESRD on HD, CAD, HTN, asthma p/w significant hyperkalemia & associated arrhythmias.*” After expanding the abbreviations and acronyms as well as correcting the misspellings, they are much more likely to understand that this sentence belongs to the description of the patient's *active problem*. It tells that the patient is a *72 year old female*

with dependence on hemodialysis, coronary heart disease, hypertensive disease, and asthma. Her current medical problem (i.e., *presenting problem*) is *significant hyperkalemia and associated arrhythmias*. An improved understanding of related concepts in discharge summaries can be achieved by normalising all health conditions to standardised, computer-processable language. In SNOMED-CT, the CUIs *C0003811*, *C0004096*, and *C0020461* correspond to synonyms of arrhythmia, asthma, and hyperkalemia, respectively.<sup>1</sup>

The patient's and her next-of-kin's understanding of health conditions can be supported not only by these expansions, corrections, and normalisations, but also by linking the words to a patient-centric search on the Internet. Already without electronic linkage with discharge summaries, nearly 70 per cent of search engine users in the USA in 2012 searched for information about health conditions [3]. In 2007, nearly 47 per cent of Europeans considered the Internet as an important source of health information [4] and over 42 per cent of Australian searches were related to health and medical information [5]. The search engine could, for example, link hyperkalemia and its synonyms to definitions in Wikipedia, Consumer Health Vocabulary, and other patient-friendly sources.<sup>2</sup> This would explain the connection between hyperkalemia and arrhythmia: *Extreme hyperkalemia (having too much potassium in the blood) is a medical emergency due to the risk of potentially fatal arrhythmias (abnormal heart rhythms)*. The engine should also assess the reliability of information (e.g., guidelines by healthcare service providers vs. uncurated but insightful experiences on discussion forums).

This paper presents an overview of the ShARe/CLEFeHealth2013 evaluation lab<sup>3</sup> to address these approaches in making clinical text easier to understand and targeting patients' information needs in search on the Internet. The novel lab aimed to develop processing techniques and data for these approaches and an evaluation setting that includes statistical metrics of correctness and end-user engagement by asking nurses and laypeople to represent patients' preferences in expansions, normalisations, and search. It offered a mentoring track for graduate students working on related fields and shared tasks on NLP and ML: identification and normalisation of disorders (1a and 1b) [6] and normalisation of abbreviations and acronyms (2) [7] in clinical reports with respect to terminology standards in healthcare as well as IR (3) [8] to address questions patients may have when reading clinical reports<sup>4</sup>. This attracted 34 teams to submit 113 systems<sup>5</sup>; demonstrated the capabilities of these systems in contributing to patients' understanding and information needs; and made data, guidelines, and tools available for future research and development. The lab workshop was in CLEF on 23–26 Sep 2013.

<sup>1</sup> Systematized Nomenclature of Medicine Clinical Terms, Concept Unique Identifiers.

<sup>2</sup> <http://en.wikipedia.org/> and <http://www.consumerhealthvocab.org/>

<sup>3</sup> [http://nicta.com.au/business/health/events/clefehealth\\_2013](http://nicta.com.au/business/health/events/clefehealth_2013), [Shared Annotated Resources](#), <http://clinicalnlpannotation.org>, and Conference and Labs of the Evaluation Forum.

<sup>4</sup> Natural Language Processing, Machine Learning, and Information Retrieval.

<sup>5</sup> Note: in this paper we refer to systems, experiments, and runs as *systems*.

## 2 Background

For over forty years, NLP and other techniques based on computational linguistics and ML have been recognised as ways to automate text analysis in healthcare. PubMed<sup>6</sup> returns 12,860 references, including pioneering studies [9–12] and recent reviews [13–18]. Some techniques have progressed from research to use in practice. As US examples, MedLEE<sup>7</sup> used in the New York Presbyterian Hospital normalises patient records to UMLS<sup>8</sup> [19] and Autocoder at the Mayo Clinic in Rochester assigns diagnosis codes to patient records, reducing workload by 80 per cent [20]. However, the development and progress has been substantially hindered, but shared tasks address these barriers [21]. The barriers can be classified as lack of access to shared data for system research, development and evaluation; insufficient common conventions and standards for data, technologies, and evaluations; the formidability of reproducibility; limited collaboration; and lack of user-centered development and scalability.

The first shared tasks related to clinical NLP were in TREC<sup>9</sup>. The 2000 Filtering Track [22] focused on building user profiles to separate relevant and irrelevant documents. Data contained around 350,000 abstracts from the MEDLINE database over five years, manually created topics, and a topic set based on the standardised MeSH.<sup>10</sup> The Genomics Track [23] had in 2003–2007 annual IR tasks on genomics data in biomedical papers and clinical reports. The tasks ranged from ad-hoc IR to classification, passage IR, and entity-based question-answering. The Medical Records Track [24] in 2011 and 2012 aimed to develop an IR technique for finding patient cohorts that are relevant to a given criteria for recruitment as populations in comparative effectiveness studies. Their data consisted of de-identified medical records, queries that resemble eligibility criteria of clinical studies, and associated relevance assessments.

In 2005, ImageCLEFmed<sup>11</sup> [25, 26] introduced annual tasks on accessing to biomedical images in papers and on the Internet. In 2005–2013, it targeted language-independent techniques for annotating images with concepts; multi-modal IR combining visual and textual features; and multilingual IR techniques.

In 2006, i2B2<sup>12</sup> [27] began its tasks on clinical NLP: text de-identification and identification of smoking status in 2006; recognition of obesity and co-morbidities in 2008; medication information extraction in 2009; concept, assertion, and relation recognition in 2010; co-reference analysis in 2011; and temporal-relation analysis in 2012. Data originated from the USA, were in English, and included approximately 1,500 de-identified discharge summaries with their annotations.

<sup>6</sup> The query of (*natural language processing*) OR (*text mining*) on 27 Jun 2013.

<sup>7</sup> Medical Language Extraction and Encoding System.

<sup>8</sup> Unified Medical Language System.

<sup>9</sup> Text Retrieval Conference, <http://trec.nist.gov/data/filtering.html>, <http://ir.ohsu.edu/genomics/>, and <http://trec.nist.gov/data/medical.html>

<sup>10</sup> Medical Literature Analysis and Retrieval System Online and Medical Subject Headings.

<sup>11</sup> <http://ir.ohsu.edu/image/>

<sup>12</sup> Informatics for Integrating Biology and the Bedside, <https://www.i2b2.org/>

Medical NLP Challenges<sup>13</sup> by the Computational Medicine Center in 2007 [28] and 2011 [29] addressed automated diagnosis coding of radiology reports and classifying the emotions found in suicide notes. In 2007, 1,954 de-identified radiology reports in English from a US radiology department for children were used. In 2011, over 1,000 suicide notes in English were used.

In 2013, the Health Design Challenge<sup>14</sup> challenged to re-imagine the visuals and layout of health/medical records. The purpose was to make the records more usable by and meaningful to patients, their families, and others who take care of them. The challenge was motivated by the continuum of care but did not address NLP and ML. Over 230 teams submitted their designs. The winning designs were announced in Jan 2013 and are showcased on the Internet.

In Nov 2012 – Feb 2013, NTCIR ran MedNLP<sup>15</sup> on information extraction from simulated medical reports in Japanese. It had text de-identification, complaint/diagnosis recognition, and open tasks.

Targeting patients' information needs through NLP, ML and IR is important, novel, and difficult. Meeting these needs is critical because of the empowering effects the right information and the negative effects missing or incorrect information may have on health outcomes. The focus on patients' and next-of-kins' information needs as opposed to the specialised information needs of healthcare workers is the main distinguishing feature of the ShARe/CLEFeHealth 2013 evaluation lab compared to previous shared tasks. This is, however, technically more difficult, as they represent a wider and more heterogeneous subject population. The variance in, for example, their health profiles, health knowledge, abilities to interpret health information, computer skills, and search queries is greater [30].

### 3 Materials and Methods

#### 3.1 Text Documents

For Tasks 1–3, de-identified clinical reports were from US intensive care and originated from the ShARe corpus<sup>16</sup> which has added layers of annotation over the clinical notes in the version 2.5 of the MIMIC II database<sup>17</sup>. The corpus consisted of discharge summaries and electrocardiogram, echocardiogram, and radiology reports. They were authored in the intensive care setting. Although the clinical reports were de-identified, they still needed to be treated with appropriate care and respect. Hence, all participants were required to register to the lab, obtain a US human subjects training certificate<sup>18</sup>, create an account

<sup>13</sup> <http://computationalmedicine.org/challenge/>

<sup>14</sup> <http://healthdesignchallenge.com>

<sup>15</sup> [NII Test Collection for IR Systems, http://mednlp.jp/medistj-en](http://mednlp.jp/medistj-en)

<sup>16</sup> <https://www.clinicalnlpannotation.org>

<sup>17</sup> [Multiparameter Intelligent Monitoring in Intensive Care, Version 2.5, http://mimic.physionet.org](http://mimic.physionet.org)

<sup>18</sup> The course was available free of charge on the Internet, for example, via the CITI Collaborative Institutional Training Initiative at <https://www.citiprogram.org/Default.asp> or the US National Institutes of Health (NIH) at <http://phrp.nihtraining.com/users/login.php>



to a password-protected site on the Internet, specify the purpose of data usage, accept the data use agreement, and get their account approved.

For Task 3, a large crawl of health resources on the Internet was used. It contained about one million documents [31] and originated from the Khresmoi project<sup>19</sup>. The crawled domains were predominantly of health and medicine sites, which were certified by the HON Foundation as adhering to the HONcode principles (appr. 60–70 per cent of the collection), as well as other commonly used health and medicine sites such as Drugbank, Diagnosia and Trip Answers.<sup>20</sup> Documents consisted of pages on a broad range of health topics and targeted at both the general public and healthcare professionals. They were made available for download on the Internet in their raw HTML format along with their URLs to registered participants on a secure password-protected server.<sup>21</sup>

### 3.2 Human Annotations, Queries, and Relevance Assessments

For Task 1, annotation of disorder mentions in clinical reports was carried out as part of the ongoing ShARe project. For this task in the evaluation lab, the focus was on the annotation of disorder mentions only. As such, there were two parts to the annotation: identifying a span of text as a disorder mention and mapping the span to a UMLS CUI. Each note was annotated by two professional coders trained for this task, followed by an open adjudication step. UMLS<sup>22</sup> represented over 130 lexicons/thesauri with terms from a variety of languages. It integrated resources used world-wide in clinical care, public health, and epidemiology. It also provided a semantic network in which every concept is represented by its CUI and is semantically typed [32]. A disorder mention was defined as any span of text which can be mapped to a concept in SNOMED-CT and which belongs to the Disorder semantic group.<sup>23</sup> A concept was in the Disorder semantic group if it belonged to one of the following UMLS semantic types: Congenital Abnormality; Acquired Abnormality; Injury or Poisoning; Pathologic Function; Disease or Syndrome; Mental or Behavioral Dysfunction; Cell or Molecular Dysfunction; Experimental Model of Disease; Anatomical Abnormality; Neoplastic Process; and Signs and Symptoms. The annotations covered about 181,000 words.

For Task 2, a gold standard of acronyms and abbreviations normalised to CUIs from the UMLS was developed. It was generated in the following three phases: First, one Australian and nine Finnish nursing professionals as well as an Australian senior researcher in clinical NLP and ML were trained for the task using

<sup>19</sup> Medical Information Analysis and Retrieval, <http://www.khresmoi.eu>

<sup>20</sup> Health on the Net, <http://www.healthonnet.org>,  
<http://www.hon.ch/HONcode/Patients-Conduct.html>, <http://www.drugbank.ca>,  
<http://www.diagnosia.com>, and <http://www.tripanswers.org>

<sup>21</sup> HyperText Markup Language and Uniform Resource Locators.

<sup>22</sup> <https://uts.nlm.nih.gov/home.html>

<sup>23</sup> Note that this definition of Disorder semantic group did not include the Findings semantic type, and as such differed from the one of UMLS Semantic Groups, available at <http://semanticnetwork.nlm.nih.gov/SemGroups>

annotation guidelines and the eHOST<sup>24</sup> annotation tool [33]; provided reports from Task 1 with disorder annotations; and instructed to span clinical acronym and abbreviations in the clinical reports. When possible, a spanned concept was assigned one CUI from the UMLS; otherwise, it was assigned “CUI-less”. Second, Phase 1 annotations were adjudicated by a US biomedical informatician as the silver standard. Third, Phase 2 annotations were adjudicated by a US biomedical informatician certified as a respiratory therapist creating the final gold standard. The Phase 3 annotations covered approximately 7,500 abbreviations in total.

For Task 3, queries and the respective result sets and relevance assessments were associated with the text documents [34]. Two Finnish nursing professionals created 55 queries from highlighted disorders identified in Task 1 (a manually extracted set). They also generated a mapping between queries and the matching clinical report in Task 1. This was provided to the participants but they were also free to use the clinical report, if they had access to them. Relevance assessments were performed by domain experts and technological experts using the Relevation system<sup>25</sup> [35] for collecting relevance assessments of documents contained in the assessment pools. Documents and queries were uploaded to the system via the Internet interface; judges could browse the uploaded documents and queries and provide their relevance assessments. The domain experts included six Finnish nursing professionals and five Australian nursing professionals or students in health sciences. The technological experts included three Irish, one Australian, and one Swedish senior researcher in clinical NLP and ML. Assessments compared the query and its mapping to the content of the retrieved document on a four-point scale (Fig. 1). The relevance of each document was assessed by one expert. The 55 queries were divided between training and testing. Assessments for the 5 training queries were performed by the same two Finnish nursing professionals who generated the queries. As we received 48 systems, we had to limit the pool depth for the test set of 50 queries and distribute the relevance assessment workload between domain experts and technological experts. System outputs for 33 test queries were assessed by the domain experts and the remaining 17 test queries by the technological experts.

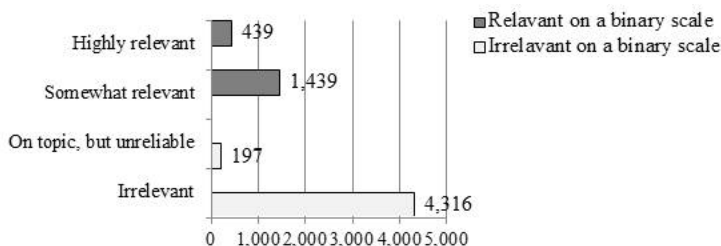
### 3.3 Evaluation Methods

The following evaluation criteria were used: correctness in identification of the character spans of disorders (1a), correctness in mapping disorders to SNOMED-CT codes (1b), correctness in mapping pre-identified acronyms/abbreviations to UMLS codes (2), and relevance of the retrieved documents to patients or their representatives.

In Tasks 1a, 1b, and 2, each participating team was permitted to upload the outputs of up to two systems. Task 1b was optional for Task 1 participants.

<sup>24</sup> Extensible Human Oracle Suite of Tools, <http://orbit.nlm.nih.gov/resource/ehost-extensible-human-oracle-suite-tools>

<sup>25</sup> <https://github.com/bevankoopman/relevation>, open source, based on Python’s Django Internet framework, uses a simple Model-View-Controller model that is designed for easy customisation and extension.



**Fig. 1.** Distribution of the relevance assessments on 4-point and binary scales

Teams were allowed to use additional annotations in their systems, but this counted towards the permitted systems; systems that used annotations outside of those provided for Tasks 1 and 2 were evaluated separately. In Task 3, teams were asked to submit up to seven ranked outputs (typically called *runs*): a mandatory baseline (referred to as  $\{\text{team}\}.\text{run1}$ ): only title and description in the query could be used without any additional resources (e.g., clinical reports, corpora, or ontologies); up to three outputs from systems which use the clinical reports (referred to as  $\{\text{team}\}.\text{run2}$ – $\{\text{team}\}.\text{run4}$ ); and up to three outputs from systems which do not use the clinical reports (referred to as  $\{\text{team}\}.\text{run5}$ – $\{\text{team}\}.\text{run7}$ ). One of the runs 2–4 and one of the runs 5–7 needed to use only the fields title and description from the queries. The ranking corresponded to priority (referred to as  $\{\text{team}\}.\{\text{run}\}.\{\text{rank}\}$  with ranks 1–7 from the highest to lowest priority).

Teams received training and test datasets in Feb–May, 2013. The evaluation for all tasks was conducted using the blind, withheld test data (reports for Tasks 1 and 2 and queries for Task 3). Teams were asked to stop development as soon as they downloaded the test data. The training set (test set) was released on 15 Feb (17 Apr), 21 Mar (1 May), and 25 Mar – 15 Apr (24 Apr) for Tasks 1, 2, and 3, respectively. Outputs for the test set were due by (evaluation results were announced to the participants on) 24 Apr (14 May), 8 May (17 May), and 1 May (1 Jun) to Tasks 1, 2, and 3, respectively.

In Tasks 1 and 2, participants were provided a training set containing clinical text as well as pre-annotated spans and named entities for disorders (Tasks 1a and 1b) or acronyms/abbreviations (Task 2). For Task 1a, participants were instructed to develop a system that predicts the spans for disorder named entities. For Tasks 1b and 2, participants were instructed to develop a system that predicts the SNOMED-CT (Task 1b) or UMLS (Task 2) CUI code (or CUI-less) for unknown pre-annotated spans. The outputs needed to follow the annotation format. The corpus of reports was split into 200 training and 100 testing.

In Task 3, post-submission relevance assessment of systems trained on the 5 training queries and the matching result set was conducted on the 50 test queries to generate the complete result set. The outputs needed to follow the TREC format. The top ten documents obtained from the participants' baseline, the highest priority output from the runs 2–4, and the highest priority output from the runs 5–7 were pooled with duplicates removed. This resulted in a pool

of 6,391 documents (Fig. 1). Pooled sets for the training queries were created by merging the top 30 ranked documents returned by the two IR models (Vector Space Model [36] and BM25 [37]) and removing duplicates.

The system performance was evaluated against the criteria by using the F1 score in Task 1a, Accuracy in Tasks 1b and 2, and Precision at 10 in Task 3. We relied on non-parametric statistical significance tests called random shuffling [38] in Tasks 1 and 2, and the Wilcoxon test [39] in Task 3 to better compare the measure values for the systems and benchmarks.

In Task 1a, the F1 score was defined as the harmonic mean of Precision (P) and Recall (R); P as  $n_{TP}/(n_{TP} + n_{FP})$ ; R as  $n_{TP}/(n_{TP} + n_{FN})$ ;  $n_{TP}$  as the number of instances, where the spans identified by the system and gold standard were the same;  $n_{FP}$  as the number of spurious spans by the system; and  $n_{FN}$  as the number of missing spans by the system. We referred to the Exact (Relaxed) F1-score if the system span is identical to (overlaps) the gold standard span.

In Tasks 1b and 2, the Accuracy was defined as the number of pre-annotated spans with correctly generated code divided by the total number of pre-annotated spans. In both tasks, the Exact Accuracy and Relaxed Accuracy were measured. In the Exact Accuracy for Task 1b, *total* was defined as the total number of gold standard named entities. In this case, the system was penalised for incorrect code assignment for annotations that were not detected by the system. In the Relaxed Accuracy for Task 1b, *total* was defined as the total number of named entities with strictly correct span generated by the system. In this case, the system was only evaluated on annotations that were detected by the system. In the Exact Accuracy for Task 2, correctly generated code was defined as the total number of pre-annotated acronyms/abbreviations with the top code selected by Phase 2 annotator from Phase 1 annotations (the best). In the Relaxed Accuracy for Task 2, *correctly generated code* was defined as the total number of pre-annotated acronyms/abbreviations for which the code is contained in a list of possibly matching codes generated by the Phase 2 and 3 annotators (*n*-best).

In Task 3, the official primary and secondary measures were P@10 and NDCG@10 [40], respectively.<sup>26</sup> Both measures were calculated over the top ten documents retrieved by a system for each query, and then averaged across the whole set of queries. To compute P@10, graded relevance assessments were converted to a binary scale (Fig. 1); NDCG@10 was computed using the original relevance assessments on a 4-point scale. The `trec_eval` evaluation tool<sup>27</sup> was used to calculate these evaluation measures<sup>28</sup>. Participants were also provided with other standard measures calculated by `trec_eval`<sup>29</sup>.

The organisers provided the following evaluation tools on the Internet: an evaluation tool for calculation of the evaluation measures of Tasks 1a, 1b, and

<sup>26</sup> Precision at 10 and Normalised Discounted Cumulative Gain at 10.

<sup>27</sup> [http://trec.nist.gov/trec\\_eval/](http://trec.nist.gov/trec_eval/)

<sup>28</sup> NDCG was computed with the standard settings in `trec_eval`, and by running the command `trec_eval -c -M1000 -m ndcg_cut qrels runName`

<sup>29</sup> Including P@5, NDCG@5, Mean Average Precision (MAP), and `rel_ret` (i.e., the total number of relevant documents retrieved by the system over all queries).

2 as well as printing the results to a file; a Graphical User Interface (GUI) for calculation of the evaluation measures of Tasks 1a, 1b, and 2, as well as for visualisation of system annotations against gold standard annotations; and a pointer to the `trec_eval` evaluation tool.

## 4 Results

The number of people who registered their interest in Tasks 1, 2, and 3 was 64, 56, and 55, respectively, and in total 34 teams with 18 unique affiliations submitted to the shared tasks (Table 1). No team participated in all three tasks. Teams represented China, France, India, Ireland, Republic of Korea, Spain, UK, 2 Australian states, and 8 US states. They had from 1 to 7 members (mean = 3.15, median = 3, and standard deviation = 1.52).

Teams submitted 113 systems (Table 2). 27 (7) were for Task 1a without (with) additional annotations. 21 (5) were for Task 1b without (with) additional annotations. 3 (2) were for Task 2 without (with) external annotations. 9 were participants' baseline systems for Task 3. In Task 3, 23 systems were not using the clinical reports nor additional annotations; 15 systems were using the clinical reports but without external annotations; and 1 system was using additional annotations but no clinical reports.

The number of teams that participated in Task 1a was 22. 5 of them were using additional annotations. 17 teams took the optional Task 1b. 4 of these teams were using additional annotations. 5 teams participated in Task 2, with 2 using additional annotations. 2 of the teams that participated in Task 2 also took Task 1a (but not Task 1b). 9 teams participated in Task 3 and only one of them was using additional annotations. All 9 participating teams submitted a baseline and systems not using the clinical reports nor additional annotations. 5 of the 9 teams also submitted systems using the clinical reports but without external annotations. 1 team submitted systems using external annotations but no clinical reports. 1 team participated in Tasks 2 and 3 and 1 team participated in Tasks 1a and 3, but these teams did not take any other tasks.

The best systems had an F1 score of 0.75 (0.80 Precision, 0.71 Recall) in Task 1a; Accuracies of 0.59 and 0.72 (0.66 without additional annotations) in Tasks 1b and 2; and P@10 of 0.52 in Task 3 (Tables 3–5). The use of additional annotations contributed to the system performance only in Task 2. In Task 3, the best system used the clinical reports. The best system that did not use the clinical reports came from the same team and had P@10 of 0.50.

The goal of the student mentoring track was to aid graduate students, regardless of which stage in their education they were in, and to provide additional feedback as a complement to their original advisors. This track was aimed at graduate students who would like to present and get more in-depth feedback on work related to the ShARe/CLEFeHealth2013 shared tasks or other relevant work in this research area, and included a peer-review process along with the assignment of one mentor (senior researcher) to provide constructive feedback in the CLEF conference on an extended abstract submission (2 pp.). The track received one submission.

**Table 1.** Participating teams. Some teams evolved during the shared tasks. For example, the Western Virginia University (WVU) had first six student-lecturer teams in Task 1 (i.e., WVU.AJ&VJ, WVU.AL&VJ, WVU.DG&VJ, WVU.FP&VJ, WVU.RK&VJ, and WVU.SS&VJ). Then the six students and their lecturer combined their forces for Task 2. Moreover, from the same organisation, many teams with changing team members participated (e.g., teams AEHRC and Mayo). Finally two teams with no members in common from the Seoul National University College of Medicine (i.e., MEDINFO and SNUBME) participated, knowing or not knowing about each other. In order to ease comparisons of organisations and countries, the organisers renamed the teams based on their affiliation (e.g., SNUBME.A for MEDINFO and SNUBME.B for SNUBME). This renaming was based on author names, affiliations, and team descriptions associated with the team submission.

ID	Organisers' team	Original team	Affiliation	Location	Number of participants
1	AEHRC.A	AEHRC	The Australian e-Health Research Centre, QLD, CSIRO, and Queensland University of Technology	Australia	3
2	AEHRC.B	AEHRC	"	"	3 (1 + 2 from A)
3	CLEAR	CLEAR	University of Colorado Boulder	CO, USA	1
4	CORAL	CORAL	The University of Alabama at Birmingham	AL, USA	3
5	HealthLanguageLABS	HealthLanguageLABS	Health Language Laboratories and The University of Sydney	NSW, Australia	3
6	KPSCMI	KPSCMI	Kaiser Permanente	CA, USA	3
7	LIMSI	LIMSI	LIMSI-CNRS	France	5
8	Mayo.A	Mayo	Mayo Clinic	MN, USA	4
9	Mayo.B	Mayo	"	"	5 (A + 1)
10	Mayo.C	Mayo	"	"	5 [3 + 2 from B (with 1 from A)]
11	NCBI	NCBI	National Center for Biotechnology Information, NLM/NIH/HHS	MD USA	3

Table 1. (*Continued*)

ID	Organisers' team	Original team	Affiliation	Location	Number of participants
12	NIL-UCM	NIL-UCM	Universidad Complutense de Madrid	Spain	4
13	OHSU	ohsu	Oregon Health and Science University	OR, USA	3
14	QUT	QUT-TOPSIG	Queensland University of Technology	QLD, Australia	1
15	RelAgent	RelAgent	RelAgent Private Lt	India	2
16	SNUBME.A	SNUBME	Seoul National University College of Medicine	Republic of Korea	3
17	SNUBME.B	MEDINFO	"	"	2
18	THCIB.A	THCIB	Tsinghua University and Canon Information Technology (Beijing)	China	4
19	THCIB.B	THCIB	"	"	6 (A + 2)
20	UCDCSI.A	UCDCSI	University College Dublin	Ireland	3
21	UCDCSI.B	UCDCSI	"	"	2 (A - 1)
22	UCSC.CW&RA	UCDCSI	University of California, Santa Cruz	CA, USA	2
23	UCSC.KC&RA	KC	"	"	2
24	UOG	uogTr	University of Glasgow	UK	1
25	UTHealthCCB.A	UTHealth_CCB	The University of Texas Health Science Center at Houston	TX, USA	4
26	UTHealthCCB.B	UTHealth_CCB	"	"	5 (2 + 3 from A)
27	UTHealthCCB.C	UTHealth_CCB	"	"	6 (A + 2)
28	WVU	WVU	West Virginia University	VW, USA	7
29	WVU.AJ&VJ	ArvindWVU	"	"	2
30	WVU.AL&VJ	alamb	"	"	2
31	WVU.DG&VJ	Diganesan	"	"	2
32	WVU.FP&VJ	FAYOLA	"	"	2
33	WVU.RK&VJ	Rahul	"	"	2
34	WVU.SS&VJ	steven_seeger	"	"	2

**Table 2.** The tasks that the teams participated in. The suffix “.add” refers to using additional annotations. In Task 3, “\*” indicates that clinical reports were used. The CORAL systems for Task 1b were not in the results announced on May 14 due to a missing registration until 17 Jun.

ID Team	Number of submitted systems per task										
	1a	1a.add	1b	1b.add	2	2.add	3 baseline	3	3*	3.add	
1 AEHRC.A	2		2								
2 AEHRC.B							1		3		
3 CLEAR	2		2								
4 CORAL	2		2								
5 HealthLanguageLABS	1				1						
6 KPSCMI	2		1								
7 LIMSI	2				1						
8 Mayo.A	1		2								
9 Mayo.B	1										
10 Mayo.C							1		3	3	
11 NCBI	2		2								
12 NIL-UCM	2		2								
13 OHSU							1		1	1	
14 QUT							1		2	3	
15 RelAgent		2									
16 SNUBME.A	2										
17 SNUBME.B							1		3	3	
18 THCIB.A		1		1							
19 THCIB.B						1	1		3	3	
20 UCDCSI.A	2										
21 UCDCSI.B			2								
22 UCSC.CW&RA		2		2							
23 UCSC.KC&RA							1		2	3	
24 UOG							1		3		
25 UHealthCCB.A	2		2								
26 UHealthCCB.B					1						
27 UHealthCCB.C							1		3		
28 WVU						1					
29 WVU.AJ&VJ	1		1								
30 WVU.AL&VJ		1		1							
31 WVU.DG&VJ	1		1								
32 WVU.FP&VJ	1		1								
33 WVU.RK&VJ		1		1							
34 WVU.SS&VJ	1		1								
Systems:	27	7	21	5	3	2	9		23	15	1
Teams:	17	5	13	4	3	2	9		9	5	1



**Table 3.** Evaluation in Task 1a. For the column of Strict F1 score, “\*” indicates that the F1 score of the system was significantly better than the one immediately below (random shuffling,  $p < 0.01$ ).

System ID ({team}.{system})	Strict Evaluation			Relaxed Evaluation		
	Precision	Recall	F1 score	Precision	Recall	F1 score
<i>No additional annotations:</i>						
(UTHealthCCB.A).2	0.800	0.706	0.750*	0.925	0.827	0.873
(UTHealthCCB.A).1	0.831	0.663	0.737*	0.954	0.774	0.854
NCBI.1	0.768	0.654	0.707*	0.910	0.796	0.849
NCBI.2	0.757	0.658	0.704*	0.904	0.805	0.852
CLEAR.2	0.764	0.624	0.687*	0.929	0.759	0.836
(Mayo.A).1	0.800	0.573	0.668*	0.936	0.680	0.787
(UCDCSI.A).1	0.745	0.587	0.656	0.922	0.758	0.832
CLEAR.1	0.755	0.573	0.651*	0.937	0.705	0.804
(Mayo.B).1	0.697	0.574	0.629*	0.939	0.766	0.844
CORAL.2	0.796	0.487	0.604	0.909	0.554	0.688
HealthLanguageLABS.1	0.686	0.539	0.604*	0.912	0.701	0.793
LIMSI.2	0.814	0.473	0.598*	0.964	0.563	0.711
LIMSI.1	0.805	0.466	0.590	0.962	0.560	0.708
(AEHRC.A).2	0.613	0.566	0.589*	0.886	0.785	0.833
(WVU.DG&VJ).1	0.614	0.505	0.554*	0.885	0.731	0.801
(WVU.SS&VJ).1	0.575	0.496	0.533	0.848	0.741	0.791
CORAL.1	0.584	0.446	0.505	0.942	0.601	0.734
NIL-UCM.2	0.617	0.426	0.504	0.809	0.558	0.660
KPSCMI.2	0.494	0.512	0.503*	0.680	0.687	0.684
NIL-UCM.1	0.621	0.416	0.498	0.812	0.543	0.651
KPSCMI.1	0.462	0.523	0.491*	0.651	0.712	0.680
(AEHRC.A).1	0.699	0.212	0.325*	0.903	0.275	0.422
(WVU.AJ&VJ).1	0.230	0.318	0.267*	0.788	0.814	0.801
UCDCSI.2	0.268	0.175	0.212*	0.512	0.339	0.408
SNUBME.2	0.191	0.137	0.160*	0.381	0.271	0.317
SNUBME.1	0.302	0.026	0.047	0.504	0.043	0.079
(WVU.FP&VJ).1	0.024	0.446	0.046	0.088	0.997	0.161
<i>Additional annotations:</i>						
(UCSC.CW&RA).2	0.732	0.621	0.672	0.883	0.742	0.806
(UCSC.CW&RA).1	0.730	0.615	0.668*	0.887	0.739	0.806
RelAgent.2	0.651	0.494	0.562*	0.901	0.686	0.779
RelAgent.1	0.649	0.450	0.532	0.913	0.636	0.750
(WVU.AL&VJ).1	0.492	0.558	0.523*	0.740	0.840	0.787
(THCIB.A).1	0.445	0.551	0.492*	0.720	0.713	0.716
(WVU.RK&VJ.1	0.397	0.465	0.428	0.717	0.814	0.762

**Table 4.** Evaluation in Tasks 1b and 2. For the column of Strict Accuracy, “\*” indicates that the Accuracy of the system was significantly better than the one immediately below (random shuffling,  $p < 0.01$ ).

System ID ({team}.{system})	Strict Accuracy	Relaxed Accuracy
<i>Task 1b, no additional annotations:</i>		
NCBI.2	0.589*	0.895
NCBI.1	0.587*	0.897
(Mayo.A).2	0.546*	0.860
(UTHealthCCB.A).1	0.514*	0.728
(UTHealthCCB.A).2	0.506	0.717
(Mayo.A).1	0.502*	0.870
KPSCMI.1	0.443*	0.865
CLEAR.2	0.440*	0.704
CORAL.2	0.439*	0.902
CORAL.1	0.410*	0.921
CLEAR.1	0.409*	0.713
NIL-UCM.2	0.362	0.850
NIL-UCM.1	0.362*	0.871
(AEHRC.A).2	0.313*	0.552
(WVU.SS&VJ).1	0.309	0.622
(UCDCSI.B).1	0.299*	0.509
(WVU.DG&VJ).1	0.241	0.477
(AEHRC.A).1	0.199*	0.939
(WVU.AJ&VJ).1	0.142	0.448
(WVU.FP&VJ).1	0.112*	0.252
(UCDCSI.B.2)	0.006	0.035
<i>Task 1b, additional annotations:</i>		
(UCSC.CW&RA).2	0.545*	0.878
(UCSC.CW&RA).1	0.540*	0.879
(THCIB.A).1	0.470*	0.853
(WVU.AL&VJ).1	0.349*	0.625
(WVU.RK&VJ).1	0.247	0.531
<i>Task 2, no additional annotations:</i>		
(UTHealthCCB.B).1	0.719*	0.725
(UTHealthCCB.B).2	0.683*	0.689
LIMS1.1	0.664*	0.672
TeamHealthLanguageLABS.1	0.467	0.488
<i>Task 2, additional annotations:</i>		
(THCIB.B).1	0.657*	0.685
WVU.1	0.426	0.448

**Table 5.** Evaluation in Task 3. Result which are significantly worse than the baseline for P@10 are indicated by "\*" (Wilcoxon test with 95% confidence). No submitted results are significantly better than the baseline. BM25 is the baseline provided by the organisers, using BM25 retrieval model and relevance feedback (BM25\_FB). The format of Run ID ({team}. {run}. {rank}) is defined in Section 3.3. The best P@10 values for each team is *emphasised*.

Run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	rel_ret
(Mayo.C).1.3	0.4800	0.4720	0.4370	0.4408	0.3040	1619
(Mayo.C).2.3	0.4960	<i>0.5180</i>	0.4391	0.4665	0.3108	1673
(Mayo.C).3.3	0.5280	0.4880	0.4742	0.4584	0.2900	1689
(Mayo.C).4.3	0.5240	0.4820	0.4837	0.4637	0.2967	1689
(Mayo.C).5.3	0.5120	<i>0.5040</i>	0.4645	0.4618	0.3061	1689
(Mayo.C).6.3	0.5160	0.4940	0.4639	0.4579	0.2953	1689
(Mayo.C).7.3	0.4920	0.4700	0.4348	0.4332	0.2981	1689
(AEHRC.B).1.3	0.4440	<i>0.4540</i>	0.3814	0.3980	0.2462	1286
(AEHRC.B).5.3	0.4560	<i>0.4840</i>	0.3957	0.4226	0.2732	1495
(AEHRC.B).6.3	0.4440	0.4240	0.4117	0.3993	0.2442	1477
(AEHRC.B).7.3	0.2080	0.2200*	0.1926	0.1984	0.1589	1425
(SNUBME.B).1.3	0.4600	<i>0.4800</i>	0.4189	0.4377	0.3131	1663
(SNUBME.B).2.3	0.4040	0.3980*	0.3467	0.3546	0.2454	1609
(SNUBME.B).3.3	0.4280	0.4040*	0.3703	0.3639	0.2584	1622
(SNUBME.B).4.3	0.4200	<i>0.4060*</i>	0.3667	0.3691	0.2601	1618
(SNUBME.B).5.3	0.3960	0.4040*	0.3407	0.3561	0.2426	1609
(SNUBME.B).6.3	0.3880	0.3600*	0.3326	0.3284	0.2343	1605
(SNUBME.B).7.3	0.3560	0.3480*	0.3061	0.3075	0.2174	1551
UOG.1.3	0.4240	<i>0.4360</i>	0.3708	0.3807	0.2438	1005
UOG.5.3	0.4280	<i>0.4400</i>	0.3663	0.3840	0.2429	983
UOG.6.3	0.4120	0.4040	0.3470	0.3528	0.2186	978
UOG.7.3	0.3640	0.3500*	0.3229	0.3207	0.1923	961
(THCIB.B).1.3	0.4360	0.3960*	0.3923	0.3716	0.1028	198
(THCIB.B).2.3	0.4440	0.3980	0.4026	0.3808	0.1106	199
(THCIB.B).3.3	0.4400	0.4020	0.3966	0.3811	0.1031	201
(THCIB.B).4.3	0.3160	0.3080*	0.2800	0.2910	0.0786	154
(THCIB.B).5.3	0.4800	<i>0.4200</i>	0.4352	0.4044	0.1217	210
(THCIB.B).6.3	0.4560	<i>0.4140</i>	0.4100	0.3904	0.1155	207
(THCIB.B).7.3	0.3360	0.3080*	0.2984	0.2928	0.0729	154
(UCSC.KC&RA).1.3	0.4040	<i>0.4040*</i>	0.3587	0.3637	0.2666	1646
(UCSC.KC&RA).2.3	0.0720	0.0600*	0.0589	0.0548	0.0178	217
(UCSC.KC&RA).3.3	0.2040	0.1920*	0.1759	0.1765	0.1590	1465
(UCSC.KC&RA).4.3	0.2520	0.2320*	0.2133	0.2062	0.1634	1433
(UCSC.KC&RA).5.3	0.0680	0.0580*	0.0586	0.0549	0.0197	250
(UCSC.KC&RA).6.3	0.3440	<i>0.3640*</i>	0.3144	0.3281	0.2270	1561
(UTHealthCCB.C).1.3	0.3920	<i>0.3740</i>	0.3444	0.3406	0.1482	458
(UTHealthCCB.C).5.3	0.2600	0.2540*	0.2681	0.2587	0.0953	296
(UTHealthCCB.C).6.3	0.2760	<i>0.2560*</i>	0.2384	0.2337	0.1124	337
(UTHealthCCB.C).7.3	0.1680	0.1460*	0.1442	0.1368	0.0546	204
QUT.1.3	0.3680	<i>0.3620*</i>	0.3376	0.3419	0.2014	1492
QUT.2.3	0.3680	<i>0.3640*</i>	0.3281	0.3368	0.2009	1492
QUT.3.3	0.3200	0.3320*	0.2808	0.2948	0.1872	1458
QUT.4.3	0.0720	0.0560*	0.0669	0.0617	0.0342	450
QUT.5.3	0.3200	0.3320*	0.2808	0.2944	0.1859	1458
QUT.6.3	0.0960	0.0900*	0.0876	0.0819	0.0745	1195
OHSU.1.3	0.2800	<i>0.2300*</i>	0.2719	0.2436	0.0953	625
OHSU.5.3	0.2840	<i>0.2600*</i>	0.2350	0.2344	0.0999	333
OHSU.6.3.add	0.1920	0.1620*	0.1895	0.1706	0.0816	461
BM25_FB	0.4840	0.4860	0.4205	0.4328	0.2945	1636
BM25	0.4520	0.4700	0.3979	0.4169	0.3043	1651

## 5 Discussion

This paper reported on a novel evaluation lab with an aim to support the continuum of care by developing methods and resources that make clinical reports in English easier to understand for patients. This ShARe/CLEFeHealth2013 lab had a mentoring track for graduate students and three shared tasks: identification and normalisation of disorders in clinical reports with respect to terminology standards in healthcare; normalisation of abbreviations and acronyms in clinical reports with respect to terminology standards in healthcare; and IR to address questions patients may have when reading clinical reports. The focus on patients' information needs as opposed to the specialised information needs healthcare workers was the main distinguishing feature of the lab from previous shared tasks on NLP and ML. The lab attracted a substantial amount of interest and demonstrated the capabilities of submitted systems and participating teams in making clinical reports easier to understand for patients. Over 30 teams from America, Asia, Australia, and Europe submitted altogether 113 systems to the shared tasks. The best systems had the F1 score of 0.75 in Task 1a; Accuracies of 0.59 and 0.72 in Tasks 1b and 2; and Precision at 10 of 0.52 in Task 3.

The significance of the lab was emphasised by the organisers' making the text documents, annotations, queries, mappings between queries and the matching clinical report, the matching result sets, relevance assessments, and evaluation tools available for future research and development. The lab developed new annotated datasets, including English text from clinical reports and the Internet. De-identified clinical reports for Task 1–3 were from US intensive care and Task 3 also used documents from the Internet. Task 1 annotations originated from the ShARe annotations, but for Tasks 2 and 3, new annotations, queries, and relevance assessments were created. Guidelines<sup>30</sup> for human subjects training, ethics clearance, research permission, registration, user access, data/annotation format, tools, and contact people were made available.

These three tasks have all aimed at supporting the patient, potential patient or next of kin to understand and have a better picture of their health condition. By working towards easier-to-understand translations of clinical text, we support the patient empowerment and patients' ability to make informed decisions concerning their own health and care.

**Acknowledgement.** The ShARe/CLEFeHealth2013 evaluation lab has been supported in part by (in alphabetical order) NICTA, funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program; the CLEF Initiative; the European Science Foundation (ESF) project ELIAS, Evaluating Information Access Systems; the Khresmoi project, funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 257528; the ShARe project funded by the United States National Institutes of Health (R01GM090187); the US

<sup>30</sup> <https://sites.google.com/site/shareclefehealth/>

Department of Veterans Affairs (VA) Consortium for Healthcare Informatics Research (CHIR); the US Office of the National Coordinator of Healthcare Technology, Strategic Health IT Advanced Research Projects (SHARP) 90TR0002; and the Vårdal Foundation (Sweden).

We acknowledge the generous support of time and expertise that the annotators as well as members of the organising and mentoring committees have invested in this first pilot year of the evaluation lab. We want to thank the following individuals: Allan Hanbury (Vienna University of Technology, Austria); Anni Järvelin and Dimitrios Kokkinakis (University of Gothenburg, Sweden); Digvijay Khangarot, Thomas Souchen, Timothy Sladden, and Warren Brown (Australian E-Health Research Centre, CSIRO, QLD, Australia); Erika Siirala, Filip Ginter, Heljä Lundgren-Laine, Jenni Lahdenmaa, Laura Maria Murtola, Lotta Kauhanen, Marita Ritmala-Castren, Riitta Danielsson-Ojala, Saija Heikkinen, and Sini Koivula (University of Turku, Finland); Jussi Karlgren (Gavagai and KTH Royal Institute of Technology, Sweden); Hans Moen (Norwegian University of Science and Technology, Norway); Henning Müller (University of Applied Sciences Western Switzerland); Hercules Dalianis (DSV Stockholm University, Sweden); Maricel Angel (NICTA, ACT, Australia); Özlem Uzuner (State University of New York, NY, USA); Pamela Forner (CELCT Center for the Evaluation of Language and Communication Technologies and CLEF, Trento, Italy); Preben Hansen (Stockholm University, Sweden); Qing Treitler Zeng and Tyler Forbush (University of Utah, SLC VA, UT, USA); and Rune Saetre (Norwegian University of Science and Technology, Norway).

## References

1. Allvin, H., Carlsson, E., Dalianis, H., Danielsson-Ojala, R., Daudaravicius, V., Hassel, M., Kokkinakis, D., Lundgren-Laine, H., Nilsson, G., Nytro, O., Salanterä, S., Skeppstedt, M., Suominen, H., Velupillai, S.: Characteristics of Finnish and Swedish intensive care nursing narratives: A comparative analysis to support the development of clinical language technologies. *Journal of Biomedical Semantics* 2(suppl. 3), S1 (2011)
2. Suominen, H. (ed.): *The Proceedings of the CLEFeHealth2012 — the CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis*. NICTA (2012)
3. Fox, S.: *Health Topics: 80% of internet users look for health information online*. Technical report, Pew Research Center (February 2011)
4. Kummervold, P., Chronaki, C., Lausen, B., Prokosch, H., Rasmussen, J., Santana, S., Staniszewski, A., Wangberg, S.: eHealth trends in Europe 2005–2007: A population-based survey. *Journal of Medical Internet Research* 10(4), e42 (2008)
5. Experian Hitwise: Google Receives 87.81 Percent of Australian Searches in June 2008 (2008), <http://www.hitwise.com/au/press-centre/press-releases/2008/ap-google-searches-for-june/>
6. Pradhan, S., Elhadad, N., South, B., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W., Savova, G.: Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In: *Online Working Notes of CLEF, CLEF (2013)*

7. Mowery, D., South, B., Christensen, L., Murtola, L., Salanterä, S., Suominen, H., Martinez, D., Elhadad, N., Pradhan, S., Savova, G., Chapman, W.: Task 2: ShARe/CLEF eHealth Evaluation Lab 2013. In: Online Working Notes of CLEF, CLEF (2013)
8. Goeriot, L., Jones, G., Kelly, L., Leveling, J., Hanbury, A., Müller, H., Salanterä, S., Suominen, H., Zucco, G.: ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. In: Online Working Notes of CLEF, CLEF (2013)
9. Becker, H.: Computerization of patho-histological findings in natural language. *Pathologia Europaea* 7(2), 193–200 (1972)
10. Anderson, B., Bross, I., Sager, N.: Grammatical compression in notes and records: Analysis and computation. *American Journal of Computational Linguistics* 2(4), 68–82 (1975)
11. Hirschman, L., Grishman, R., Sager, N.: From text to structured information: automatic processing of medical reports. In: American Federation of Information Processing Societies: 1976 National Computer Conference. AFIPS Conference Proceedings, vol. 45, pp. 267–275. Association for Computational Linguistics, New York (1976)
12. Collen, M.: Patient data acquisition. *Medical Instrumentation* 12, 222–225 (1978)
13. Sarkar, I.: Biomedical informatics and translational medicine. *Journal of Translational Medicine* 8, 22 (2010) (review)
14. Demner-Fushman, D., Chapman, W., McDonald, C.: What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics* 42(5), 760–772 (2009) (review)
15. Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J.: Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, 128–144 (2008) (review)
16. Reiner, B., Knight, N., Siegel, E.: Radiology reporting, past, present, and future: the radiologist's perspective. *Journal of the American College of Radiology: JACR* 4(5), 313–319 (2007) (review)
17. Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salakoski, T., Salanterä, S.: Applying language technology to nursing documents: pros and cons with a focus on ethics. *International Journal of Medical Informatics* 76(suppl. 2), S293–S301 (2007) (review)
18. Zweigenbaum, P., Demner-Fushman, D., Yu, H., Cohen, K.: Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics* 8(5), 358–375 (2007) (review)
19. Mendonça, E., Haas, J., Shagina, L., Larson, E., Friedman, C.: Extracting information on pneumonia in infants using natural language processing of radiology reports. *Journal of Biomedical Informatics* 38(4), 314–321 (2005)
20. Pakhomov, S., Buntrock, J., Chute, C.: Automating the assignment of diagnosis codes to patient encounters using example based and machine learning techniques. *Journal of the American Medical Informatics Association: JAMIA* 13(5), 516–525 (2006)
21. Chapman, W., Nadkarni, P., Hirschman, L., D'Avolio, L., Savova, G., Uzuner, Ö.: Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association: JAMIA* 18, 540–543 (2011) (editorial)
22. Robertson, S., Hull, D.: The TREC-9 filtering track final report. In: NIST Special Publication 500-249: The 9th Text REtrieval Conference (TREC 9), pp. 25–40 (2000)
23. Roberts, P.M., Cohen, A.M., Hersh, W.R.: Tasks, topics and relevance judging for the TREC genomics track: five years of experience evaluating biomedical text information retrieval systems. *Information Retrieval* 12, 81–97 (2009)

24. Voorhees, E.M., Tong, R.M.: Overview of the TREC 2011 medical records track. In: Proceedings of TREC, NIST (2011)
25. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A., Tsirikla, T.: The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of CLEF 2011 (Cross Language Evaluation Forum) (2011)
26. Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds.): Experimental Evaluation in Visual Information Retrieval. The Information Retrieval Series, vol. 32. Springer (2010)
27. Uzuner, Ö., South, B., Shen, S., DuVall, S.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association: JAMIA 18, 552–556 (2011)
28. Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K., Duch, W.: A shared task involving multi-label classification of clinical free text. In: BioNLP Workshop of the Association for Computational Linguistics, pp. 97–104. Association for Computational Linguistics (2007)
29. Pestian, J., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, Ö., Wiebe, J., Cohen, K., Hurdle, J., Brew, C.: Sentiment analysis of suicide notes: A shared task. Biomedical Informatics Insights 5(suppl. 1), 3–16 (2012)
30. Boyer, C., Gschwandtner, M., Hanbury, A., Kritz, M., Pletneva, N., Samwald, M., Vargas, A.: Use case definition including concrete data requirements (D8.2). public deliverable, Khresmoi EU project (2012)
31. Hanbury, A., Müller, H.: Khresmoi – multimodal multilingual medical information search. In: MIE Village of the Future (2012)
32. Bodenreider, O., McCray, A.: Exploring semantic groups through visual approaches. Journal of Biomedical Informatics 36, 414–432 (2003)
33. South, B.R., Shen, S., Leng, J., Forbush, T.B., DuVall, S.L., Chapman, W.W.: A prototype tool set to support machine-assisted annotation. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, BioNLP 2012, pp. 130–139. Association for Computational Linguistics, Stroudsburg (2012)
34. Goeuriot, L., Kelly, L., Jones, G., Zuccon, G., Suominen, H., Hanbury, A., Müller, H., Leveling, J.: Creation of a New Evaluation Benchmark for Information Retrieval Targeting Patient Information Needs. In: Song, R., Webber, W., Kando, N., Kishida, K. (eds.) Proceedings of the 5th International Workshop on Evaluating Information Access (EVIA), A Satellite Workshop of the NTCIR-10 Conference. National Institute of Informatics/Kijima Printing, Tokyo/Fukuoka (2013)
35. Koopman, B., Zuccon, G.: Relevation! an open source system for information retrieval relevance assessment. arXiv preprint (2013)
36. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
37. Robertson, S.E., Jones, S.: Simple, proven approaches to text retrieval. Technical Report 356, University of Cambridge (1994)
38. Yeh, A.: More accurate tests for the statistical significance of result differences. In: Proceedings of the 18th Conference on Computational Linguistics (COLING), Saarbrücken, Germany, pp. 947–953 (2000)
39. Smucker, M., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM 2007), pp. 623–632. Association for Computing Machinery, New York (2007)
40. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20(4), 422–446 (2002)

# Overview of CLEF-IP 2013 Lab

## Information Retrieval in the Patent Domain

Florina Piroi, Mihai Lupu, and Allan Hanbury

Vienna University of Technology,  
Institute of Software Technology and Interactive Systems,  
Favoritenstrasse 9-11, 1040 Vienna, Austria

**Abstract.** The first CLEF-IP test collection was made available in 2009 to support research in IR methods in the intellectual property domain; only one type of retrieval task (Prior Art Search) was given to the participants. Since then the test collection has been extended with both more content and varied types of tasks, reflecting various specific parts of patent experts' workflows. In 2013 we organized two tasks – Passage Retrieval Starting from Claims and Structure Recognition – on which we report in this work.

## 1 Introduction

The patent system is designed to encourage disclosure of new technologies and novel ideas by granting exclusive rights on the use of inventions to their inventors, for a limited period of time [23]. An important requirement for a patent to be granted is that the invention it describes is novel. That is, there is no earlier patent, publication or public communication of a similar idea. To ensure the novelty of an invention, patent offices as well as other Intellectual Property (IP) service providers perform thorough searches called 'prior art searches' or 'validity searches'. Since the number of patents in a company's patent portfolio affects the company market value, well-performed prior art searches that lead to solid, difficult to challenge patents are of high importance.

Patent data has attracted researchers' interest as early as 1977 when, while studying local clustering in full-text searches using local feedback, experiments were done on a database of US patents [5]. In [5], Attar and Fraenkel did an experiment that was a 'technology survey'-like search on a set of 76 US patents. Two decades later an 'invalidity search' was performed on 60000 US patents. Similar to the Prior Art Search task in CLEF-IP 2009–2011, the topics of the invalidity search were patents and citations were used to generate relevance assessments [21].

In the last decades, research in IR methods for the IP domain has intensified. Workshops, conferences and evaluation tracks were organized in an effort to bring IR and IP communities together (see [11,13,27,10]). The National Institute of Informatics (NII), Japan, initiated a series of workshops and evaluations



using patent data as part of the NTCIR project (the NII Test Collections for IR Systems, currently renamed to the NII Testbeds and Community for Information access Research), focusing on Japanese and Chinese patents, and their translations into English.

In 2009, two further evaluation activities using patent data were launched: TREC-CHEM and CLEF-IP. TREC-CHEM ran from 2009 to 2011 and was organized as a chemical IR track in TREC (Text Retrieval Conference) addressing the challenges in chemical and patent IR [15]. The collection corpus was limited to chemical patent documents and chemical journal articles.

The purpose of the CLEF-IP track, part of the Cross-Language Evaluation Forum (CLEF), is to encourage and facilitate research in the area of multilingual patent retrieval by providing a large, clean data set for experimentation. The data set contains patents in three European languages, patents published by the European Patent Office (EPO), as well as queries and associated relevance judgements.

In 2013, the CLEF-IP lab proposed two tasks: a passage retrieval task where we asked for passages relevant to a given (set of) patent claim(s) and a structure recognition task where we asked to extract the textual representation of flowcharts occurring in patents and represented in black and white images.

## 2 The 2013 CLEF-IP Benchmark

We begin this section by establishing the patent terminology used throughout this paper and shortly describing the patenting process such that the rationale behind the lab's activities are better understood.

The main phases of obtaining a patent for an invention are<sup>1</sup>:

**The Pre-application Phase:** a person with a new idea will write down its description as detailed as necessary. Then she or he will usually perform a survey-like search in the domain of the invention. This preliminary search will allow the inventor to avoid unnecessary effort in case a similar invention already exists and will help him to draft the invention claims. The drafted document generally has three parts: an *abstract*, a *description* of the invention with technical drawings, and a *claims* section which states the extent of the protection sought for the described invention.

**The Application and Examination Phase:** after filing the invention description at a patent office the document (now called a *patent application document*) is given to a patent examiner. He or she will inspect the document and verify that it respects certain criteria, namely: novelty, the existence of a non-obvious

---

<sup>1</sup> The process described by these phases is typical for EPO patent applications. Though very similar, processes at other patent offices may reveal important differences. For example, the US Patents and Trademark Office, USPTO, makes use of Examiner's Letter or Action to record considered citations and does not publish a distinct search report ([3], chapter 707).

inventive step, and realizability. During the novelty check the patent examiner will search for and create a list of existing patents, a.k.a. prior art, that are relevant to the application document under inspection. At the EPO the list of relevant documents is published as a *search report* document. The search report contains also the relevant documents provided by the inventor as background information to the invention. In the IP vocabulary, the documents listed in the search report are called *patent citations*, the citations provided by the applicant being known, additionally, as *applicant citations*.

In this document, whenever the word ‘citation’ occurs we mean the patent citations, that is, the documents in the search reports which were considered relevant by patent examiners. This is different from the research community’s understanding of ‘citation’ which refers to later publications citing a research article. A patent citation is more similar to what in the research community is known as a reference at the end of an article<sup>2</sup>. In the IP community, differentiating between the patent citations and later references to patents is done by using the notions of *forward* and *backward citations* [1]. Given a patent application document, the patent citations listed in a search report are known as *backward citations*, while the patent application itself is a *forward citation* for any of the patents listed in the search report.

The patent citations usually have various degrees of relevance to the application document. The main three types of citations are:

- citations that describe prior work but which do not destroy the novelty of the application;
- citations which, taken alone, make a patent application not novel;
- citations that, in combination with other citations, destroy the novelty of an application.<sup>3</sup>

At the end of the examination phase the patent application document and its associated search report are published by the patent office. At the same time, the patent application is given a classification code that assigns the patent to a specific technological area<sup>4</sup>.

**The Granting and Opposition Phase:** Based on the search report a dialogue between the patent office and the patent applicant is initiated. There are various outcomes to this dialogue: an application may be retracted, rejected, or modified in order not to infringe existing patents. If the patent office reaches the decision to grant a patent, after various fee payments made by the applicant, a *patent document* is published. From this point on, for a certain amount of time (9 months at the EPO) oppositions to a granted patent may be filed to the patent

---

<sup>2</sup> This observation is critical in understanding how we have selected the topics and how the relevance judgement were created.

<sup>3</sup> The last two patent citation types are referred to as *highly relevant citations* in the CLEF-IP Labs.

<sup>4</sup> We do not expand here on the subject of patent classification codes. See [2] for a description of the classification system we mention later in this work.

office. Note that opposition procedures at a patent office are different from the legal actions to invalidate patents which are taken in justice courts.

The rest of this section describes the main connection between the CLEF-IP tasks and a patent expert's work, and the CLEF-IP test collection: document corpus, topic sets, and judgements.

## 2.1 The Retrieval Tasks

There are many aspects of the *search for innovation* use case domain that previous evaluation campaigns, including CLEF-IP, have focused on in their retrieval tasks. Creating technical surveys on various chemical subjects (TREC-CHEM [18]) or creating patent translations to be used by non-speakers of certain languages [8] are two such examples.

This year in CLEF-IP we proposed two tasks. The first one models the type of searches examiners do to establish the non-obviousness of an invention, where they closely inspect the claims in the patent application against other existing patent documents. At the EPO, search reports generally show not only the prior art documents, but also the claims in the patent application to which the patent citation pertains and which passages in the citation are particularly of interest (see Figure 1). The retrieval task was designed to investigate the degree of support an IR system offers patent experts in finding relevant documents and text passages to a set of claims in a patent application.

The second task in the lab is not one that models part of an expert's work, but it is designed to support his or her work during patent examination. Technical drawings are often crucial not only in illustrating the embodiments of an invention, but also to quickly filter out non-relevant patents by rapid glances to images in them. The aim of the structure recognition task is to make the content of the images textually searchable and comparable. Out of the many types of images that may occur in patents we limited this retrieval task to images representing flow-charts.

## 2.2 The Collection Corpus

One of our aims when embarking on the CLEF-IP endeavor was to create a test collection fit for experimenting with patent data, a collection that faithfully mirrors the features and challenges of the data used in the actual working cycles of a patent professional. For this we use actual patent documents published by the EPO and WIPO (World Intellectual Property Organization). These documents contain most of the information that is actively used by patent practitioners in their daily work with patent data.

The bulk of the collection's corpus is made up of patent documents stored as XML files. Since its first release in 2009, consecutive additions were made to the CLEF-IP test collection, so that it currently contains almost 1.5 million patents published before 2002, stored into approximately 3.5 million XML documents.

These patents are an extract from the larger MAREC<sup>5</sup> collection which contains documents representing over 19 million patents published at the EPO, USPTO, WIPO and JPO (Japan Patent Office) stored in a common normalized XML format. The main elements of the XML representations are the ones shown in the simplified listing below:

```
<patent-document>
  <bibliographic-data> ... </bibliographic-data>
  <abstract> ... </abstract>
  <description> ... </description>
  <claims> ... </claims>
</patent-document>
```

The `<abstract>`, `<description>`, and `<claims>` elements store the textual content of the disclosed invention. These fields may occur more than once when, for example, both the English and the German versions of the abstract are stored in a patent document. The abstract, description and claim fields are the parts of the patent file mostly used by the textual retrieval methods. The `<bibliographic-data>` element contains the administrative data related to a patent. In this XML element we will find the application and publication dates and references, family identifiers, the classification symbols, inventors, assignees, postal addresses of the inventors and/or assignees, the invention's title (in three languages), and the citations relevant to the invention in this document.

In the corpus of European patent documents with application date prior to 2002, a high percentage of the patent documents refer to applications internationally filed under the Patent Cooperation Treaty [22], also known as 'EuroPCTs', in which case, the EPO does not republish the whole patent application, but only a bibliographic entry linking to the original application published by the WIPO. Using text-based methods to retrieve such documents is problematic, and therefore, for these patent documents the current CLEF-IP collection contains their WIPO equivalent. Determining that the EuroPCT patent documents refer to a certain invention disclosed in a document published by WIPO is done by the family identifier which for the two documents must be the same.

One of the most important features of the CLEF-IP corpus is its multilingualism. Patent applications to the EPO are written in one of the three official EPO languages (German, English, French), with the additional requirement that, once the decision to grant a patent is made, the claims section of the patent document must be submitted in all these three languages. Although the English language is overrepresented<sup>6</sup> in the CLEF-IP collection, not least due to the EuroPCT applications written in their large majority in English, the collection entails large amounts of content that is in German and French, making the collection suitable for carrying out multilingual retrieval experiments.

<sup>5</sup> The MAtrixware REsearch Collection. <http://ifs.tuwien.ac.at/imp/marec>

<sup>6</sup> Almost 70% of the documents in the collection are written in English, about 23% have German as the document language, and about 7% are in French.

## 2.3 Passage Retrieval Starting From Claims

The topics of this retrieval task are sets of claims occurring in patent application documents. Participants were asked to return documents from the CLEF-IP corpus which were considered relevant and, within these documents, mark the most relevant passages to the set of claims.

We have provided over 150 training topics and the test set contained 149 topics. A third of both the test and the training sets contained topics in English, another third contained topics in German, and yet another third had the topic language French. We did not provide translations of topics from one language into any of the other two. The structure of a CLEF-IP topic is as follows:

```
<tid> topic_id </tid>
<tfile> patent_ucid.xml </tfile>
<tfam-docs> patent_ucid.xml </tfam-docs>
<tclaims> xpathes_to_claims </tclaims>
```

where

- `tid` is the topic identifier;
- `tfile` is the XML file which stores the source patent application;
- `tclaims` is the list of XPathes to the claims selected as topic from the source patent document;
- `tfam-docs` contains the XML files that are part of the source patent's family<sup>7</sup> and published prior to the source patent document.

Providing previously published patent documents that are family members of the source patent application is motivated by the patenting process rules and by the practices of the patent examiners at patent offices. More concretely, when an applicant files for a patent grant at, let's say, EPO he is required to provide information on whether he has already applied for a patent grant, *for the same invention*, at other patent offices in the world. Later, when the patent application is examined, the patent examiner pulls whatever search reports are available in the patent databases related to the previous publications of the inventions in order to re-use that information.

Below is an example of a topic in the CLEF-IP 2013 Passage Retrieval Task:

```
<tid>PSG-2</tid>
<tfile>EP-1445439-A1.xml</tfile>
<tfam-docs>FI-116479-B1.xml,FI-20030196-A.xml,FI-20030196-D0.xml</tfam-docs>
<tclaims>/patent-document/claims/claim[1] /patent-document/claims/claim[2]
/patent-document/claims/claim[3] /patent-document/claims/claim[4]</tclaims>
```

---

<sup>7</sup> A *patent family* denotes the collection of patent documents that refer to the same invention and are published by different patent offices around the world.

**Topic Selection.** We created a pool of patent application documents out of which we extracted the set of test topics for this task. The pool of patent applications was extracted from the MAREC collection with the requirements that:

- it was not part of the CLEF-IP corpus (i.e. published after 2002);
- it was a patent published by the EPO;
- there is at least one previously published document in the patent’s family;
- it has content for all document parts (claims, abstract, description), and the document word count is lower than 300,000 (we included the XML tags and attributes in this number)<sup>8</sup>;
- there are at least two and at most 10 citations in the corresponding search report, and the cited documents occur in the CLEF-IP corpus.

Some technological areas are overly represented in the patent domain. For example, the number of patents filed in US, last year, in the technological area of Electrical Engineering (including Computer Technologies) outnumbered the number of patents filed in any of the other technological areas [29]. To avoid overrepresentation of patents in certain technological classes in the topic set we restricted the sampling process in the following way: we grouped the documents in the pool by the number of citations the documents have, and we randomly selected 20 documents from each group, with the restriction that each selected patent belongs to a different IPC class (there are 121 IPC classes in the topic pool). At this point we have a pool of 462 patent application documents out of which to extract topics.

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X	WO 98 07379 A (LARSEN ERIC ;HOEGSETH SOLFRID (NO)) 26 February 1998 (1998-02-26)	1-7,14, 15	A61B18/20
Y	* page 5, paragraph 1 - page 6, paragraph 2; figures 2,3 *	8-11	
X	WO 01 26573 A (COHERENT INC) 19 April 2001 (2001-04-19) * page 13, line 30 - page 15, line 16; figure 3 *	1-3,7	
Y	EP 1 101 450 A (PULSION MEDICAL SYSTEMS AG) 23 May 2001 (2001-05-23) * page 5, line 9 - line 22; figure 2 *	8	

**Fig. 1.** Extract from a search report

The next step in the topic selection process is, for each patent application document in the pool, to manually retrieve its European search report

<sup>8</sup> We chose this limit in order to avoid pooling documents of excessive length which make some retrieval algorithms fail [19]. Some patent applications are more than 100 pages long which we wanted to avoid being part of the topic test set.

(that is, the search report published by the EPO, see Figure 1) and inspect each citation document with respect to the claims it is relevant to (the third column in Figure 1) and the relevant passage recorded in the report (second column in the same figure). For each citation document in the search report *and* in our data corpus, we extracted the claim numbers the citation referred to<sup>9</sup>. These formed the sets of claims for a candidate topic. Looking, now, at the passages noted as relevant, further decisions had to be made whether a candidate topic is retained. Rejecting candidate topics was done when:

- the relevant documents referred to figures only;
- there was no mention of relevant passages, or only ‘whole document’ mentions were recorded;
- the search report had the mention ‘Incomplete search’ which generally means that the patent expert, for various reasons, did not perform the prior art search for all the claims in the patent application.

From one patent application document it was possible to extract several sets of claims as topics, often with completely different sets of relevance judgments. The process just shown has been first used in the CLEF-IP 2012 Lab and is also described in [24].

This has been a lengthy process – being done manually – and we managed to inspect over 200 application documents. The final set of topics (148) was extracted out of 69 patent applications. The citation distribution for the topics and application documents is shown in Table 1, where the topic source documents belong to 66 different IPC classes.

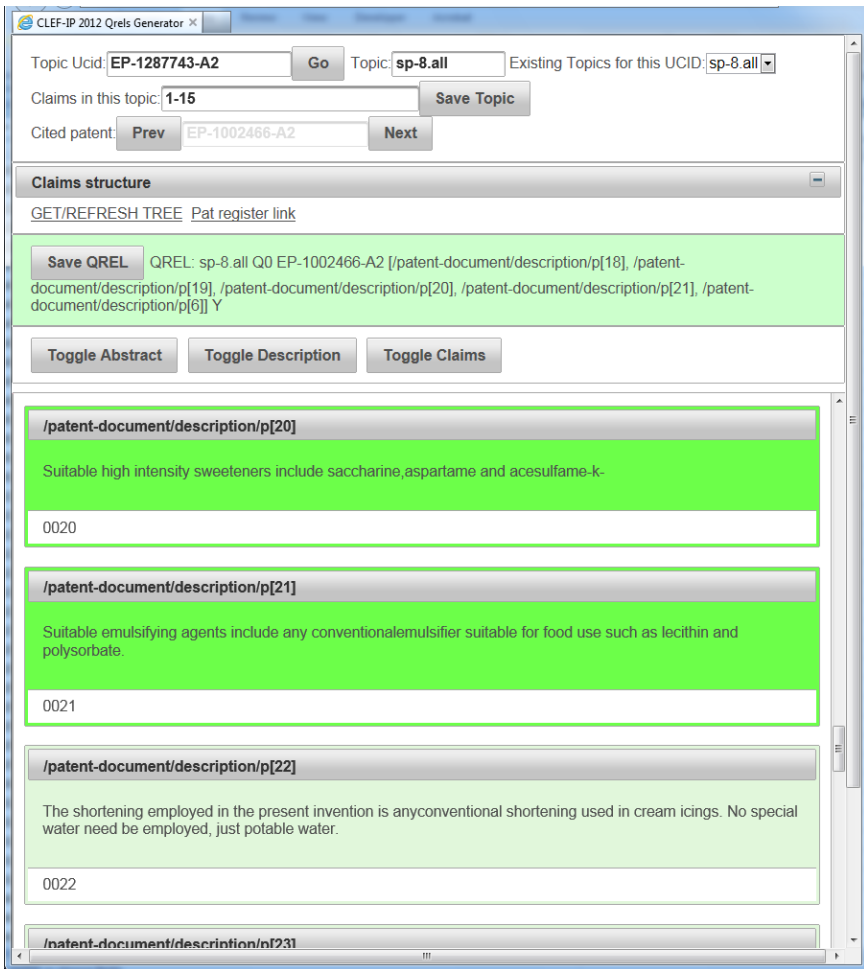
**Table 1.** Citation distribution in the topic set

Number of citations	3	4	5	6	7	8	9	10	Total
Number of topics	24	25	30	23	16	19	8	3	148
Number of documents	13	11	13	10	10	8	3	1	69

**Obtaining the Relevance Judgements.** When a topic candidate made it into the final topic set, the next phase was to create its relevance judgements. Judging the retrieved results cannot be confidently done by non-patent experts, therefore, pooling the results and judging them post-submission is not a solution that can be used in CLEF-IP, primarily because engaging patent experts is very costly for a research project. The solution chosen by us, as well as most other evaluation tracks using patent data, is to make use of the patent search reports, which constitute a very reliable source of relevance judgements. More reasons to support this decision can be found in [9].

Patent citation information can be rather easily obtained in a machine processable format (e.g. relational tables). For our task, however, we need relevant passage information which we had to extract manually by matching the passage

<sup>9</sup> In patent documents, claims are numbered for ease of reference.



**Fig. 2.** A system for extracting and storing qrels

indications in the search reports with the textual content of the patent documents in our corpus. When matched, we extracted the XPath of the identified content and saved them to a database.

To assist us in this tedious process we used an in-house developed system, developed in 2012 (Figure 2), which read the XML patent citation documents and displayed the individual XPath passages. We see in Figure 2 three screen areas: In the upper part, given a topic source document ('Topic UCID'), we define the topic id ('sp-8.all' – an intermediary topic identifier), and the claim numbers that are to be part of this topic. In the middle part of the screen we have buttons for toggling (marking as selected) all passages in the abstract, description, or claim sections at once, and a button ('Save QREL') for storing the currently selected XPaths to the database. In the lower part of the screen each textual content at the end of an XPath in the citation document selected



for the source topic id is displayed and can be selected into the topic’s qrels (the green text). Displaying the different citations for the source patent application is done with the navigation buttons ‘Prev’ and ‘Next’ of the top screen area.

Below is an excerpt from the qrel files obtained with the help of the system:

```
PSG-5 EP-1078736-A1 /patent-document/description/p[20]
PSG-5 EP-1078736-A1 /patent-document/description/p[21]
PSG-5 EP-1078736-A1 /patent-document/description/p[18]
PSG-5 EP-1078736-A1 /patent-document/description/p[15]
PSG-5 EP-1078736-A1 /patent-document/claims/claim[1]
PSG-5 EP-1078736-A1 /patent-document/abstract/p
PSG-5 EP-1078736-A1 /patent-document/claims/claim[2]
...
```

## 2.4 Structure Recognition from Patent Images

From the outset, non-textual patent content, like tables, technical drawings, formulae, was not a part of the CLEF-IP campaigns. But these non-textual items have an important role in taking quick decisions about the relevance of a document to an information need. During a patent search, plenty of documents may be returned as the result of a query. An experienced patent professional will often be able to expeditiously dismiss non-relevant documents by glances at images in the patent documents.

In 2012 we designed a task that aimed at making the patent images searchable and comparable by textual means. Two separate sets of images were given, flow-charts and chemical structures. This year we continued this task only with a set of flow-chart images that contained more complicated graphical structures than in 2012.

The topics of this task are black and white images representing flow-charts, images occurring in patents. We made available the 2012 training and test topic sets as training data (150 images). The test set we used in 2013 contains 747 images of flow-charts<sup>10</sup>. The retrieval task required the participants to extract the information stored in the image files and store it into a textual form that encoded the graph-like structure of the flow-charts, where the text is seen as node or edge labels.

**Topic Selection.** By comparison with the topic selection process in the Passage Retrieval Task, shown above, selecting the topics for the structure recognition task was ‘a walk in the park’: we re-used the set of flow-chart images that were part of the Patent Image Classification task in the CLEF-IP 2011 Lab [23]. We slightly modified the encodings used last year to accommodate for the more complicated flow-charts in this year’s topic set.

**Relevance Judgements.** Before creating the qrels we have to establish a textual encoding of the flow-charts. For the purpose of this task, we decided that a text file encoding a flow-chart is a sequence of text lines, each line being one of the below:

MT for ‘Meta’, refers to meta information in the flow-chart:

- MT Title ‘figure’s title’: title of the chart, in double quotes

<sup>10</sup> In CLEF-IP 2012, the set of flow-charts selected for the Structure Recognition Task was filtered to contain less complex flow-charts, w.r.t. type of nodes, edges, and lines enclosing other nodes—meta nodes in 2013.

- MT NO <number>: number of nodes in the flow-chart;
  - MT DE <number>: number of directed edges in the flow-chart;
  - MT UE <number>: number of undirected edges in the flow-chart
- NO for ‘Node’. Lines starting with NO describe the node of the chart. Each node description line must contain an identifier of the node (unique in the chart), a node-type that describes the shape of the node (oval, rectangle, etc.), the text of the node (empty string of no text is present), and a pair of coordinates marking the graphical location of the node’s center. The coordinates are intended for later use with graph representation tools to graphically display the encoded graph and visually compare it with the original image.
- MN for ‘Meta-node’. Lines beginning with MN describe a meta-node of the chart. Each such node must have a unique identifier (different from the NO’s identifiers), a comma separated list of NO nodes identifiers enclosed in square brackets, a text attached to the meta-node (or the empty string).
- DE |UE for ‘directed’ and ‘undirected’ edges. The lines starting with one of these identifiers describe the edges connecting the flow-chart nodes. Each such line must contain the identifiers of the start and end nodes of the edge, the type of the edge (plain, wiggly, dotted, etc.), and the label attached to the edge, if any.
- CO for ‘Comment’. These lines are not to be considered by the evaluation scripts.

Figure 3 shows an example of a flow-chart textually encoded using the format given above.

### 3 Submitted Runs

Three participants submitted a total of 19 retrieval experiments, we shortly describe the main retrieval approaches used.

**Georgetown University, USA.** The participants from Georgetown University focused on formulating representative queries using patent metadata (embedded in the collection’s XML patent documents). The queries were then submitted to a Lemur search engine [14]. Several indexes were created: one for the stemmed content words in the CLEF-IP collection, and several other for specific patent metadata (title, inventor, application date). The retrieval engines used were TF-IDF based, Language Modelling based, and Okapi BM25.

Six experiments were submitted, each of them using a different approach to obtaining query terms. Extracting the words occurring in claims, titles (experiment with the id `GU.OnlyClaimLM`, `GU.coOnlyTt1LM`), the hyphenating phrases, Part of Speech tagging (`GU.HypCoTt1NoIdfUpperBoundLM`) and a combination of `idf` filterings on the extracted query terms (`GU.HypCoTt1WithIdfUpperBoundLM`, `GU.HypDuTt1NoIdfUpperBoundBM`, `GU.HypDuTt1WithIdfUpperBoundBM`) are among the tested options. The queries thus generated were used to retrieve relevant documents. The passages in these documents were ranked using a `tf-idf` weighting scheme, returning the top 10 ranked passages.

**Innovandio S.A.** The participants from Chile submitted five runs to the Passage Retrieval starting from Claims task. The general approach used a two step model in which relevant documents were first retrieved, which were further processed to extract relevant passages. The best placed run was obtained using a Vector Space Model with word 1-grams, and `tf-idf` weighting scheme for the word/dimensions (runID: `In.cos`). Using cosine similarity computations, the first 100 patent documents were retrieved,

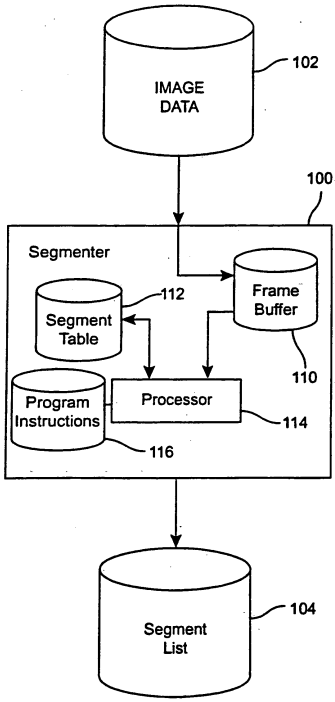


FIG. 2

```

MT Title "FIG. 2"
MT NO 16
MT DE 5
MT UE 7
CO -----
CO ----- Nodes in the chart ----
NO 1 cylinder "IMAGE DATA" (100,100)
NO 2 cylinder "Frame Buffer" (150, 180)
NO 3 rectangle "Processor" (100, 250)
NO 4 cylinder "Segment Table" (70,200)
NO 5 cylinder "Program Instructions" (30,200)
NO 6 point "" (100,120)
NO 7 no-box "110" (180, 240)
...
CO ---- List of edges ----
DE 1 6 plain ""
DE 6 2 plain ""
DE 2 3 plain ""
UE 2 7 wiggly ""
...
CO --- Meta nodes ----
MN 16 [2,3,4,5] "Segmenter" (180,180)
CO --- Done ----
    
```

Fig. 3. A flow chart and an excerpt of its textual encoding

then another cosine similarity was computed at the passage level, between the passages of these 100 documents and the topic's passage vector representations. A similar retrieval approach was done using character 3-gram computations (*In.c3g*).

To tackle the multilingual aspect of the topics and collection, a method that tested the CL-ESA Wikipedia-based multilingual retrieval model was applied [26,4]. 10,000 Wikipedia articles with the most amount of available translations were used to create CL-ESA vector representations, which, together with the *tf-idf* weighting scheme, were used in similarity computations (run *In.clesa*).

In another approach, using the open-source Apache/Solr framework, the entire collection corpus was indexed and the topic content was used to generate a sequence of queries per topic which were sent to the framework. The top 100 documents retrieved were indexed at the passage level (including their XPath) and using the queries formed out of the most frequent words (10 per query) the Solr was tapped to retrieve the most relevant passages (*In.solr*).

In the last of the submitted runs, a combination of the Solr index and word 1-gram solution was aimed for. We suspect (as do the participants) that due to a mistake the wrong data was processed, since all computed scores were 0. This run is not shown in the figures displaying the evaluation scores below.

In all retrieval approaches stopwords and diacritics were removed and a stemmer was applied.

**Vienna University of Technology - University of Macedonia, Thessaloniki.** The TM team participated in the Passage Retrieval task and used a distributed IR system that queried a split CLEF-IP collection. The split is done by exploiting the hierarchical structure of the IPC system. By dividing the collection into several sub-collections (by IPC class `TM.split3`, subclass `TM.split4`, and subgroup `TM.split5`) the patents are organized according to their technological topic. Because patents may be assigned several IPC codes, these splits are not disjoint.

The documents in the CLEF-IP collection were preprocessed to remove the stop-words, and to apply the Porter stemmer. Different documents referring to one patent were merged to form a single (virtual) document to represent the patent. Then the Lemur indexer was used to index the title, abstract, description (first 500 words), claims, inventor, applicant and IPC class information [7].

The CORI and a multilayer method were used for selecting the sources (sub-collections) on which the retrieval should be performed as well as for joining the results.

We note that the TM team did only document level retrieval, therefore the passage specific metric scores in the next section are zero.

## 4 Evaluation Results

We present in this section the measures and the numeric values of these measures we obtain when evaluating the participant's submissions against the task's relevance judgements.

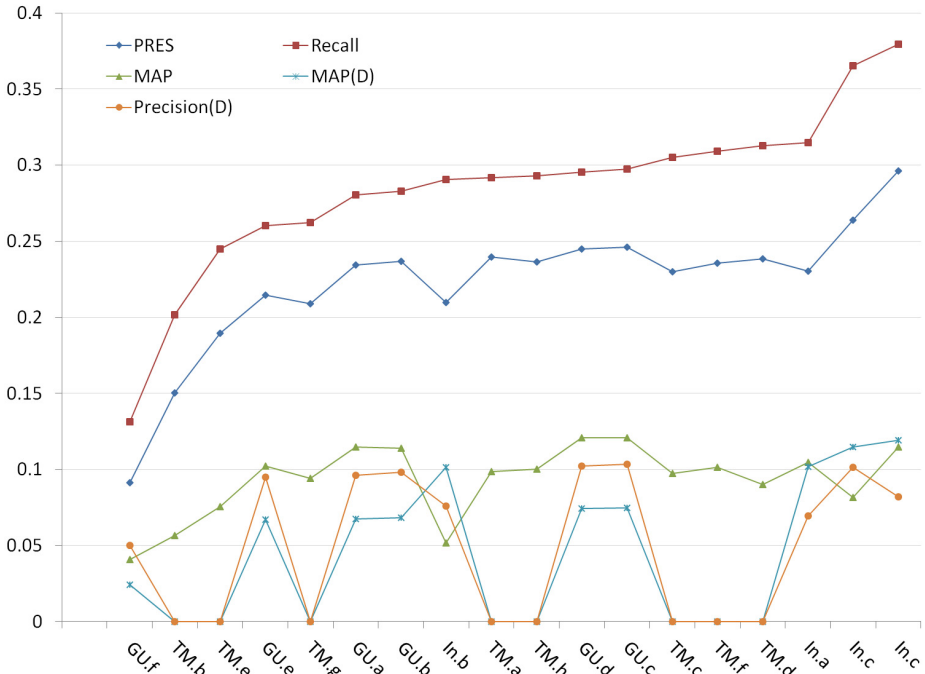


Fig. 4. Evaluation results, ordered by Recall

**Passage Retrieval Starting from Claims.** Considering the submission requirements, where both the patent document relevant to a topic as well as the most important passages in the document are given in the retrieval experiments, we proceeded to do evaluations at two levels: the document level and the passage level.

The evaluation at the document level ignored the passage information in the submitted runs. The metrics computed were PRES (Patent Retrieval Evaluation Score [20]), Recall, and MAP (Mean Average Precision).

At the passage level we compute, for each relevant document retrieved, the Precision and Average Precision scores of the retrieved passages. We then average over the number of relevant document retrieved to get the passage retrieval scores per topic. Averaging over all topics we obtain then the Precision(D) and Mean Average Precision MAP(D) for the retrieval experiment. For more details on these computations see [24]. The idea behind the solutions chosen to compute Precision(D) and MAP(D) are based on the measures used in the ‘Relevant in Context’ task of the INEX evaluation track [12].

Before running any evaluation scripts, we did a clean-up of the submissions by checking that the data follows the required format, that no duplicates occur, and that the retrieval results for one topic were not scattered in the submission file (this caused the evaluation script to exit with an error code). We also removed all XPath referring to headings since we deliberately left them out of the relevance judgements as well. On the qrels side we found that out of 149 two topics were erroneous (topic 78 and topic 101) so we removed them from the evaluation data.

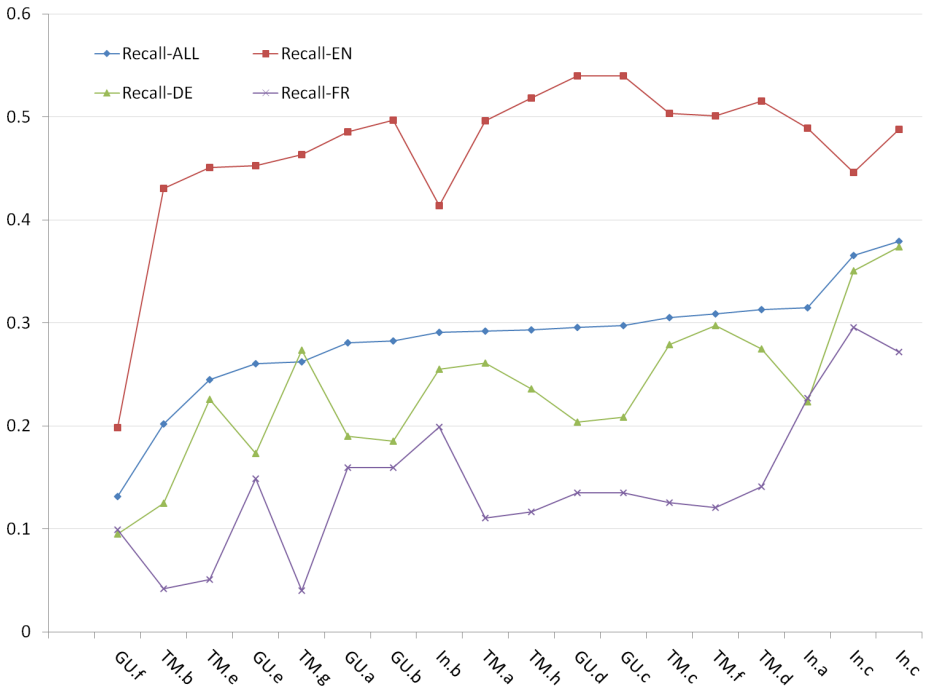


Fig. 5. Evaluation results, document level Recall per language

In the figures below we will use a shortened name for the experiment files. The mapping between the short and the original file name is shown in the appendix.

In our evaluations we considered all documents equally relevant and did evaluations on four sets of topics: the set of all 147 topics, the subset of 50 English topics (1-50), the subset of 49 German topics (51-100, with topic 78 removed), and the subset of 48 French topics (102-149, topic 101 previously removed). The results of the evaluation for the whole topic set are shown in Figure 4, and the document level Recall scores per languages are plotted in Figure 5. One participant submitted retrieval results to the document level, only, which is the reason for the zero Precision(D) and MAP(D) scores.

Further evaluations were done depending on the relevance degree of the patent citation documents, evaluations presented in [25]. A thorough statistical analysis of the retrieval result scores is yet to be done in the near future and will be reported on.

**Structure Recognition Evaluations.** To our dismay, there were no submissions to this task. Nevertheless, in the eventuality that image information extraction experiments were submitted, we were prepared to do evaluation using a set of measures to assess the effectiveness of flowchart recognition. The first set of measures are based on a graph distance metric using the notion of ‘most common subgraph’ (see [6,28] for a definition of the metric and [24] for how it was used in the evaluation last year). Using the experimental data participants submitted in 2012 we also investigated a functional view of the flow-chart recognition results (see [17]).

## 5 Final Words

The CLEF-IP Lab and its tasks have evolved considerably over the last five years, from a rough approximation of a prior art search task in 2009, to, in 2013, a good simulation of the passage-level search carried out by patent searchers. Along the way we have also investigated other important aspects of patent search such as patent classification and patent image search.

The increase in the realism of the tasks over the five years has also raised the bar for participation. In 2009, the CLEF-IP task was similar to a standard ad-hoc retrieval task, and participants could straightforwardly apply general IR solutions and achieve good results. As the tasks have been more closely modelled on actual patent search workflows, participants have been required to invest increasing time in understanding how the patent system works and in developing more granular retrieval solutions.

These factors have likely led to the decline in CLEF-IP submissions in recent years. In 2013, although the number of initial registrations to the tasks was promising, the small number of result submissions is visible in this paper. These factors have made us decide not to pursue the organisation of another round of CLEF-IP evaluations.

The comprehensive, curated test collection containing patent data, with tasks closely related to various activities of a patent expert’s daily workflow, created during the five years of running the CLEF-IP Lab, will remain available to the research community. This should give researchers more than the few months available in the CLEF cycle to develop solutions meeting the demanding requirements of professional patent searchers. A conclusion of CLEF-IP is that patent IR is certainly not a solved problem — many challenges [16] in applying IR solutions in the intellectual property domain remain to be overcome.

**Acknowledgements** This work was partly supported by the EU Network of Excellence PROMISE(FP7-258191) and the Austrian Research Promotion Agency (FFG) FIT-IT project IMPEX(No. 825846).

## References

1. \*\*\*. Citations, <http://www.intellogist.com/wiki/Citations> (last retrieved: July 2013)
2. \*\*\*. International Patent Classification (IPC), <http://www.wipo.int/classifications/ipc/en/> (last retrieved: March 2013)
3. Manual of Patent Examining Procedure (MPEP), revision 2012 (2012) (last retrieved: June 2013)
4. Anderka, M., Stein, B.: The ESA Retrieval Model Revisited. In: Allan, J., Aslam, J.A., Sanderson, M., Zhai, C.X., Zobel, J. (eds.) Proceedings of SIGIR, pp. 670–671. ACM (2009)
5. Attar, R., Fraenkel, A.S.: Local Feedback in Full-text Retrieval Systems. *J. ACM* 24(3), 397–417 (1977)
6. Bunke, H., Shearer, K.: A Graph Distance Metric Based on the Maximal Common Subgraph. *Pattern Recognition Letters* 19(3-4), 255–259 (1998)
7. Giachanou, A., Salampasis, M., Satratzemi, M., Samaras, N.: Report on the CLEF-IP 2013 Experiments: Multilayer Collection Selection on Topically Organized Patents. In: CLEF (Notebook Papers/LABs/Workshops) (2013)
8. Goto, I., Lu, B., Chow, K.P., Sumita, E., Tsou, B.K.: Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In: Proceedings of NTCIR, vol. 9, pp. 559–578 (2011)
9. Graf, E., Azzopardi, L.: A Methodology for Building a Patent Test Collection for Prior art Search. In: Proceedings of the Second International Workshop on Evaluating Information Access (EVIA) (2008)
10. Hanbury, A., Zenz, V., Berger, H.: 1st international workshop on advances in patent information retrieval (AsPIRe 2010). *SIGIR Forum* 44(1), 19–22 (2010)
11. Iwayama, M., Fujii, A., Kando, N., Marukawa, Y.: An Empirical Study on Retrieval Models for Different Document Genres: Patents and Newspaper Articles. In: Proc. 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2003, pp. 251–258. ACM (2003)
12. Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., Robertson, S.: INEX 2007 Evaluation Measures. In: Fuhr, N., Kamps, J., Lalmas, M., Trotman, A. (eds.) INEX 2007. LNCS, vol. 4862, pp. 24–33. Springer, Heidelberg (2008)
13. Kando, N., Leong, M.-K.: Workshop on Patent Retrieval (SIGIR 2000 Workshop Report). *SIGIR Forum* 34(1), 28–30 (2000)
14. Luo, J., Yang, H.: Query Formulation for Prior Art Search - Georgetown University at CLEF-IP 2013. In: CLEF (Notebook Papers/LABs/Workshops) (2013)
15. Lupu, M., Huang, J., Zhu, J., Tait, J.: TREC-CHEM: Large Scale Chemical Information Retrieval Evaluation at TREC. *SIGIR Forum* 43(2) (December 2009)
16. Lupu, M., Hanbury, A.: Patent Retrieval. FnTIR. NOW Publishers (2012)
17. Lupu, M., Piroi, F., Hanbury, A.: Evaluating Flowchart Recognition for Patent Retrieval. In: Song, R., Webber, W., Kando, N., Kishida, K. (eds.) The Fifth International Workshop on Evaluating Information Access (EVIA), pp. 37–44 (2013)

18. Lupu, M., Piroi, F., Huang, X., Zhu, J., Tait, J.: Overview of the TREC 2009 Chemical IR Track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, 2009, Gaithersburg, Maryland, USA, November 17-20, Special Publication 500–278. National Institute of Standards and Technology, NIST (2009)
19. Lv, Y., Zhai, C.: When Documents are Very Long, BM25 Fails! In: Ma, W.-Y., Nie, J.-Y., Baeza-Yates, R.A., Chua, T.-S., Croft, W.B. (eds.) Proceedings of SIGIR, pp. 1103–1104. ACM (2011)
20. Magdy, W., Jones, G.J.F.: PRES: A Score Metric for Evaluating Recall-oriented Information Retrieval Applications. In: SIGIR 2010 (2010)
21. Osborn, M., Strzalkowski, T., Marinescu, M.: Evaluating Document Retrieval in Patent Database: A Preliminary Report. In: Proceedings of the Sixth International Conference on Information and Knowledge Management, CIKM 1997, pp. 216–221. ACM, New York (1997)
22. PCT. Patent Cooperation Treaty (1970), <http://www.wipo.int/pct/en/treaty/about.html> (last retrieved: March 2013)
23. Piroi, F., Lupu, M., Hanbury, A., Zenz, V.: CLEF-IP 2011: Retrieval in the Intellectual Property Domain (September 2011)
24. Piroi, F., Lupu, M., Hanbury, A., Sexton, A.P., Magdy, W., Filippov, I.V.: Clef-IP 2012: Retrieval Experiments in the Intellectual Property Domain. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
25. Piroi, F., Lupu, M., Hanbury, A.: Passage Retrieval Starting from Patent Claims. A CLEF-IP2013 Task Overview. In: CLEF (Online Working Notes/Labs/Workshop) (2013)
26. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-Based Multilingual Retrieval Model. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 522–530. Springer, Heidelberg (2008)
27. Tait, J., Harris, C., Lupu, M.: The 3rd International Workshop on Patent Information Retrieval (PaIR 2010) (October 2010)
28. Wallis, W.D., Shoubridge, P., Kraetzl, M., Ray, D.: Graph Distances Using Graph Union. Pattern Recognition Letters 22(6/7), 701–704 (2001)
29. WIPO Economics & Statistics Series. 2013: PCT Yearly Review. WIPO Publication No. 901E/2013 (2013), [http://www.wipo.int/export/sites/www/freepublications/en/patents/901/wipo\\_pub\\_901\\_2013.pdf](http://www.wipo.int/export/sites/www/freepublications/en/patents/901/wipo_pub_901_2013.pdf)



## Appendix

**Table 2.** Original and short experiment file names

Original file ID	Short ID	Comment
GU.HypCoTtlNoIdfUpperBoundLM	GU.a	
GU.HypCoTtlWithIdfUpperBoundLM	GU.b	
GU.HypDuTtlNoIdfUpperBoundBM	GU.c	
GU.HypDuTtlWithIdfUpperBoundBM	GU.d	
GU.OnlyClaimLM	GU.e	
GU.coOnlyTtlLM	GU.f	
In.c3g	In.a	
In.clesa	In.b	
In.cos	In.c	
In.solr-cos	In.d	Probably an error in generating this experiment file.
In.solr	In.e	
TM.10-100.CORI.CORI.split3	TM.a	no relevant passages returned
TM.10-100.CORI.SSL.split4	TM.b	-"-
TM.10-100.CORI.SSL.split5	TM.c	-"-
TM.10-100.Multilayer.CORI.split4	TM.d	-"-
TM.10-100.Multilayer.CORI.split5	TM.e	-"-
TM.20-50.CORI.CORI.split5	TM.f	-"-
TM.20-50.Multilayer.CORI.split5	TM.g	-"-
TM.clefp-2013-centralised	TM.h	-"-

# ImageCLEF 2013: The Vision, the Data and the Open Challenges

Barbara Caputo<sup>1</sup>, Henning Muller<sup>2</sup>, Bart Thomee<sup>3</sup>, Mauricio Villegas<sup>4</sup>,  
Roberto Paredes<sup>4</sup>, David Zellhofer<sup>5</sup>, Herve Goeau<sup>6</sup>, Alexis Joly<sup>7</sup>,  
Pierre Bonnet<sup>8</sup>, Jesus Martinez Gomez<sup>9</sup>,  
Ismael Garcia Varea<sup>9</sup>, and Miguel Cazorla<sup>10</sup>

<sup>1</sup> Idiap Research Institute, Martigny, Switzerland

<sup>2</sup> University of Applied Sciences Western Switzerland in Sierre, Switzerland

<sup>3</sup> Yahoo! Research, Barcelona, Spain

<sup>4</sup> ITI/DSIC, Universitat Politècnica de València, Spain

<sup>5</sup> Brandenburg University of Technology, Germany

<sup>6</sup> INRIA-IMEDIA, Paris, France

<sup>7</sup> INRIA-ZENITH, Montpellier, France

<sup>8</sup> CIRAD, UMR AMAP, Montpellier, France

<sup>9</sup> University of Castilla-La Mancha, Albacete, Spain

<sup>10</sup> University of Alicante, Alicante, Spain

**Abstract.** This paper presents an overview of the ImageCLEF 2013 lab. Since its first edition in 2003, ImageCLEF has become one of the key initiatives promoting the benchmark evaluation of algorithms for the cross-language annotation and retrieval of images in various domains, such as public and personal images, to data acquired by mobile robot platforms and botanic collections. Over the years, by providing new data collections and challenging tasks to the community of interest, the ImageCLEF lab has achieved a unique position in the multi lingual image annotation and retrieval research landscape. The 2013 edition consisted of three tasks: the photo annotation and retrieval task, the plant identification task and the robot vision task. Furthermore, the medical annotation task, that traditionally has been under the ImageCLEF umbrella and that this year celebrates its tenth anniversary, has been organized in conjunction with AMIA for the first time. The paper describes the tasks and the 2013 competition, giving an unifying perspective of the present activities of the lab while discussing the future challenges and opportunities.

## 1 Introduction

Since its first edition in 2003, the ImageCLEF lab initiative has focused on providing an evaluation forum for the cross-language annotation and retrieval of images [1]. The main motivation behind ImageCLEF is the need to support multilingual users from a global community accessing the ever growing body of visual information. Thus, the main goal of ImageCLEF is to support the advancement of the field of visual media analysis, indexing, classification, and retrieval, by developing the necessary infrastructure for the evaluation of visual

	2009	2010	2011	2012	2013
Total Participations	65	47	43	51	42
Total Registrations	84	112	141	209/106	219/95

**Fig. 1.** Number of registered groups versus number of groups that submitted at least one valid run since 2009. In 2012 and 2013, we report also the total number of groups that initiated the registration process but that, for several reasons, were not able to complete it in time.

information retrieval systems operating in monolingual, language-independent and multi-modal contexts, providing reusable resources for such benchmarking purposes.

To meet these objectives, ImageCLEF organises tasks that benchmark the annotation and retrieval of diverse images such as general photographic and medical images, as well as domain-specific tasks such as plant identification and robot vision. These evaluation tasks aim to support and promote research that addresses key challenges in the field including: 1) visual image annotation with concepts at various levels of abstraction that relies not only on manual, and thus reliable, training data but also on automatically acquired and thus noisy, labelled samples, 2) scientific multimedia data management through the particular case of botanical data identification, and 3) the shift in the area of robot vision from visual place recognition to multimodal place recognition. Moreover, the ImageCLEF 2013 lab has maintained its decade long traditional commitment to medical informatics by helping organizing a challenge on modality classification and retrieval in the medical domain. The aim is to move closer to clinical practice and routine through classification tasks that consider complex, hierarchically organised classes of modalities and retrieval tasks that support medical practitioners in their decision making. This challenge has moved for the first time in 2013 from ImageCLEF in conjunction with the American Medical Informatics Association (AMIA) annual symposium.

Over the years, ImageCLEF has had a significant influence on the visual information retrieval field by benchmarking various retrieval and annotation tasks and by making available the large and realistic test collections built in the context of its activities. Many research groups have participated over the years in its evaluation campaigns and even more have acquired its datasets for experimentation. Figure 1 shows the number of registered groups, and of groups that eventually submitted a run, since 2008. In 2013, over 200 research groups registered, with 42 of those submitting runs officially to the ImageCLEF tasks.

The impact of ImageCLEF can also be seen by its significant scholarly impact indicated by the substantial numbers of its publications and their received citations [2].

The rest of the paper is organized as follows: section 2 describes the three subtasks of the 2013 edition: the photo annotation and retrieval task (section 2.1), the plant identification task (section 2.2), and the robot vision task (section 2.3). Section 2.4 describes the AMIA associated medical tasks. We conclude with an overall discussion, and pointing towards the challenges ahead and possible new directions for ImageCLEF 2014.

## 2 ImageCLEF 2013: The Tasks, the Data and Participation

The 2013 edition of ImageCLEF consisted of three main tasks, plus one task associated with the AMIA 2013 meeting: the photo annotation and retrieval task, the plant identification task, the robot vision task and, jointly with AMIA, the medical task. These tasks had the goal to benchmark the annotation and retrieval of diverse images such as general photographic, as well as domain-specific tasks such as plant identification and robot vision. The overall aim is to support and promote research that addresses key challenges in the field including:

- visual image annotation with concepts at various levels of abstraction that relies not only on manual, and thus reliable, training data, but also on automatically acquired, and thus noisy, labelled samples,
- scientific multimedia data management through the particular case of botanical data identification, and
- the shift in the area of robot vision from visual place recognition to multimodal place recognition.

In the rest of the section, we give an overview account, for each task, of its historical perspective within ImageCLEF, of its 2013 objective and task, and of the task participation and relative results.

### 2.1 The Photo Annotation and Retrieval Task

Automatic concept detection within images is a challenging research problem, as of today yet unsolved. Despite considerable research efforts the so-called semantic gap has not yet been successfully breached, in terms of being able to detect semantic concepts within any kind of imagery for any kind of concept as accurately as real people can. ImageCLEF’s photo annotation and retrieval task aims to advance the state of the art in multimedia research by acting as a platform to foster interaction and collaboration between researchers and by providing a realistic and challenging benchmark for visual concept detection, annotation and retrieval in the context of personal photo and web image collections.

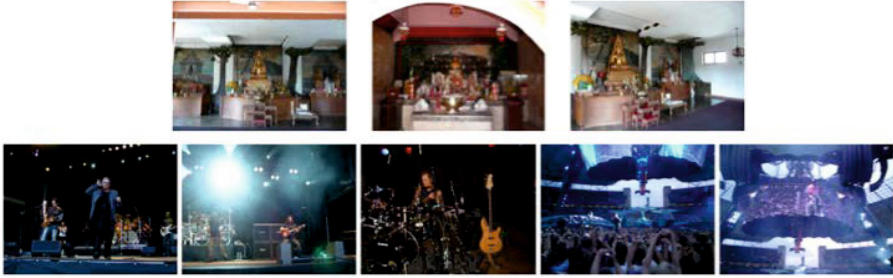


**Fig. 2.** Exemplar images for the photo annotation task. The top row shows images obtained from a web search query of ‘rainbow’; the bottom row shows images from a web search query of ‘sun’.

**Past Editions.** Annotation and retrieval of web images and personal photographs has been part of ImageCLEF since its very first edition in 2003. In the early years the focus was on retrieving relevant images from a web collection given (multilingual) queries, while from 2006 onwards annotation tasks were also held, initially aimed at object detection, but more recently also covering semantic concepts. Between 2009 and 2012 the photo annotation and retrieval tasks were based upon various subsets of the MIRFLICKR collection [3, 4], where every year the list of concepts to detect was updated in order to cover a wider selection of concept types, thus making the task more challenging. With the aim of providing new challenges to the research community, in 2012 two novel subtasks were introduced, one on annotation without requiring any manually labeled training data [5], and the other on retrieval in the context of personal photo collections [6]. These two paths have been continued for this year’s task, and they are described in more details in the following.

**Objective and Task for 2013 Edition.** This year’s task has been divided into two separate subtasks, one entitled *Scalable Concept Image Annotation* and the other *Personal Photo Retrieval*. Each of the subtasks focuses on the two directions of research in this field on which the subtask organizers agreed that deserve more attention.

**Annotation Subtask:** Image concept detection generally has relied on training data that has been manually, and thus reliably annotated, an expensive and laborious endeavor that cannot easily scale, particularly as the number of concepts grows. However, images for any topic can be cheaply gathered from the web, along with associated text from the webpages that contain the images. The degree of relationship between these web images and the surrounding text varies



**Fig. 3.** Exemplar images for the personal photo annotation task. The top row shows samples of the Visual Concept ‘Asian Temple Interior’; the bottom row shows samples of the Event Class ‘Rock Concert’.

greatly, i.e., the data is very noisy, but overall this data contains useful information that can be exploited to develop annotation systems. Likewise there are other resources available that can help to determine the relationships between text and semantic concepts, such as dictionaries or ontologies. The goal of this subtask was to evaluate different strategies to deal with the noisy data so that it can be reliably used for annotating images from practically any topic. Participants were provided with a training set composed of images and corresponding webpage text, and for the given development/test set they had to detect the corresponding concepts for each image using only the input image, the provided training set and any other automatically obtained resources.

*Data.* The data used in this subtask is mostly the same as the one from last year’s task [5], although there are differences [7]. The training set is composed of visual and textual features for 250,000 images downloaded from the web by querying popular search engines. The development and test sets have 1,000 and 2,000 images, respectively, which include only visual features and the corresponding hand labeled concepts ground truth. Figure 2 shows some exemplar images that illustrate the type of challenges addressed in the task. For further details, please refer to [7].

**Personal Photo Retrieval Subtask:** This year’s subtask has a focus on different retrieval usage scenarios and user groups. That is, the subtask reveals whether the tested algorithms are stable in terms of retrieval quality for different user groups. In order to associate relevance assessments with different user groups, the assessors had to answer a questionnaire (see [8]). The subtask is ad-hoc, i.e., no additional training data is released. The participants have to rely on multiple QBE documents and/or browsing data and are asked to find the best matching documents illustrating an event or depicting a visual concept. Thus, an additional objective of this task is to find out whether the participating retrieval systems can exploit data from different search strategies, i.e., query-by-example

and browsing data, in order to find both visual concepts and photos depicting events. To solve the task, the participants have access to pre-extracted visual low-level features, metadata, but are also free to use their own techniques.

*Data.* The subtask uses the same document corpus as in 2012 [6], i.e., 5,555 images that have been sampled from 19 personal photo collections of layperson photographers. In contrast to the last year’s pilot phase, the amount of queries has been increased and the queries are no longer subdivided into events and visual concepts. Additionally, the participants have access to a baseline system that can be used for feature extraction. Figure 3 shows some exemplar images that illustrate the type of challenges addressed in the task. More detailed information is available in a separate publication [8].

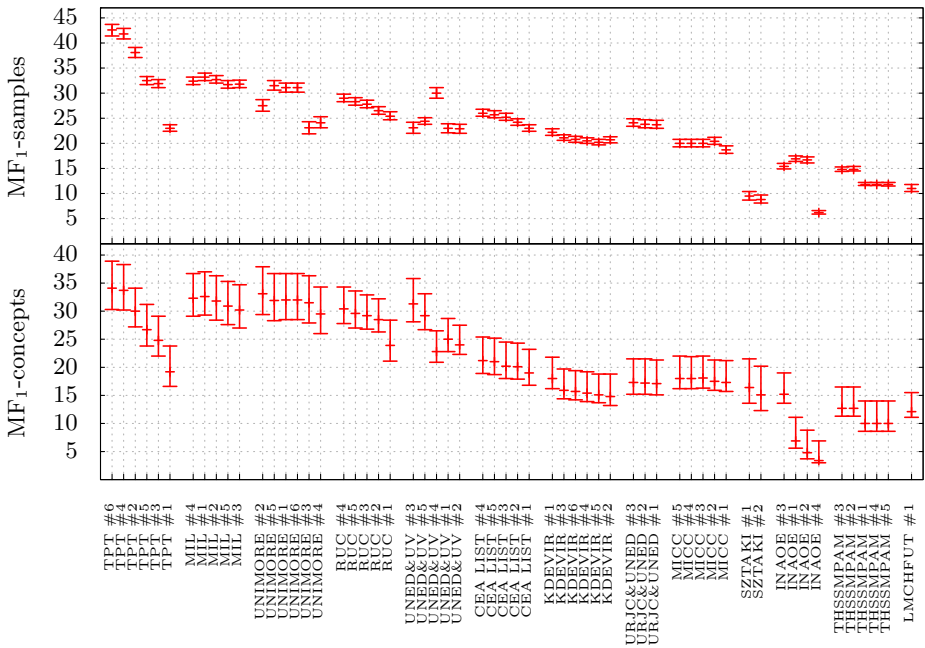
**Participants and Results.** Generally speaking, the participation was excellent. In total, 18 groups took part in the task and submitted 84 runs, of which 26 runs were submitted by 7 groups to the retrieval subtask, whereas the remaining 58 runs were submitted by 13 groups to the annotation subtask. The following is a very brief summary of the results obtained for each subtask. For further details and analysis, the readers should refer to the corresponding overview paper, [7] or [8].

*Annotation Subtask Results:* In comparison to last year (the first edition of this subtask), this year’s results have been much more interesting, even though the challenge has remained mostly the same. The main reason for this is the significantly greater number of participants and submissions. The participating groups have explored several interesting ideas to tackle the proposed problem, which gives hand to a more richer discussion. Figure 4 presents a graph that compares all of the submitted runs using the annotation mean F-measure ( $MF_1$ ), measured both for the test samples and for the concepts. Most of the groups obtained a very impressive improvement in performance compared to the baselines. The most interesting aspect of the results was that even though one system outperformed the rest, many of the ideas proposed by the participants are complementary, so considerable improvements could be expected in future works. For further details, please refer to the subtask overview paper [7].

*Personal Photo Retrieval Subtask Results:* The best performing groups – ISI and DBIS – used visual low-level features and metadata to solve the task. While ISI used relevance feedback for all of their runs, DBIS used this technique only for run #3. In accordance with the findings of the last years’ ImageCLEF tasks, there is evidence that the utilization of multiple modalities increases the retrieval effectiveness. Table 1 shows an excerpt of the average results in order to provide an overview over the general retrieval effectiveness achieved by the participants of the subtask. The user group-specific results are available at the subtask’s website<sup>1</sup>. Regarding the effectiveness variance over the different user groups, the

---

<sup>1</sup> <http://imageclef.org/2013/photo/retrieval#results>



**Fig. 4.** Graphs showing the test set performance measures (in %) for all the submissions for the annotation subtask. The error bars correspond to the 95% confidence intervals computed using Wilson’s method.

results are not very clear. There are only minor differences between the user groups. For a discussion of this effect and a complete overview over the results, please refer to [8].

## 2.2 The Plant Identification Task

If agricultural development is to be successful and biodiversity is to be conserved, then accurate knowledge of the identity, geographic distribution and uses of plants is essential. Unfortunately, such basic information is often only partially available for professional stakeholders, scientists and citizens. So that simply identifying plant species is usually a very difficult task, even for professionals. Using image retrieval technologies is nowadays considered by botanists as a promising direction in reducing this taxonomic gap. ImageCLEF plant identification task, funded by the French project PI@ntNet and the EU coordination action CHORUS+, is aimed at evaluating recent advances of the multimedia IR community on this challenging problem.

**Past Editions.** Each year since 2011, the task is becoming closer to a real-world scenario thanks to the observations feed of a French social network specialized in botany (Tela Botanica). The underlying citizen science project aims at covering

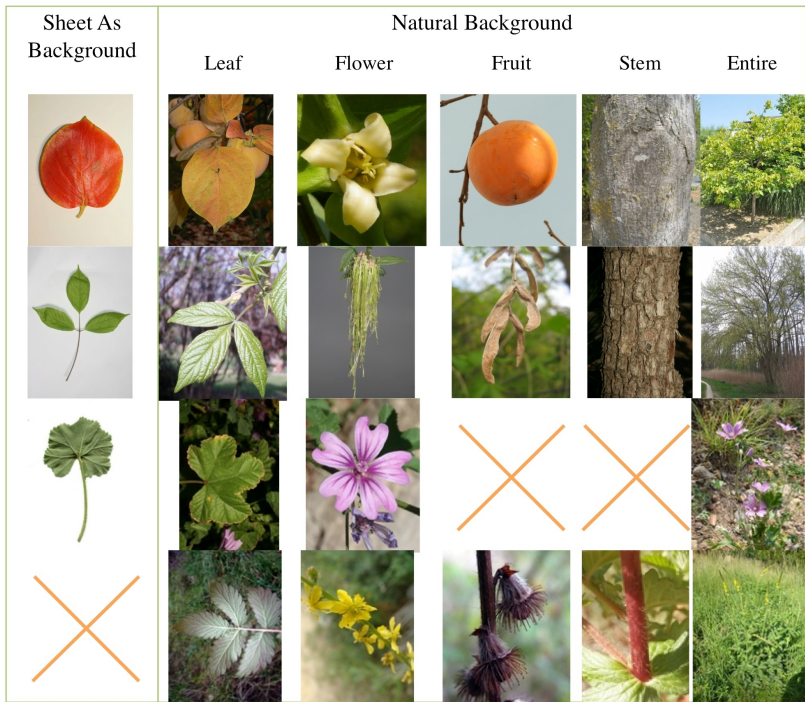


**Table 1.** Summary of the averaged results for the personal photo retrieval subtask (excerpt of the best submissions per group)

Group	Run ID	map_cut_100	ndcg_cut_10	ndcg_cut_20	ndcg_cut_30
DBIS	run3	0.3954	0.7197	0.6798	0.6546
FINKI	run2	0.1375	0.5510	0.4398	0.3881
IPL	IPL13_visual_r4	0.1162	0.5152	0.4173	0.3713
ISI	4	0.5034	0.2167	0.3132	0.3716
ThssMpam4	5000_TL_CR	0.070	0.4005	0.3051	0.2676
ThssMpam4	5000_TL_NCR	0.070	0.4009	0.3050	0.2675
VCTLab	2	0.0783	0.3574	0.3047	0.2754
WIDE_IO	WideIO	0.0584	0.3253	0.2501	0.2192

the entire French flora with a sufficiently rich and balanced collection of pictures. The dataset used for the 2013 campaign covered 250 species of herbs and trees living in France area (i.e. the most represented ones in the whole collected social data since 2011). Contrary to the two previous years that were exclusively focused on leaf images (of tree species only), the coverage of the 2013 task was extended to six different types of view of the plant: leaf scans (or scan-like), leaf photographs, flower photographs, fruit photographs, bark photographs, and the entire view of the plant. A separate evaluation score was computed for the two main categories of images, i.e. scans (or scan-like pictures) vs. photographs (with natural background). Proportions were around 42% of scans and scan-like pictures of leaves vs. 58% of photographs with a natural background (more precisely 16% of leaves, 18% of flowers, 8% of fruits, 8% of stems and 8% of entire). The whole database contained around 26k images collected by 327 distinct contributors, living in different regions in France, equipped with various cameras and at different periods of the year. This makes the task much more realistic than any previous data built for the evaluation of content-based plant identification methods.

**Objective and Task for the 2013 Edition.** The precise goal of the 2013 task was to retrieve the correct species among the top  $k$  species of a ranked list of returned species, one list for each image of a test dataset. Participants received a first training set of annotated images in order to explore different techniques and train their system. Six weeks later participants received the test set containing images without species labels. Then participants were allowed to submit until 4 run files, most of the time related to variations of one same method. A particular attention was paid when splitting the data into training and test subsets to avoid any bias. Several pictures in the dataset might actually depict the same individual plant (or neighboring plants) observed in the same conditions (same person, day, device, lightening conditions, etc.). Randomly splitting images in a naive way would therefore favor having such near-duplicate images in both the training and the test subsets, making the recognition much more easy. To avoid this bias, we therefore performed our random split at the observation level rather than at the image level thanks to associated metadata (observation id



**Fig. 5.** Examples of the different views used in the database: scan or scan-like images of leaves associated to a SheetAsBackground category, and photographs of leaves, flowers, fruits, stems or the entire plants associated to a Natural Background category. Tree species like kaki or maple have generally more pictures and kind of views than herbaceous species like the mallow or the agrimony.

when available, author, date, etc.). The training data finally resulted in 20985 images while the test data resulted in 5092 images. According to similar concerns, the primary metric used to evaluate the submitted runs uses a two-stage average of raw image scores thanks to the users and observations ids associated to each test image. The raw image score itself is the inverse of the rank of the correct species in the list of retrieved species.

**Participation and Results.** With 12 finalist groups over 9 countries and 33 submitted runs, the 2013 edition of the task confirmed its increasing attractiveness (respectively 10 and 11 groups crossed the finish line in 2011 and 2012) although its complexity was higher (with heterogeneous view types). Concerning the scan and scan-like images of leaves (called SheetAsBackground), the results of the 2013 task show that relatively high identification scores can be reach using leaf shape boundary features (between 0.6 and 0.5) but we cant notice a great step of improvement compared to the 2012 campaign. This can be explained by the fact that the queries were more difficult this year with more shadows, weaker

lighting conditions, more old dried leaves and not so uniform background. Concerning the NaturalBackground category, results are as expected lower than the SheetAsBackground category. The highest scores reached equivalent values than the 2012 task, but without any human intervention in the workflow contrary to last year best runs involving some semi-automatic segmentation mechanisms. The detailed results by organ did show that most methods were clearly more accurate on the flower images rather than other organs. It corroborates a well-know usage of botanists for identifying plants and this is good news in a sense that computer vision methods go in the same direction. After flowers, there was no clear second best organ or view type. Bark images provided surprisingly good results relatively to the botanist knowhow on using bark morphology as an identification criterion. Identification results on the entire plant views are also rather surprising regarding their higher complexity and variability. Overall, an important remark is that the ranking of the runs did not change much from an organ to another one, fostering the idea that generic methods might solve heterogeneous fine-grained classification problems. Regarding metadata, one run did show that using the observation date complementary to the visual content was a simple and efficient way to obtain a gain of up to 5 points on the flower category (thanks to the relatively short flourishing season of many species). On the other side, the GPS information was not successfully exploited probably because the database doesn't contain dense enough observations to build an accurate geographic repartition of the species. With the emergence of more and more plant identification apps [9] [10], [11], [12] and the ecological urgency to build real-world and effective identification tools, we believe that the detailed results and conclusions of the task will be of high interest for the community [13].

### 2.3 The Robot Vision Task

The Robot Vision task addresses two main problems related to semantic robot localization: place classification and object recognition. Participants are asked to answer the questions “where are you?” and “which object can you recognize in the scene?” when presented with a test sequence. Such test sequence contains depth and visual images acquired by a mobile robot with a RGB-D camera in a previously seen indoor environment.

**Past Editions.** The Robot Vision task started in 2009 [14], with the main objective to compare different approaches to robot localization in a common scenario. The localization problem has always been managed from a semantic point of view, where no topological information is provided or required. Since its origin, new challenges have been introduced each new edition, from detection of unknown rooms [14], to generalization across floors [15? ], to categorization problems [16] to multimodal data analysis [17]. At this fifth edition, 2013, the proposed challenge is the object recognition problem.

**Objective and Task for the 2013 Edition.** For the 2013 edition, the semantic representation of the space is described by two elements which will determine

the expected behaviour of people or robots in such scene. These two elements are: (1), the semantic category of the room (determines the activities we usually perform there, like Kitchen or Corridor) and (2), the list of objects the room contains (like Frigde or Desk). In a similar way topological localization (in conjunction with navigation and mapping) allows robots to move to a desired position, semantic localization is expected to provide robots with new capabilities. These capabilities are the identification of the most appropriate behaviour and the recognition of the objects that are suitable for interaction.

In this task edition, the relationship between room categories and objects is explicitly given. Using the labelling information, we can compute the conditional probability for a room category, given the list of objects in the scene  $P(C = c_1 | o_1, o_2, o_6)$  or vice versa  $P(o_1 | C = c_1)$ . This can be used to create a high level reasoning layer to be used in conjunction with low level classifiers. For example, the probability of detecting an Urinal in a Secretary is very low. Let us assume that we have classified a test frame as Secretary with high confidence but the object classifier cannot detect the presence or lack for the Urinal. In this case, the prior knowledge could be used to classify Urinal as not present. The use of this knowledge from participants is one of the goals of the challenge.

*Task Description.* Participants are provided with two training sequences imaging all the rooms and object categories. They are expected to generate algorithms capable of providing information from test frames. Concretely, algorithms have to list all the objects that appear in the scene and classify the room category. The number of times a specific object appears in a frame is not relevant and, for each object, we have a binary problem. Room classification is a multi-class problem.


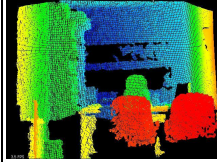

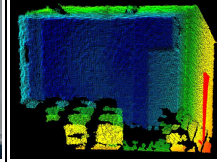
Table 2 shows a global description for the task, where left columns correspond with the training stage of the challenge and the right one with the test. For both sequences, at each frame two different images are presented to the participants: a visual image and a point cloud. In the training stage, all the information for the room and the objects in the scene is provided. This information should be used by the participants teams to generate their algorithms. They have to classify the room category into one of the 10 available classes and say if each of the 8 possible objects are present or not. Due to wrong classifications will obtain negative scores, participants are allowed to not provide information about room category or object presence.

*Performance Evaluation.* The proposals of the participants are compared using a score obtained from their submissions. The final score for a run will be the sum of all the scores obtained for the test frames included in the sequence. The following rules are used when calculating the final score for a frame:

Room Category (single multi-class problem)

- The room category has been correctly classified: +1.0 points
- The room category has been wrongly classified: -0.5 points
- The room category has not been classified: 0.0 points

**Table 2.** Task Description

Training		Test	
Visual and Depth Images		Visual and Depth Images	
			
Labels (provided)		Labels (required)	
Room Category	Objects	Room Category	Objects
Professor Office	Extinguisher: NO Computer: YES Printer: NO Urinal: NO Chair: YES Screen: NO Trash: YES Fridge: NO	Class in Rooms or Unknown	Extinguisher: Y/N/-? Computer: Y/N/-? Printer: Y/N/-? Urinal: Y/N/-? Chair: Y/N/-? Screen: Y/N/-? Trash: Y/N/-? Fridge: Y/N/-?

Object (8 different binary problems)

- For each correctly classified object within the frame: +0.125 points
- For each misclassified object within the frame: -0.125 points
- For each object that was not classified: 0.0 points

*The Data.* The dataset provided for the task consists of different sequences of depth (in Point Cloud Data (PCD) format [18]) and visual images acquired within a department building at the University of Alicante, Spain. Concretely, there are two labelled sequences for training, another labelled sequence provided for validation, and one unlabelled sequence for testing. Every image has been manually labelled with its corresponding room category and with a list of eight different objects to appear or not within it. The 10 different room categories are: corridor, hall, professorOffice, studentOffice, technicalRoom, toilet, secretary, visioconference, elevator area and warehouse. The 8 different objects are: extinguisher, computer, chair, printer, urinal, screen, trash and fridge. The frequency distribution for room categories and objects are depicted in Table 3 and Table 4 respectively.

Corridor is the most common class in all sequences, due to the space distribution of the building used in the acquisition. This turns room classification into an unbalanced problem with higher probabilities for classifying frames as Corridor than for the rest of room categories. The validation sequence was released some months after the training sequences. The main objective of this sequence was to prevent the extreme lighting conditions of the test sequence. Due to it was acquired only in the first floor of the building, it does not contains any frame for three rooms: Warehouse, VisioConference and Hall.

**Table 3.** Frequency distribution of room categories for dataset sequences

Room Category	Number of frames			
	Training 1	Training 2	Validation	Test
Corridor	891	1262	764	1317
Hall	103	228	000	297
ProfessorOffice	124	192	200	222
StudentOffice	155	276	282	318
TechnicalRoom	136	281	214	240
Toilet	121	242	188	198
Secretary	098	195	181	201
VisioConference	149	300	000	306
Warehouse	070	166	000	127
ElevatorArea	100	174	040	289
All	1947	3316	1869	3515

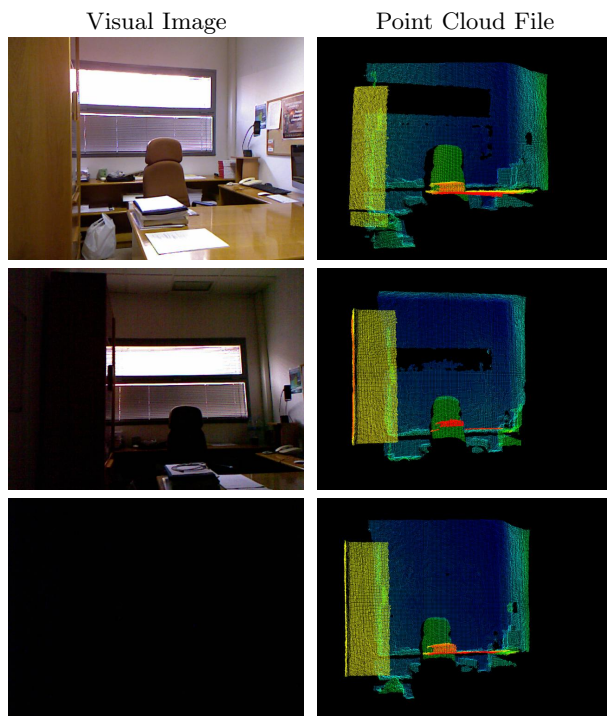
**Table 4.** Frequency distribution of object presences or lacks for dataset sequences

Room Category	Number of presences / lacks			
	Training 1	Training 2	Validation	Test
Extinguisher	259 / 1688	529 / 2787	286 / 1583	520 / 2995
Computer	289 / 1658	466 / 2850	416 / 1453	473 / 3042
Chair	470 / 1477	767 / 2549	567 / 1302	889 / 2626
Printer	210 / 1737	292 / 3024	255 / 1614	279 / 3236
Urinal	054 / 1893	110 / 3206	070 / 1799	090 / 3425
Screen	081 / 1866	190 / 3126	000 / 1869	151 / 3364
Trash	406 / 1541	451 / 2865	253 / 1616	662 / 2853
Fridge	057 / 1890	104 / 3212	099 / 1770	114 / 3401
All	1826 / 13750	2909 / 23610	1946 / 13006	3178 / 24942

Fig. 6 shows the same scene represented in three different sequences: training1 (top), validation (middle) and test (bottom). The scene was acquired using visual images (left) and point cloud data files (right). Training, validation and test sequences were acquired within the same building at two different floors but with some variations in the lighting conditions (as can be observed in Fig. 6) and in the acquisition procedure (clockwise and counter clockwise, ground floor first or ground floor last). Participants were provided with running code for computing several feature descriptors [19–21] as well as SVM-based online [22, 23] and cure integration classifiers [24, 25].

**Participation and Results.** In 2013, 39 participants registered to the Robot Vision task but only 6 submitted, at least, one run accounting for a total of 16 different runs. These participants were:

- NUDT: National University of Defense Technology, Changsha, China.
- MIAR ICT: Beijing, China.



**Fig. 6.** Visual and 3D point cloud files for the same scene under different lighting conditions

- MICA: Hanoi university of Science and Technology, Hanoi, Vietnam
- REGIM: University of Sfax National School of Engineers, Tunisia
- GRAM: University of Alcalá de Henares, Spain
- SIMD: University of Castilla-La Mancha, Albacete, Spain.
  - Out of competition organizers contribution using proposed techniques

The scores obtained by all the submitted runs are shown in Table 5. The maximum score that could be achieved was 7030 and the winner (MIAR ICT) obtained a score of 6033.5 points. NUDT and SIMD teams ranked second and third respectively and their score was higher than 71% of the maximum score (the one obtained with the baseline system, SIMD result in the table).

\* SIMD organizers submission was out-of-competition, it was provided to be considered a baseline score. The organizers only used the techniques proposed in the webpage of the Robot Vision challenge<sup>2</sup>. Concretely, PHOW [19] features were extracted from visual images and then, a Support Vector Machine was trained.

According to the obtained results we can conclude that the introduction of the object recognition task was not as challenging as we expected: most of the

<sup>2</sup> <http://www.imageclef.org/2013/robot>

**Table 5.** Overall ranking of the runs submitted by the participant groups to the 2013 Robot Vision task

Rank	Group Name	Score	% Max. Score
1	MIAR ICT	6033.500	85.83
2	MIAR ICT	5924.250	84.27
3	MIAR ICT	5924.250	84.27
4	MIAR ICT	5867.500	83.46
5	MIAR ICT	5867.000	83.46
6	NUDT	5722.500	81.40
7	SIMD*	5004.750	71.19
8	REGIM	4368.250	65.98
9	MICA	4479.875	63.73
10	REGIM	3763.750	53.54
11	MICA	3316.125	47.17
12	MICA	2680.625	38.13
13	GRAM	-487.000	<0.00
14	GRAM	-497.000	<0.00
15	GRAM	-497.000	<0.00
16	NUDT	-866.250	<0.00

participants were able to identify those object properly. With respect to the scores obtained by the different runs, almost half of them improved the baseline results provided by the organizers, obtaining score higher than the 80% of the maximum score.

## 2.4 AMIA: The Medical Task

The main objective of the medical ImageCLEF task is to compare content-based image retrieval (CBIR) systems in medicine, and in particular to determine how associated cross-language text can be used in combination with CBIR to improve retrieval and ranking. ImageCLEFmed evaluates retrieval systems with visual, semantic and mixed topics in several languages using since 2008 a data collection from the biomedical literature.

**Past Editions.** ImageCLEFmed started in 2004 with only an image-based retrieval task [26]. In 2005, an automatic annotation task was introduced [27]. The goal of this task was to find out how well the techniques can identify body orientation, body region, and biological system examined based on the images. The database consisted of 10,000 radiographs fully annotated with IRMA code, taken randomly from medical routine. Between 2006 and 2009, ImageCLEFmed kept these two tasks in similar formats format but using larger and more complex databases each year [28–32]. From 2008 to 2010, the database contained images from articles published in Radiology and Radiographics including the text of the captions and a link to the html of the full text articles. In 2009, a lung nodule detection task was tested. The goal of this task was to compare the performance of lung nodule detection techniques with a gold standard of manually identified



nodules. The data for this task was a subset of the LIDC (Lung Image Database Consortium) database. From 2010 to 2012, there were three types of task: the traditional image-based retrieval, modality classification and case-based retrieval [33–35]. The modality classification task was introduced since previous studies have shown that imaging modality is an important aspect of the image for medical retrieval. Using the modality classification the search results can be improved significantly. In 2010, the images had to be classified into one of 8 modalities (CT, MR, XR, etc.); in 2011 into 18 and in 2012–2013 into 31. In the case-based retrieval task, a case description, with patient anamnesis, limited symptoms and test results including imaging studies is provided (but not the final diagnosis). The goal is to retrieve cases including images that are useful for a differential diagnosis or even match the exact diagnosis of the query.

**Objective and Task for 2013 Edition.** In 2013, the 10th year of the medical task is celebrated [36]. The ImageCLEFmed meeting will be organized at the annual AMIA meeting in the form of a workshop. This means that the workshop will be organized outside of Europe for the first time. ImageCLEFmed is running in a similar format as in 2012 but with a new task, the compound figure separation that became important as a large fraction of around 40% of the database of PubMed Central used contain compound figures and the sub images are otherwise not accessible for research. Another novelty in 2013 is that the modality classification task includes a large amount of compound images to make the task more difficult and realistic. The following tasks were offered in 2013:

- Modality Classification: In user-studies, clinicians have indicated that modality is one of the most important filters that they would like to be able to limit their search by. Many image retrieval websites (Goldminer, Yottalook) allow users to limit the search results to a particular modality. However, this modality is typically extracted from the caption and is often not correct or present. Studies have shown that the modality can be extracted from the image itself using visual features. Additionally, using the modality classification, the search results can be improved significantly. In 2013, a larger number of compound figures will be present making the task significantly harder but corresponding much more to the reality of biomedical journals.
- Compound figure separation: As up to 40% of the figures in PubMed Central are compound figures, a major step in making the content of the compound figures accessible is the detection of compound figures and then their separation into sub figures that can subsequently be classified into modalities and made available for research. The task makes available training data with separation labels of the figures, and then a test data set where the labels were made available after the submission of the results.
- Ad-hoc image-based retrieval: This is the classic medical retrieval task, similar to those in organized since 2004. Participants were given a set of 30 textual queries with 2–3 sample images for each query. The queries were classified into textual, mixed and semantic, based on the methods that are expected to yield the best results.

- Case-based retrieval: This task was first introduced in 2009. Unlike the ad-hoc task, the unit of retrieval here is a case, not an image. For the purposes of this task, a "case" is a PubMed ID corresponding to the journal article. In the results submissions the article DOI should be used as several articles do neither have PubMed IDs nor Article URLs.

The medical image classification and retrieval tasks in 2013 cover image modality classification, compound image separation and image retrieval with visual, semantic and mixed topics in several languages using a data collection from the biomedical literature.

**Participation and Results.** In total over 60 groups registered for the medical tasks and obtained access to the data sets. 10 of the registered groups submitted results to the medical tasks with a total of 166 valid runs submitted. 8 groups participated in the modality classification task with 51 runs; 3 groups participated in the compound figure separation task with 4 runs; 9 groups participated in the image retrieval task with 66 runs and 7 groups participated in the case-based retrieval task with 45 runs. As in previous years, the largest number of runs was submitted for the image-based retrieval task although the number submitted runs at the modality classification task increased to 51 (43 in 2012 and 34 in 2011). There are still different situations as to whether visual, textual or combined techniques perform better depending on the task. For further information you can see the ImageCLEFmed overview [36].

### 3 Conclusion

This paper presented an overview of the activities in the 2013 edition of the ImageCLEF lab. The sustained interest in the lab, witnessed by the growing number of registration and the sustained number of groups actually participating to the lab, make ImageCLEF an important resource in the multi lingual image annotation and retrieval research landscape. The ever growing amount of data available through the internet, and the growing demand of tools for accessing and exploiting them, will become one of the key focus for the 2014 edition of ImageCLEF, where we look forward to welcome back the medical task under the ImageCLEF umbrella.

**Acknowledgments.** This work has been partially supported by the Halser Foundation (B. C.), by the LiMoSINE FP7 project under grant # 288024 (B. T.), by the Khresmoi (grant # 257528) and PROMISE ( grant # 258191) FP 7 projects (H.M.) and by the tranScriptorium FP7 project under grant # 600707 (M. V., R. P.).

### References

1. Muller, H., Clough, P., Deselaers, T., Caputo, B.: ImageCLEF: experimental evaluation in visual information retrieval. Springer (2010)

2. Tsirikla, T., Seco de Herrera, A.G., Müller, H.: Assessing the scholarly impact of imageCLEF. In: Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., de Rijke, M. (eds.) CLEF 2011. LNCS, vol. 6941, pp. 95–106. Springer, Heidelberg (2011)
3. Huiskes, M., Lew, M.: The MIR Flickr retrieval evaluation. In: Proceedings of the 10th ACM Conference on Multimedia Information Retrieval, Vancouver, BC, Canada, pp. 39–43 (2008)
4. Huiskes, M., Thomee, B., Lew, M.: New trends and ideas in visual concept detection. In: Proceedings of the 11th ACM Conference on Multimedia Information Retrieval, Philadelphia, PA, USA, pp. 527–536 (2010)
5. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task. In: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy (2012)
6. Zellhöfer, D.: Overview of the Personal Photo Retrieval Pilot Task at ImageCLEF 2012. In: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy (2012)
7. Villegas, M., Paredes, R., Thomee, B.: Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes, Valencia, Spain (2013)
8. Zellhöfer, D.: Overview of the ImageCLEF 2013 Personal Photo Retrieval Subtask. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes, Valencia, Spain (2013)
9. Leafsnap (2011)
10. Plantnet (2013)
11. Mobile flora (2013)
12. Folia (2012)
13. Goëau, H., Bonnet, P., Joly, A., Bakic, V., Boujemaa, N., Barthelemy, D., Molino, J.F.: The imageclef 2013 plant identification task. In: ImageCLEF 2013 Working Notes (2013)
14. Pronobis, A., Xing, L., Caputo, B.: Overview of the CLEF 2009 robot vision track. In: Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy-Cramer, J., Müller, H., Tsirikla, T. (eds.) CLEF 2009. LNCS, vol. 6242, pp. 110–119. Springer, Heidelberg (2010)
15. Pronobis, A., Caputo, B.: The robot vision task. In: Muller, H., Clough, P., Deselaers, T., Caputo, B. (eds.) ImageCLEF. The Information Retrieval Series, vol. 32, pp. 185–198. Springer, Heidelberg (2010)
16. Pronobis, A., Christensen, H.I., Caputo, B.: Overview of the imageCLEF@ICPR 2010 robot vision track. In: Ünay, D., Çataltepe, Z., Aksoy, S. (eds.) ICPR 2010. LNCS, vol. 6388, pp. 171–179. Springer, Heidelberg (2010)
17. Martinez-Gomez, J., Garcia-Varea, I., Caputo, B.: Overview of the imageclef 2012 robot vision task. In: CLEF 2012 Working Notes (2012)
18. Rusu, R., Cousins, S.: 3d is here: Point cloud library (pcl). In: 2011 IEEE International Conference on Robotics and Automation (ICRA), pp. 1–4. IEEE (2011)
19. Bosch, A., Zisserman, A., Muñoz, X.: Image classification using random forests and ferns. In: International Conference on Computer Vision, pp. 1–8. Citeseer (2007)
20. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
21. Linde, O., Lindeberg, T.: Object recognition using composed receptive field histograms of higher dimensionality. In: Proc. ICPR. Citeseer (2004)
22. Orabona, F., Castellini, C., Caputo, B., Luo, J., Sandini, G.: Indoor place recognition using online independent support vector machines. In: Proc. BMVC, vol. 7 (2007)

23. Orabona, F., Castellini, C., Caputo, B., Jie, L., Sandini, G.: On-line independent support vector machines. *Pattern Recognition* 43, 1402–1412 (2010)
24. Orabona, F., Jie, L., Caputo, B.: Online-Batch Strongly Convex Multi Kernel Learning. In: *Proc. of Computer Vision and Pattern Recognition, CVPR* (2010)
25. Orabona, F., Jie, L., Caputo, B.: Multi kernel learning with online-batch optimization. *Journal of Machine Learning Research* 13, 165–191 (2012)
26. Clough, P., Müller, H., Sanderson, M.: The CLEF 2004 cross-language image retrieval track. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) *CLEF 2004. LNCS*, vol. 3491, pp. 597–613. Springer, Heidelberg (2005)
27. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 cross-language image retrieval track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, pp. 535–557. Springer, Heidelberg (2006)
28. Müller, H., Deselaers, T., Deserno, T., Clough, P., Kim, E., Hersh, W.: Overview of the imageCLEFmed 2006 medical retrieval and medical annotation tasks. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) *CLEF 2006. LNCS*, vol. 4730, pp. 595–608. Springer, Heidelberg (2007)
29. Müller, H., Deselaers, T., Deserno, T., Kalpathy–Cramer, J., Kim, E., Hersh, W.: Overview of the imageCLEFmed 2007 medical retrieval and medical annotation tasks. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) *CLEF 2007. LNCS*, vol. 5152, pp. 472–491. Springer, Heidelberg (2008)
30. Müller, H., Kalpathy–Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Kahn Jr., C.E., Hersh, W.: Overview of the CLEF 2009 medical image retrieval track. In: Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy–Cramer, J., Müller, H., Tsirikla, T. (eds.) *CLEF 2009, Part II. LNCS*, vol. 6242, pp. 72–84. Springer, Heidelberg (2010)
31. Müller, H., Kalpathy–Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Kahn Jr., C.E., Hersh, W.: Overview of the CLEF 2009 medical image retrieval track. In: Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy–Cramer, J., Müller, H., Tsirikla, T. (eds.) *CLEF 2009, Part II. LNCS*, vol. 6242, pp. 72–84. Springer, Heidelberg (2010)
32. Tommasi, T., Caputo, B., Welter, P., Güld, M.O., Deserno, T.M.: Overview of the CLEF 2009 medical image annotation track. In: Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy–Cramer, J., Müller, H., Tsirikla, T. (eds.) *CLEF 2009. LNCS*, vol. 6242, pp. 85–93. Springer, Heidelberg (2010)
33. Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds.): *ImageCLEF – Experimental Evaluation in Visual Information Retrieval. The Springer International Series on Information Retrieval*, vol. 32. Springer, Heidelberg (2010)
34. Kalpathy–Cramer, J., Müller, H., Bedrick, S., Eggel, I., García Seco de Herrera, A., Tsirikla, T.: The CLEF 2011 medical image retrieval and classification tasks. In: *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)* (2011)
35. Müller, H., García Seco de Herrera, A., Kalpathy–Cramer, J., Demner Fushman, D., Antani, S., Eggel, I.: Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: *Working Notes of CLEF 2012 (Cross Language Evaluation Forum)* (2012)
36. García Seco de Herrera, A., Kalpathy–Cramer, J., Demner Fushman, D., Antani, S., Müller, H.: Overview of the ImageCLEF 2013 medical tasks. In: *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)* (2013)

# Overview of INEX 2013

Patrice Bellot, Antoine Doucet, Shlomo Geva<sup>1</sup>, Sairam Gurajada, Jaap Kamps<sup>2</sup>, Gabriella Kazai, Marijn Koolen, Arunav Mishra, Véronique Moriceau, Josiane Mothe, Michael Preminger, Eric SanJuan, Ralf Schenkel<sup>3</sup>, Xavier Tannier, Martin Theobald, Matthew Trappett, and Qiuyue Wang

<sup>1</sup> INEX co-chair and QUT, Australia

<sup>2</sup> INEX co-chair and University of Amsterdam, The Netherlands

<sup>3</sup> INEX co-chair and University of Passau, Germany

**Abstract.** INEX investigates focused retrieval from structured documents by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results. This paper reports on the INEX 2013 evaluation campaign, which consisted of four activities addressing three themes: *searching professional and user generated data* (Social Book Search track); *searching structured or semantic data* (Linked Data track); and *focused retrieval* (Snippet Retrieval and Tweet Contextualization tracks). INEX 2013 was an exciting year for INEX in which we consolidated the collaboration with (other activities in) CLEF and for the second time ran our workshop as part of the CLEF labs in order to facilitate knowledge transfer between the evaluation forums. This paper gives an overview of all the INEX 2013 tracks, their aims and task, the built test-collections, and gives an initial analysis of the results.

## 1 Introduction

Traditional IR focuses on pure text retrieval over “bags of words” but the use of structure—such as document structure, semantic metadata, entities, or genre/topical structure—is of increasing importance on the Web and in professional search. INEX has been pioneering the use of structure for focused retrieval since 2002, by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results.

INEX 2013 was an exciting year for INEX in which we joined forces with CLEF and ran our workshop as part of the CLEF labs in order to foster further collaboration and facilitate knowledge transfer between the evaluation forums. In total four research tracks were included, which studied different aspects of focused information access:

**Social Book Search Track** investigating techniques to support users in searching and navigating collections of digitised or digital books, metadata and complementary social media. The *Social Book Search Task* studies the relative value of authoritative metadata and user-generated content using a

test collection with data from Amazon and LibraryThing. The *Prove It Task* asks for pages confirming or refuting a factual statement, using a corpus of the full texts of 50k digitized books.

**Linked Data Track** investigating retrieval over a strongly structured collection of documents based on DBpedia and Wikipedia. The *Ad Hoc Search Task* has informational requests to be answered by the entities in DBpedia/Wikipedia. The *Jeopardy Task* asks for the (manual) formulation of effective SPARQL queries with additional keyword filters, aiming to express natural language search cues more effectively.

**Tweet Contextualization Track** investigating tweet contextualization, helping a user to understand a tweet by providing him with a short background summary generated from relevant Wikipedia passages aggregated into a coherent summary.

**Snippet Retrieval Track** investigate how to generate informative snippets for search results. Such snippets should provide sufficient information to allow the user to determine the relevance of each document, without needing to view the document itself.

Both Tweet Contextualization and Snippet retrieval use the same XML'ified corpus of Wikipedia, and address focused retrieval in the form of constructing some concise selection of information in a form that is of interest to NLP researchers (tweet contextualization) and to IR researchers (snippet retrieval).

In the rest of this paper, we discuss the aims and results of the INEX 2013 tracks in relatively self-contained sections: the Social Books Search track (Section 2), the Linked Data track (Section 3), and the paired Tweet Contextualization (Section 4) and Snippet Retrieval (Section 5) tracks.

## 2 Social Book Search Track

In this section, we will briefly discuss the INEX 2013 Social Book Search Track (addressing the searching professional and user generated data theme). Further details are in [7].

### 2.1 Aims and Tasks

Prompted by the availability of large collections of digitized books, the Social Book Search Track aims to promote research into techniques for supporting users in searching, navigating and reading full texts of digitized books and associated metadata. This year, the track ran two tasks: the Social Book Search task and the Prove It task:

1. The *Social Book Search* (SBS) task, framed within the scenario of a user searching a large online book catalogue for a given topic of interest, aims at exploring techniques to deal with both complex information needs of searchers—which go beyond topical relevance and can include aspects such as genre, recency, engagement, interestingness, quality and how well-written

a book is—and heterogeneous information sources including user profiles, personal catalogues, professional metadata and user-generated content.

2. The *Prove It* (PI) task aims to test focused retrieval approaches on collections of books, where users expect to be pointed directly at relevant book parts that may help to confirm or refute a factual claim;

In addition to these task, the *Structure Extraction* (SE) task runs at ICDAR in 2013 [3] and aims at evaluating automatic techniques for deriving structure from OCR and building hyperlinked table of contents. The extracted structure could then be used to aid navigation inside the books.

## 2.2 Test Collections

For the Social Book Search task a new type of test collection has been developed. Unlike traditional collections of topics and topical relevance judgements, the task is based on rich, real-world information needs from the LibraryThing (LT) discussion forums and user profiles. The collection consists of 2.8 million book descriptions from Amazon, including user reviews, and is enriched with user-generated content from LT. For the information needs we used the LT discussion forums. We selected 386 discussion threads which focus on members asking for book recommendations on a certain topic. The initial messages in these threads often contain detailed descriptions of what they are looking for. The relevance judgements come in the form of suggestions from other LT members in the same discussion thread. We paid trained annotators to indicate for each book suggestion in the thread whether the person suggesting the book has read it and whether they are positive, neutral or negative to it. These opinions are used to derive relevance values for the books. Opinions from the topic creator are the most important, then those of others who have read the book and finally those of members who have not. The final set of judgements contain suggestions for 380 topics with an average of 16 judgements per topic. The judgements are independent of the submitted runs, which avoids pooling bias. Previously we investigated the reliability of using forum suggestions for evaluation and found they are complete enough, but different in nature from editorial judgements based on topical relevance [6].

The PI task builds on a collection of over 50,000 digitised out-of-copyright books (about 17 million pages) of different genre (e.g., history books, text books, reference works, novels and poetry) marked up in XML. The task was first run in 2010 and was kept the same for 2011 and 2012. This year the aim is to evaluate book-pages not only on whether they contain information confirming or refuting a statement, but also whether the book is authoritative and of an appropriate genre and subject matter such that a reader would trust the confirming or refuting information.

The SE task relies on a subset of the 50,000 digitized books of the PI task. In 2013, the participants were to extract the tables of contents of 1,000 books extracted from the whole PI book collection. In previous years, the ground truth was constructed collaboratively by participating institutions. For the first time

in 2013, the ground truth production was performed by an external provider, and partly funded by the Seventh Framework Program (FP7) of the EU Commission. This centralized construction granted better consistency. In addition, it also validated the collaborative process used since 2009, as the results this year were in line with those of the previous rounds.

### 2.3 Results

Eight teams together submitted 33 runs to the SBS task and two teams submitted 12 runs to the Prove It! task. The *Social Book Search* task evaluation has shown that the most effective systems use all available book information—professional metadata and user-generated content—and incorporate either the full topic statement, which includes the title of the topic thread, the name of the discussion group, the full first message that elaborates on the request and the query generate by annotators, or a combination of the title and the query. None of the groups used user profile information for the runs they submitted. The best performing run is *run3.all-plus-query.all-doc-fields* by **RSLIS**, which used all topic fields combined against an index containing all available document fields. The second best group is **UAmS (ILLC)** with run *inex13SBS.tiqu.bayes.avg.LT.rating*, which uses only the topic titles and moderated query ran against an index containing the title information fields (title, author, edition, publisher, year), user-generated content fields (tags, reviews and awards) and the subject headings and Dewey decimal classification titles from the British Library and Library of Congress. The retrieval score of each book was then multiplied by a prior probability based on the Bayesian average of LT ratings for that book. The third group is **ISMD**, with manual run *run\_ss\_bsqstw\_stop\_words\_free...*. This run is generated after removing Book Search Query Stop Words (bsqstw), standard stopwords and the member field from the topics and running against an index where stopwords are removed and the remaining terms are stemmed with the Krovetz stemmer. If we ignore the manual runs, ismd is still the third group with the fully automatic run ism run ss free text 2013, which is generated using free text queries on Krovetz stemmed and stopwords removed index.

For the *Prove It* task, we expect to have relevance judgments from Mechanical Turk with book appropriateness and evaluation results in time for the INEX proceedings. Evaluation results with relevance judgments for the statements split into their atomic aspects indicate that performance increases when matching named entities (persons and locations) from the statements with named entities in the pages.

The Structure Extraction task is conjoint with ICDAR and therefore ran a bit earlier than the other tasks, with a run submission deadline in May. A total of 9 organizations signed up, 6 of which submitted runs. This increase in active participants is probably a direct result of both 1) the availability of training data and 2) the removal of the requirement for participating organizations to create part of the ground truth. This round of the competition further provided rejoicing results, as for the first time since the competition started, one organization has beaten the baseline BookML format provided by MDCS (Microsoft



Development Center Serbia) in 2008. The University of Innsbruck indeed performed best in terms of link-based evaluation.

## 2.4 Outlook

Next year, we continue with the SBS task to further investigate the role of user information. We plan to run an additional pilot task for which we have a few options. One option is to investigate how we can use the interactivity in the forum thread to simulate interactive sessions. Another is to extend the original task by requiring systems to not only determine which book ISBNs to return, but also what information about those books to return. Book descriptions contain a mixture of professional metadata, user tags and up to 100 user reviews. A new challenge could be to determine which tags and reviews are relevant to the user in determining whether she wants to read a book or not.

The Prove It task attracted no new participants in the last two years and will not continue next year. We are considering a new task centered around entity recognition, such as identifying and mapping characters in novels.

The structure extraction task has reached a record high number of active participants, and has for the first time witnessed an improvement of the state the art. In future years, we aim to investigate the usability of the extracted ToCs, both for readers in navigating books and systems that index and search parts of books. To be able to build even larger evaluation sets, we hope to experiment with crowdsourcing methods. This may offer a natural solution to the evaluation challenge posed by the massive data sets handled in digitized libraries

## 3 Linked Data Track

In this section, we will briefly discuss the INEX 2013 Linked Data Track (addressing the searching structured or semantic data theme). Further details are in [4].

### 3.1 Aims and Tasks

The goal of the Linked Data track was to investigate retrieval techniques over a combination of textual and highly structured data, where RDF properties carry additional key information about semantic relations among data objects that cannot be captured by keywords alone. We intend to investigate if and how structural information could be exploited to improve ad-hoc retrieval performance, and how it could be used in combination with structured queries to help users navigate or explore large result sets via Ad-hoc queries, or to address Jeopardy-style natural-language queries which are translated into a SPARQL-based query format. The Linked Data track thus aims to close the gap between IR-style keyword search and Semantic-Web-style reasoning techniques. Our goal is to bring together different communities and to foster research at the intersection of Information Retrieval, Databases, and the Semantic Web.

For INEX 2013, we explored two different retrieval tasks that continue from INEX 2012:

- The classic Ad-hoc Retrieval task investigates informational queries to be answered mainly by the textual contents of the Wikipedia articles.
- The Jeopardy task employs natural-language Jeopardy clues which are manually translated into a semi-structured query format based on SPARQL with keyword conditions.

### 3.2 Test Collection

The Linked Data track used a subset of DBpedia 3.8 and YAGO2s together with a recent dump of Wikipedia core articles (dump of June 1st, 2012). Valid results are entities occurring in both Wikipedia and DBpedia (and hence in YAGO), hence we provided a complete list of valid URIs to the participants. In addition to these reference collections, we will also provide two supplementary collections: 1) to lower the participation threshold for participants with IR engines, a fusion of XML'ified Wikipedia articles with RDF properties from both DBpedia and YAGO2s, and 2) to lower the participation threshold for participants with RDF engines, a dump of the textual content of Wikipedia articles in RDF. Participants are explicitly encouraged to make use of more RDF facts available from DBpedia and YAGO2s, in particular for processing the reasoning-related Jeopardy topics.

The goal of the Ad-hoc Task is to return a ranked list of results in response to a search topic that is formulated as a keyword query. Results had to be represented by their Wikipedia page ID's, which in turn had to be linked to the set of valid DBpedia URI's. A set of 144 Ad-hoc task search topics for the INEX 2013 Linked Data track had been released in March 2013 and was made available for download from the Linked Data Track homepage. In addition, the set of QREls from the 2012 Ad-Hoc Task topics was provided for training.

These are familiar IR topics, an example is:

```
<topic id="2009002">
  <title>best movie</title>
  <description>information of classical movies</description>
  <narrative>
    I spend most of my free time seeing movies. Recently, I want to
    retrospect some classical movies. Therefore, I need information about
    the awarded movies or movies with good reputation. Any information,
    such as the description or comments of the awarded movies on famous
    filmfests or movies with good fame, is in demand.
  </narrative>
</topic>
```

As in 2012, the Jeopardy task continued to investigate retrieval techniques over a set of natural-language Jeopardy clues, which were manually translated into SPARQL query patterns with additional keyword-based filter conditions. A set of 105 Jeopardy task search topics, out of which 74 topics were taken over from 2012 and 31 topics were newly added to the 2013 setting. 72 single-entity

topics (with one query variable) were also included into the set of 144 Ad-hoc topics. All topics were made available for download in March 2013 from the Linked Data Track homepage. In analogy to the Ad-hoc Task, the set of topics from 2012 was provided together with their QREls for training. An example topic is:

```
<topic id="2013301" category="Falls">
  <jeopardy_clue>
    This river's 350-foot drop at the Zambia-Zimbabwe border creates this
    water falls.
  </jeopardy_clue>
  <keyword_title>
    river's 350-foot drop Zambia-Zimbabwe Victoria Falls
  </keyword_title>
  <sparql_ft>
    SELECT DISTINCT ?x ?o WHERE {
      ?x <http://dbpedia.org/property/watercourse>
      ?o . FILTER FTContains (?x, "Victoria Falls") .
      FILTER FTContains (?o, "river water course Victoria 350-foot drop
      Zimbabwe") .
    }
  </sparql_ft>
</topic>
```

### 3.3 Results

In total, 4 ad-hoc search runs were submitted by 3 participants and 2 valid Jeopardy! runs were submitted by 1 participant. Assessments for the Ad-hoc Task were done on Amazon Mechanical Turk by pooling the top-100 ranks from the 4 submitted runs in a round-robin fashion. Conversely, the top-10 results were pooled from the 3 Jeopardy submissions for the single-entity topics and by pooling the top-20 for the multi-entity topics, respectively, again in a round-robin fashion. A total of 72 Ad-hoc topics and 77 Jeopardy topics were assessed.

The TREC-eval tool was adapted to calculate the following well-known metrics (see [1, 5]) used in ad-hoc and entity ranking settings: Precision, Recall, Average-Precision (AP), Mean-Average-Precision (MAP), Mean-Reciprocal-Rank (MRR), and Normalized-Discounted-Cumulated-Gain (NDCG). The best scoring submission for the ad hoc task was *ruc-all-2200-paragraph-80* by the **Renmin University of China (RUC)** with a MAiP of 0.3880. The best scoring submission for the Jeopardy task was *MPIUltimatum\_Phrase* by the **Max-Planck Institute for Informatics (MPI)** with a MRR of 0.7671.

Given the low number of submissions, it is difficult to draw general conclusions from the runs, but individual participants found various interesting results demonstrating the value of the build test collection for research in this important emerging area. We hope and expect that the test collection will be (re)used by researchers for future experiments in this active area of research.

### 3.4 Outlook

The Linked Data Track was organized towards our goal to close the gap between IR-style keyword search and Semantic-Web-style reasoning techniques. The track thus continues one of the earliest guiding themes of INEX, namely to investigate whether structure may help to improve the results of ad-hoc keyword search. A key contribution is the introduction of a new and much larger supplementary XML collection, coined *Wikipedia-LOD v2.0*, with XML-ified Wikipedia articles which were additionally annotated with RDF properties from both DBpedia 3.8 and YAGO2. However, due to the very low number of participating groups, in particular for the Jeopardy, detailed comparisons of the underlying ranking and evaluation techniques can only be drawn very cautiously.

## 4 Tweet Contextualization Track

In this section, we will briefly discuss the INEX 2013 Tweet Contextualization Track (one of the two tracks addressing the focused retrieval theme). Further details are in [2].

### 4.1 Aims and Tasks

Twitter is increasingly used for on-line client and audience fishing, this motivated the proposal of a new track addressing tweet contextualization. The objective of this task is to help a user to understand a tweet by providing him with a short summary (500 words). This summary should be built automatically using local resources like the Wikipedia and generated by extracting relevant passages and aggregating them into a coherent summary. The task is evaluated considering informativeness which is computed using a variant Kullback-Leibler divergence and passage pooling. Meanwhile effective readability in context of summaries is checked using binary questionnaires on small samples of results. Running since 2010 as a complex QA track at INEX, the results showed that only systems that efficiently combine passage retrieval, sentence segmentation and scoring, named entity recognition, text POS analysis, anaphora detection, diversity content measure as well as sentence reordering are effective.

### 4.2 Test Collection

The document collection has been built based on a recent dump of the English Wikipedia from November 2011. This date is anterior to all selected topics. Since we target a plain XML corpus for an easy extraction of plain text answers, we removed all notes and bibliographic references that are difficult to handle and kept only non empty Wikipedia pages (pages having at least one section). Resulting documents consist of a title (**t**itle), an abstract (**a**) and sections (**s**). Each section has a sub-title (**h**). Abstract and sections are made of paragraphs (**p**) and each paragraph can contain entities (**t**) that refer to other Wikipedia pages.

In 2012, topics were made of 53 tweets from New York Times (NYT). In 2013, the task was enriched and evaluated topics were made of 120 tweets manually collected by organizers. These tweets were selected and checked, in order to make sure that:

- They contained “informative content” (in particular, no purely personal messages); Only non-personal accounts were considered (*i.e.* @CNN, @TennisTweets, @PeopleMag, @science...).
- The document collection from Wikipedia contained related content, so that a contextualization was possible.

From the same set of accounts, more than 1,800 tweets were then collected automatically. These tweets were added to the evaluation set, in order to avoid that fully manual, or not robust enough systems could achieve the task. All tweets were then to be processed by participants, but only the 120 short list was used for evaluation. Participants did not know which topics were selected for evaluation. These tweets were provided in a text-only format without metadata and in a JSON format with all associated metadata.

### 4.3 Measures

Tweet contextualization is evaluated on both informativeness and readability. Informativeness aims at measuring how well the summary explains the tweet or how well the summary helps a user to understand the tweet content. On the other hand, readability aims at measuring how clear and easy to understand the summary is. Informativeness measure is based on lexical overlap between a pool of relevant passages (RPs) and participant summaries. Once the pool of RPs is constituted, the process is automatic and can be applied to unofficial runs. The release of these pools is one of the main contributions of Tweet Contextualization tracks at INEX [8]. By contrast, readability is evaluated manually and cannot be reproduced on unofficial runs. In this evaluation the assessor indicates where he misses the point of the answers because of highly incoherent grammatical structures, unsolved anaphora, or redundant passages.

Three metrics were used: **Relevancy (or Relaxed) metric**, counting passages where the T box has not been checked (*Trash* box if the passage does not make any sense in the context of the previous passages); **Syntax**, counting passages where the S box was not checked either (*i.e.*, the passage has no syntactic problems), and the **Structure (or Strict) metric** counting passages where no box was checked at all. In all cases, participant runs were ranked according to the average, normalized number of words in valid passages.

### 4.4 Results

A total number of 13 teams from 9 countries (Brasil, Canada, France, India, Ireland, Mexico, Russia, Spain, USA) submitted runs to the Tweet Contextualization track in 2013. This year, the best participating system *256* from **Université de Nantes** used hashtag preprocessing. The best run by this participant

used all available tweet features including web links which was not allowed by organizers. However their second best run *258* without using linked web pages is ranked first among official runs. Second best participant on informativeness was run *275* from **IRIT, Toulouse** which score best in readability and used state of the art NLP tools. Third best participant was run *254* from **University of Minnesota Duluth** was first in 2012, suggesting that their system performs well on a more diversify set of tweets in 2013.

All participants but two used language models, however informativeness of runs that only used passage retrieval is under 5%. Terminology extraction and reformulation applied to tweets was also used in 2011 and 2012. Appropriate stemming and robust parsing of both tweets and wikipedia pages are an important issue. All systems having a run among the top five in informativeness used the Stanford Core NLP tool or the TreeTagger. Automatic readability evaluation and anaphora detection helps improving readability scores, but also informativeness density in summaries. State of the art summarization methods based on sentence scoring proved to be helpful on this task. Best runs on both measures used them. Best run in 2013 also experimented a tweet tag scoring technique while generating the summary. Finally, this time the state-of-the-art system proposed by organizers since 2011 combining LM indexation, terminology graph extraction and summarization based on shallow parsing was not ranked among the ten best runs which shows that participant systems improved on this task over the three editions.

## 4.5 Outlook

The discussion on next year's track is only starting, and there are links to related activities in other CLEF labs that need to be further explored. The use case and the topic selection should remain stable in 2014 TC Track, so that 2013 topics can be used as a training set. Nevertheless, we will consider more diverse types of tweets, so that participants could better measure the impact of hashtag processing on their approaches.

## 5 Snippet Retrieval Track

In this section, we will briefly discuss the INEX 2013 Snippet Retrieval Track (one of the tracks addressing the focused retrieval theme). Further details are in [9].

### 5.1 Aims and Task

The goal of the snippet retrieval track is to determine how to generate informative snippets for search results. Such snippets should provide sufficient information to allow the user to determine the relevance of each document, without needing to view the document itself, allowing the user to quickly find what they are looking for.

The task was to return a ranked list of documents for the requested topic to the user, and with each document, a corresponding text snippet describing the document. Each run had to return 20 documents per topic, with a maximum of 180 characters per snippet. The snippets may be created in any way – they may consist of summaries, passages from the document, or any other text at all.

## 5.2 Collection

The Snippet Retrieval Track uses the exact same collection as the Tweet Contextualization track—an XML version of the English Wikipedia, based on a dump taken on November 2012. Since the task is to generate snippets for the documents given in the reference run, a link to an archive containing only those 700 documents (as well as the reference run submission file itself) was provided.

There were 35 topics in total—10 taken from the INEX 2010 Ad Hoc Track, and 25 created specifically for this track, with the goal being to create topics requesting more specific information than is likely to be found in the first few paragraphs of a document. Each topic contains a short content only (CO) query, a phrase title, a one line description of the search request, and a narrative with a detailed explanation of the information need, the context and motivation of the information need, and a description of what makes a document relevant or not.

## 5.3 Assessment and Evaluation

To determine the effectiveness of the returned snippets at their goal of allowing a user to determine the relevance of the underlying document, manual assessment is being used. Both snippet-based and document-based assessment are being used. The documents will first be assessed for relevance based on the snippets alone, as the goal is to determine the snippet's ability to provide sufficient information about the document. The documents will then be assessed for relevance based on the full document text, with evaluation based on comparing these two sets of assessments.

We created snippet assessment packages (the size of a single submission) to assess, each participating organization will receive as many packages as they have submitted runs. For each topic, the assessor will read through the details of the topic, after which they will read through each snippet, and determine whether or not the underlying document is relevant to the topic. This is expected to take around 1-2 hours per package. Ideally, each package should be assessed by a different person if feasible. Additionally, it will be required to perform one assessment of the document assessment package. For each of the 35 topics, the assessor is shown the full text of each of the 20 documents. They must read through enough of the document to determine whether or not it is relevant to the topic. This is expected to take around 3-7 hours, depending on the assessor.

Submissions are evaluated by comparing the snippet-based relevance judgements with the document-based relevance judgements, which are treated as a

ground truth. The primary evaluation metric used is the geometric mean of recall and negative recall (GM). A high value of GM requires a high value in recall and negative recall—i.e., the snippets must help the user to accurately predict both relevant and irrelevant documents. If a submission has high recall but zero negative recall (e.g. in the case that everything is judged relevant), GM will be zero. Likewise, if a submission has high negative recall but zero recall (e.g. in the case that everything is judged irrelevant), GM will be zero. Details of additional metrics used are given in [9].

## 5.4 Results

As of this writing, only preliminary results are available. The best scoring system is *snippets\_2013\_knapsack* of **IRIT, Toulouse**, with a GM score of 0.5352. The second scoring run is *QUT\_2013\_Focused* of **Queensland University of Technology (QUT)** with a GM score of 0.4774. Further discussion of the results will be available in [9].

## 5.5 Outlook

We have discussed the setup of the track, and presented the preliminary results of the track. The preliminary results show that in all submitted runs, poor snippets are causing users to miss over half of all relevant results, indicating that a lot of work remains to be done in this area. Final results will be released at a later date, once further document assessment has been completed.

## 6 Envoi

This completes our walk-through of INEX 2013. INEX 2013 focused on three themes: *searching professional and user generated data* (Social Book Search track); *searching structured or semantic data* (Linked Data track); and *focused retrieval* (Snippet Retrieval and Tweet Contextualization tracks). The last two tracks use the same Wikipedia corpus and both address focused retrieval in the form of constructing some concise selection of information in a form that is of interest to NLP researchers (tweet contextualization) and to IR researchers (snippet retrieval). The INEX tracks cover various aspects of focused retrieval in a wide range of information retrieval tasks. This overview has only touched upon the various approaches applied to these tasks, and their effectiveness. The online proceedings of CLEF 2013 contains both the track overview papers, as well as the papers of the participating groups. The main result of INEX 2013, however, is a great number of test collections that can be used for future experiments, and the discussion amongst the participants that happens at the CLEF 2013 conference in Valencia and throughout the year on the discussion lists.



## References

- [1] Amer-Yahia, S., Lalmas, M.: XML search: languages, INEX and scoring. *SIGMOD Record* 35 (2006)
- [2] Bellot, P., Moriceau, V., Mothe, J., Sanjuan, E., Tannier, X.: Overview of the INEX 2013 tweet contextualization track. In: *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes* (2013)
- [3] Doucet, A., Kazai, G., Colutto, S., Muehlberger, G.: Overview of the ICDAR 2013 Competition on Book Structure Extraction. In: *Proceedings of the Twelfth International Conference on Document Analysis and Recognition (ICDAR 2013)*, Washington, USA (September 2013)
- [4] Gurajada, S., Kamps, J., Mishra, A., Schenkel, R., Theobald, M., Wang, Q.: Overview of the INEX 2013 linked data track. In: *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes* (2013)
- [5] Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., Robertson, S.: INEX 2007 evaluation measures. In: Fuhr, N., Kamps, J., Lalmas, M., Trotman, A. (eds.) *INEX 2007*. LNCS, vol. 4862, pp. 24–33. Springer, Heidelberg (2008)
- [6] Koolen, M., Kamps, J., Kazai, G.: Social Book Search: The Impact of Professional and User-Generated Content on Book Suggestions. In: *Proceedings of the International Conference on Information and Knowledge Management (CIKM 2012)*. ACM (2012)
- [7] Koolen, M., Kazai, G., Preminger, M., Doucet, A.: Overview of the INEX 2013 social book search track. In: *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes* (2013)
- [8] SanJuan, E., Bellot, P., Moriceau, V., Tannier, X.: Overview of the INEX 2010 question answering track (QA@INEX). In: Geva, S., Kamps, J., Schenkel, R., Trotman, A. (eds.) *INEX 2010*. LNCS, vol. 6932, pp. 269–281. Springer, Heidelberg (2011)
- [9] Trappett, M., Geva, S., Trotman, A., Scholer, F., Sanderson, M.: Overview of the INEX 2013 snippet retrieval track. In: *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes* (2013)

# Recent Trends in Digital Text Forensics and Its Evaluation

## Plagiarism Detection, Author Identification, and Author Profiling

Tim Gollub<sup>1</sup>, Martin Potthast<sup>1</sup>, Anna Beyer<sup>1</sup>, Matthias Busse<sup>1</sup>,  
Francisco Rangel<sup>2,3</sup>, Paolo Rosso<sup>3</sup>, Efstathios Stamatatos<sup>4</sup>, and Benno Stein<sup>1</sup>

<sup>1</sup> Web Technology and Information Systems, Bauhaus-Universität Weimar, Germany

<sup>2</sup> Autoritas Consulting, S.A., Spain

<sup>3</sup> Natural Language Engineering Lab, ELiRF, Universitat Politècnica de València, Spain

<sup>4</sup> Dept. of Information and Communication Systems Engineering,

University of the Aegean, Greece

pan@webis.de

<http://pan.webis.de>

**Abstract** This paper outlines the concepts and achievements of our evaluation lab on digital text forensics, PAN 13, which called for original research and development on plagiarism detection, author identification, and author profiling. We present a standardized evaluation framework for each of the three tasks and discuss the evaluation results of the altogether 58 submitted contributions. For the first time, instead of accepting *the output* of software runs, we collected *the softwares themselves* and run them on a computer cluster at our site. As evaluation and experimentation platform we use TIRA, which is being developed at the Webis Group in Weimar. TIRA can handle large-scale software submissions by means of virtualization, sandboxed execution, tailored unit testing, and staged submission. In addition to the achieved evaluation results, a major achievement of our lab is that we now have the largest collection of state-of-the-art approaches with regard to the mentioned tasks for further analysis at our disposal.

## 1 Introduction

Nowadays, people increasingly share their work online, contribute to open projects and engage in web-based social interactions. The ease and the anonymity with which all of this can be done raises concerns about verifiability and trust: is a given text an original? Is an author the one who she claims to be? Does a piece of information originate from a trusted source? Answers to these and similar questions are crucial in order to deal with and rely on information obtained online, while the scale at which answers should be given calls for an automatic means. Specific tasks that address these questions include plagiarism detection, author identification, and author profiling, whereas tackling them requires expertise from diverse areas such as text forensics, computer linguistics, machine learning, and information retrieval, rendering research on these tasks a challenge.

Besides expertise, the research and development of solutions to these tasks is clearly limited due to the absence of representative evaluation resources and solid implementations of state-of-the-art approaches [18]. Moreover, researchers frequently lack the

time and budget to acquire these resources themselves while researching their own approach. As a consequence, evaluations are often performed in an ad hoc manner, and only the data and approaches that are easily accessible are used to evaluate a new idea—a fact that impedes the comparability of published evaluation results significantly. This undesirable development can be mitigated by the development of standardized benchmarks and evaluation frameworks: in case of a widespread adoption by the community, individual researchers can compare their approaches independently, simply by following the evaluation guidelines of a given framework. However, the effort to develop and spread standardized evaluation frameworks is considerable, so that such frameworks emerge typically only for very popular tasks, whereas for the less studied tasks the quality depends on individual initiatives. To foster such initiatives, evaluation conferences are organized in order to bring together the stakeholders of a given task in the form of labs, where some develop a new evaluation framework, and others team up to develop approaches that are run against such a framework.

The typical modus operandi of such evaluation conferences can be summarized as follows: the organizers of a lab hand out a data set of instances of a task's underlying problem, which are downloaded and processed by the participants. They in turn compute and submit sets of solutions (so-called runs) to the problem instances, which are then evaluated by the organizers against an undisclosed gold standard of solutions. Beforehand, the organizers often hand out data sets comparable to the test data for training and development purposes. This process minimizes the “interface” between participants and organizers since they only need to agree on data formats. As a result, the evaluation resources provided by the task organizers may be used by other researchers later on, in order to compare their approaches against those of a lab's participants.

While organizing a lab this way requires least effort of all involved parties, there are also downsides with this approach: participants are not incentivized to *publish their software*, i.e., after the lab has passed other data sets cannot be evaluated by members of the community, and, the exact steps of how the participants obtained their results cannot be traced unless their approach is reimplemented. Taking into account that (1) even the best researchers make errors (including lab organizers), (2) devising an evaluation framework is a difficult engineering task, and (3) evaluation methodology evolves at rapid pace, the “classical” lab organization approach lacks long-term sustainability and reproducibility in first place.

In this paper we show how these shortcomings can be addressed in the context of digital text forensics: our contributions include the large-scale evaluation of 19 plagiarism detectors, 18 author identifiers, and 21 author profilers. Unlike traditional labs we do not collect software runs (outputs) but the softwares themselves, and evaluate them at our site. Our evaluation lab is the first that entirely switches to software submissions instead of run submissions. For this purpose we develop the TIRA experimentation platform, which facilitates such kinds of evaluations so that few staff can conduct the evaluation part-time. The outcome of our evaluation is not only a table of performance values and a data set, but also a collection of state-of-the-art implementations of a diversity of approaches to the three tasks. Given their original authors' consent, they can be readily used (via the web frontend of TIRA) by the community for comparison purposes, even on different data sets.

The remainder of this paper is organized as follows: after a brief discussion of related work, the Sections 2, 3, and 4 present insights into the evaluation results obtained for plagiarism detection, author identification, and author profiling respectively. Section 5 details our evaluation setup that handles the software submissions, and Section 6 draws conclusions.

## 1.1 Related Work

Before going into details, we review related work on evaluating plagiarism detectors, author identifiers, and author profilers, as well as related online evaluation platforms.

**Plagiarism Detection.** In recent years, the evaluation of plagiarism and text reuse detectors has been studied in the context of the PAN evaluation labs that have been organized annually since 2009. For the purpose of these labs, we developed the first standardized evaluation framework which comprises a series of corpora of (semi-)automatically generated plagiarism as well as detection performance measures [43].<sup>1</sup> During the first three editions of the lab, a total of 43 plagiarism detectors have been evaluated using this framework [41, 42, 44]. The two recent editions refocused on specific sub-problems of plagiarism detection, namely source retrieval and text alignment. This also included the development of new corpora for these problems. Instead of again applying a semiautomatic approach to corpus construction, a large corpus of manually generated plagiarism has been crowdsourced in order to increase the level of realism [48]. This corpus comprises 297 essays of about 5000 words length, written by professional writers. In this regard the writers were given a set of topics to choose from along with two more technical rules: (1) to use the ChatNoir search engine [46] to research their topic of choice, and (2) to reuse text passages from retrieved web pages in order to compose their essay. The resulting essays represent the to-date largest corpus of realistic text reuse cases available, and they have been employed to evaluate another 33 plagiarism detectors in the past two labs [45, 47]. Besides the mentioned corpora, there are two other ones that comprise text reuse, namely the Meter corpus [9] and the Clough09 corpus [8]. The former contains 445 cases of text reuse among 1716 news articles, whereas the latter contains 57 short cases of manually generated plagiarism. To the best of our knowledge, these corpora have not yet been used in a large-scale evaluation of text reuse or plagiarism detectors.

**Author Identification.** Author identification has many possible settings. Previous competitions on this task focused on closed-set and open-set classification problems with multiple candidate authors [2, 25, 26]. The evaluation corpora comprised a set of problems of similar form, i.e., a number of texts from a set of known authors and a number of texts of unknown authorship; the evaluation measures included traditional information retrieval measures such as micro- and macro-averaged accuracy, precision, recall and  $F_1$ . Author verification has been studied in the framework of the PAN 11 lab [2]. In contrast to our lab's setting, each problem comprised multiple test texts.

<sup>1</sup> The corpora PAN-PC-2009/2010/2011 are available at <http://www.webis.de/research/corpora>

Therefore, precision, recall, and  $F_1$  per author (problem) were used for the evaluation of the participant methods. Accuracy and macro-average  $F_1$  were also used in the evaluation of the well-known “unmasking” method [29]. In a recent work, Koppel and Winter [27] studied a similar problem where, given a pair of documents, the question is whether or not they are written by the same person. In addition to accuracy, they use recall-precision curves to provide a more complete picture of the performance of the examined models. Taking into account the nature of the practical applications involved with the task of author verification, it is crucial to estimate the ability of the attribution models to assign high confidence scores to their correct answers.

**Author Profiling.** Our lab is the first to offer author profiling as an evaluation task. Therefore we review previous evaluations and data sets, where classification accuracy has been used in almost all cases as a performance measure. Pennebaker *et al.* [40] connected language use with personality traits, studying how the variation of linguistic characteristics in a text can provide information regarding gender and age of its author. Argamon *et al.* [3] analyzed formal written texts extracted from the British National Corpus, combining function words with part-of-speech features, and achieved approximately 80% accuracy in gender prediction. Other research investigated how to obtain age and gender information from formal texts [7, 22]. With the rise of the social media, Koppel *et al.* [28] built a dataset of blog posts and studied the problem of automatically determining an author’s gender based on proposing combinations of simple lexical and syntactic features, also achieving approximately 80% accuracy. Schler *et al.* [51] collected more than 71,000 blog posts and used a set of stylistic features such as non-dictionary words, parts-of-speech, function words and hyperlinks, combined with content features, such as word unigrams with the highest information gain. They also obtained an accuracy of about 80% for gender identification, and about 75% for age identification. Goswami *et al.* [20] added some new features to Schler’s work, such as slang words and the average length of sentences, improving accuracy to 80.3% in age group detection and to 89.2% in gender detection. Peersman *et al.* [38] compiled a dataset for the purpose of gender and age prediction from Netlog.<sup>2</sup> Studying short texts, Zhang and Zhang [58] experimented with segments of blog posts and obtained 72.1% accuracy for gender prediction. Similarly, Nguyen *et al.* [35] studied the use of language and age among Dutch Twitter users. They modelled age as a continuous variable (as they had previously done in [36]), and used a prediction approach based on logistic regression. They also measured the effect of gender in the performance of age detection, considering both variables as interdependent, and achieved correlations of up to 0.74 and mean absolute errors between 4.1 and 6.8 years.

**Online Evaluation Platforms.** Based on our previous work in developing the TIRA experimentation framework [17–19], we revisit and update the related work. Our assessment of existing frameworks is based on the needs for local instantiation, web dissemination, platform independence, result retrieval, and peer to peer collaboration; Table 1 gives an overview. (1) The need for local instantiation arises from the fact that data may be kept confidential—i.e., the framework must be able to reside with the data instead of the other way around. External researchers then can use the service

---

<sup>2</sup> <http://www.netlog.com>

**Table 1.** Assessment of existing experimentation frameworks with respect to our five proposed design goals (1) local instantiation, (2) web dissemination, (3) platform independence, (4) result retrieval, and (5) peer to peer collaboration. The top six tools are non-commercial, developed out of universities, the bottom four are commercial ones.

Tool	[Reference]	Domain	Design Goal				
			1	2	3	4	5
evaluatIR	[5] <sup>1</sup>	IR	×	✓	✓	✓	×
OpenML	[6] <sup>2</sup>	ML	×	×	×	✓	×
MLComp	<sup>3</sup>	ML	×	✓	×	✓	×
myExperiment	[10] <sup>4</sup>	any	×	✓	✓	✓	×
NEMA	[12] <sup>5</sup>	IR	×	✓	×	✓	×
TunedIT	[57] <sup>6</sup>	ML, DM	✓	✓	×	✓	×
TIRA	[19] <sup>7</sup>	any	✓	✓	✓	✓	×
Google Code Jam	<sup>8</sup>	Algorithms	×	×	✓	✓	×
Kaggle	<sup>9</sup>	ML, DM	×	×	✓	×	×
TopCoder	<sup>10</sup>	any	×	×	✓	✓	×
Yahoo Pipes	<sup>11</sup>	Web	×	✓	×	×	×

<sup>1</sup> <a href="http://www.evaluatir.org">http://www.evaluatir.org</a>	<sup>7</sup> <a href="http://tira.webis.de">http://tira.webis.de</a>
<sup>2</sup> <a href="http://www.openml.org">http://www.openml.org</a>	<sup>8</sup> <a href="http://www.google.com/codejam">http://www.google.com/codejam</a>
<sup>3</sup> <a href="http://www.mlcomp.org">http://www.mlcomp.org</a>	<sup>9</sup> <a href="http://www.kaggle.com">http://www.kaggle.com</a>
<sup>4</sup> <a href="http://www.myexperiment.org">http://www.myexperiment.org</a>	<sup>10</sup> <a href="http://www.topcoder.com">http://www.topcoder.com</a>
<sup>5</sup> <a href="http://www.music-ir.org">http://www.music-ir.org</a>	<sup>11</sup> <a href="http://pipes.yahoo.com">http://pipes.yahoo.com</a>
<sup>6</sup> <a href="http://www.tunedit.org">http://www.tunedit.org</a>	

for comparison and evaluation of their own research hypotheses, whilst the experiment provider is in full control of the experiment resources. Apart from TIRA, this goal is currently only achieved by Tunedit. (2) Web dissemination is another important factor when developing an experimentation framework since it allows researchers to link the results in a paper with the experiment service used to produce them. Especially for standard preprocessing tasks or evaluations on private data, such a web service can become a frequently cited resource. However, not all frameworks currently pursue this goal. For example, Kaggle and TopCoder target commercial customers who typically refrain from sharing their assets, whereas Google Code Jam currently targets only scholars by organizing one-time competitions for education purposes. (3) The sophisticated and varying platform requirements of research experiments (as well as individual coding preferences of software developers) render the development constraints imposed by an experimentation framework critical for its success. Ideally, software developers can deploy experiments as a service that is unconstrained by the underlying operating system, parallelization paradigm, programming language, or data format. Local instantiation is a key to achieve this goal. Furthermore, the framework should operate as a layer on top of the experiment software and should use, instead of close intra-process communication such as in Tunedit, standard inter-process communication on the POSIX level to exchange information. (4) For computationally expensive retrieval tasks, the maintenance of a public result repository can become a valuable asset since it allows others to

reuse them. Almost all frameworks support this goal. (5) Finally, by fostering peer-to-peer collaboration, a framework can drive a standardization process while maintaining a central repository of related evaluation resources. Note that currently none of the experimentation platforms implements peer-to-peer collaboration, though some have related functions.

## 2 Plagiarism Detection

This section briefly reports on the results of evaluating 18 plagiarism detectors that have been submitted to our evaluation lab. An extended version of this evaluation report can be found in [47], where a more in-depth analysis of the obtained results as well as a survey of detection approaches is given. To evaluate plagiarism and text reuse detectors, we measure their performance with regard to the two tasks source retrieval and text alignment, both of which are important parts of detectors that detect plagiarism from the web [54]. In the former task, a detector retrieves likely candidates from which text may have been reused in a suspicious document. In the latter task, the suspicious document is compared to selected candidates in closer detail. In the remainder of the section, we review the evaluation resources for each task individually and present the results of using it to evaluate the submitted detectors.

### 2.1 Source Retrieval

In source retrieval, given a suspicious document and a web search engine, the task is to retrieve all source documents from which text has been reused whilst minimizing retrieval costs. The cost-effectiveness of plagiarism detectors in this task is important since using existing search engines is perhaps the only feasible way for researchers as well as small and medium-sized businesses to implement plagiarism detection against the web, whereas search companies charge considerable fees for automatic usage. To study this task, we employ a controlled, static web environment, which consists of a large web crawl and search engines indexing it. Using this setup, we built a large corpus of manually generated text reuse in the form of essays, which serve as suspicious documents and which are fed into a plagiarism detector. The detection results returned are evaluated using tailored performance measures derived from precision and recall as well as cost-effectiveness statistics. Before discussing the actual performances obtained, we describe each of these resources in some detail.

**Evaluation Setup.** Evaluating source retrieval in a reproducible, yet representative manner is a difficult endeavor, since this requires a search engine that indexes a representative portion of the web in a way so that the result sets of queries do not change, even after years. Commercial search engines are under constant development, so that they do not meet this constraint. Therefore, we resort to the current most representative research search engines Indri<sup>3</sup> and ChatNoir [46], which both index the ClueWeb09 corpus,<sup>4</sup> a 2009 web crawl of about one billion web pages, half of which are English ones.

---

<sup>3</sup> <http://lemurproject.org/clueweb09/index.php#Services>

<sup>4</sup> <http://lemurproject.org/clueweb09>

Since the ClueWeb corpus is static, the search engines that index it can be considered static as well, presuming their underlying retrieval models are not severely changed in the future. In order to independently measure the cost-effectiveness of source retrieval algorithms, we monitor access to the search engines by means of a central search proxy service. All source retrieval algorithms submitted to our lab used this service to retrieve sources for a given suspicious document. The service accepts search requests for Indri and ChatNoir and returns their search results in a unified format. Moreover, it serves web pages from the ClueWeb on demand. Besides unifying the search interfaces and result formats for the convenience of developers, all accesses to the search engines as well as the ClueWeb are logged minutely. This way, the performance of a source retrieval algorithm can be measured by analyzing the logs obtained after running it.

**Evaluation Corpus.** As a realistic evaluation corpus we employ the Webis Text Reuse Corpus 2013 (Webis-TRC-13) [48]. The corpus has been constructed entirely manually and consists of 297 essays of about 5000 words length the contents of which have been reused from ClueWeb pages. The writers who wrote these essays were instructed to find web pages that match their respective essay's topic using the aforementioned ChatNoir search engine. If they decided to reuse a certain passage from a given web page, their instructions were to edit the reused text as thoroughly as they thought necessary to avoid detection. The modifications made include paraphrasing of the text itself as well as interleaving of reused passages from different sources. The average number of edits made on an essay is 2132.4, whereas the standard deviation is 1444.9. The average number of different sources used is 15.4, and the standard deviation 10. A subset of 40 essays of the Webis-TRC-13 was chosen as training documents, and 58 essays for testing. Based on this data, the source retrieval algorithms submitted to our lab were presented with a realistic retrieval setting, since it can be assumed that plagiarists as well as plagiarism detectors use the same search infrastructure to search for sources.

**Performance Measures.** To assess the performance of a source retrieval algorithm, we measure its retrieval performance and the cost-effectiveness of obtaining its results. Retrieval performance is measured as precision, recall, and  $F_1$  of retrieved sources regarding downloaded documents for a given suspicious document. The computation of precision and recall per suspicious document, however, is not straightforward, since each individual source of a given document may have a number of duplicates in the ClueWeb. These duplicates are not known a priori, so that each downloaded document has to be checked whether or not it is a duplicate of one of the sources of the suspicious document in question. If a downloaded document turns out to be a source duplicate, it is treated as a true positive detection (i.e., as if the original source had been found). However, retrieving more than one duplicate of a source document does not increase recall beyond that of retrieving just one, since no additional information is added by finding more duplicates of the same document. Conversely, retrieving more than one duplicate of a source document does not decrease precision, since they are not false positives. A detailed definition of what constitutes a source duplicate is beyond the scope of this overview, but can be found in [47].

Cost-effectiveness is measured as average workload per suspicious document, and as average numbers of queries and downloads until the first true positive detection has



**Table 2.** Source retrieval results with respect to retrieval performance and cost-effectiveness

Team (alphabetical order)	Downloaded Sources			Total Workload		Time to 1st Detection		No Runtime Detection	
	F <sub>1</sub>	Precision	Recall	Queries	Downloads	Queries	Downloads		
Elizalde	0.17	0.12	0.44	44.50	107.22	16.85	15.28	5	241.7 m
Vesely	0.15	0.11	0.35	161.21	81.03	184.00	5.07	16	655.3 m
Gillam	0.04	0.02	0.10	16.10	33.02	18.80	21.70	38	<b>15.1 m</b>
Haggag	0.44	<b>0.63</b>	0.38	32.04	<b>5.93</b>	8.92	<b>1.47</b>	9	152.7 m
Kong	0.01	0.01	<b>0.65</b>	48.50	5691.47	2.46	285.66	<b>3</b>	4098.0 m
Lee	0.35	0.50	0.33	44.04	11.16	7.74	1.72	15	310.5 m
Nourian	0.10	0.15	0.10	<b>4.91</b>	13.54	<b>2.16</b>	5.61	27	25.3 m
Suchomel	0.06	0.04	0.23	12.38	261.95	2.44	74.79	10	1637.9 m
Williams	<b>0.47</b>	0.55	0.50	116.40	14.05	17.59	2.45	5	1163.0 m

been made. These statistics reveal if a source retrieval algorithm finds sources quickly, thus reducing the costs of using it.

**Evaluation Results.** Table 2 shows the performances of the nine plagiarism detectors that implemented source retrieval. Since there is currently no formula to organize retrieval performance and cost-effectiveness into an absolute order, the detectors are ordered alphabetically, whereas the best performance value for each metric is highlighted. As can be seen, there is no single detector that performs best on all accounts. Rather, different detectors have different characteristics. The detector of Williams *et al.* [56] achieves the best trade-off between precision and recall and therefore the best F<sub>1</sub> value. This detector is followed closely by that of Haggag and El-Beltagy [21], which achieves best precision but mediocre recall, whereas the detector of Kong *et al.* [31] achieves best recall at the cost of poor precision. It is not easy to decide which of these detectors solves the task best, since each of them may have their justification in practice. For example, the detector of Haggag and El-Beltagy downloads only about six documents on average per suspicious document and minimizes the time to first detection. Despite the excellent trade-off of Williams *et al.*'s detector, it incurs the second-highest costs in terms of queries on average, which is more than thrice as much as the other mentioned detectors. Kong *et al.*'s detector has highest download costs, but one may argue that downloads are much cheaper than queries, and that in this task recall is more important than precision.

## 2.2 Text Alignment

In text alignment, given a pair of documents, the task is to identify all contiguous passages of reused text between them. The challenge with this task is to identify passages of text that have been obfuscated, sometimes to the extent that, apart from stop words, little lexical similarity remains between an original passage and its plagiarized counterpart. Consequently, for evaluators, the challenge is to provide a representative corpus of documents that emulate this situation. To study this task, we employ a similar corpus construction methodology that has been used in previous evaluations of this task, while

**Table 3.** Text alignment results with retrieval performance and runtime

Team	PlagDet	Recall	Precision	Granularity	Runtime
R. Torrejón	0.82220	0.76190	0.89484	1.00141	1.2 m
Kong	0.81896	0.81344	0.82859	1.00336	6.1 m
Suchomel	0.74482	0.76593	0.72514	1.00028	28.0 m
Saremi	0.69913	0.77123	0.86509	1.24450	446.0 m
Shrestha	0.69551	0.73814	0.87461	1.22084	684.5 m
Palkovskii	0.61523	0.53561	0.81699	1.07295	6.5 m
Nourian	0.57716	0.43381	0.94707	1.04343	40.1 m
baseline	0.42191	0.34223	0.92939	1.27473	30.5 m
Gillam	0.40059	0.25890	0.88487	1.00000	21.3 m
Jayapal	0.27081	0.38187	0.87901	2.90698	4.8 m

fixing some of its deficiencies. We evaluate the performance of plagiarism detectors based on the traditionally employed measures.

**Evaluation Corpus.** The evaluation corpus for text alignment is also based on the aforementioned Webis-TRC-13. But instead of employing the essays of that corpus directly, pairs of documents that comprise reused passages have been constructed automatically, as was done in previous years [43]. One frequent point of criticism about automatically generating plagiarism is that it is difficult to ensure that documents between which text is plagiarized are about the same topic, so that the plagiarism could be detected simply by analyzing topic drift [48]. Using the documents that have been retrieved manually as sources for the essays of the Webis-TRC-13 as a basis for constructing plagiarism cases, however, allows us to mitigate this problem.

The corpus consists of pairs of documents about the same topic that share passages of text. These passages have been automatically obfuscated to emulate plagiarist behavior. We apply three basic obfuscation strategies, namely paraphrasing through naive random text operations and through cyclic translations, and summarization. Naive random text operations include shuffling, adding, removing, and replacing words at random while using WordNet as a source of word replacements and while optionally maintaining the original passage's part-of-speech sequence. Cyclic translations include, for example, translating a text from English to Japanese to Spanish and back to English using on-line translation services such as Google Translate. Summaries have been obtained by including an additional language resource from the Document Understanding Conference 2001 corpus for text summarization.<sup>5</sup> The corpus contains in total 1826 suspicious documents and 3169 source documents, which are grouped into 5000 pairs, so that there are 3000 pairs containing plagiarism (i.e., 1000 for each of the mentioned obfuscation strategies), 1000 containing unobfuscated plagiarism, and 1000 without plagiarism.

**Evaluation Results.** Table 3 shows the overall performance of nine plagiarism detectors that implemented text alignment. The detailed performances of each detector with regard to different kinds of obfuscation can be found in [47]. Performances are measured using precision and recall at character level as well as granularity (i.e., how

<sup>5</sup> [http://www-nlpir.nist.gov/projects/duc/data/2001\\_data.html](http://www-nlpir.nist.gov/projects/duc/data/2001_data.html)

often the same plagiarism case is detected). These values are combined into the PlagDet score by dividing  $F_1$  value of precision and recall by the granularity's logarithm. The two top-ranked detectors of Rodríguez Torrejón and Martín Ramos [49] and Kong *et al.* [31] achieve similar PlagDet scores, but differ in precision and recall. These detectors as well as that of Suchomel *et al.* [55] have been evaluated in previous years, all of which implement text alignment under the seed-and-extend paradigm: seeds, which encode positions of exact overlap between a pair of documents, are identified and then aligned into passages based on their pairwise distance. Examples for seeds include 5-grams, and stop word 8-grams [53].

### 3 Author Identification

Authorship attribution is an important problem in many areas including information retrieval and computational linguistics, but also in applied areas such as law and journalism where knowing the author of a document (such as a ransom note) may be crucial to save lives. The most common framework for testing algorithms that solve this task is a closed-set text classification problem: given a sample of documents from a small, finite set of known candidate authors, the task is to determine for a document of unknown authorship, which author, if any, wrote the document in question [24, 52]. It has been commented, however, that this may be an unreasonably easy task [30]. A more demanding problem is author verification where, given a set of documents by a single author and a document of unknown authorship, the task is to determine if the document was written by that particular author or not [29]. This setting more accurately reflects real life in the experiences of professional forensic linguists, who are often called upon to answer this kind of question. Interestingly, every author identification problem with multiple candidate authors can be transformed to a set of author verification problems. Therefore, the ability to effectively deal with author verification is fundamental in author identification research.

**Evaluation Setup.** The author identification task of our lab is set up as follows: given a small set (no more than 10, possibly as few as one) of “known” documents by a single person and an “unknown” document, the task is to determine whether the unknown document was written by the same person who wrote the known document set. The participants were given several problems of this form in three natural languages: English, Greek, and Spanish. One problem comprises a set of known documents by a single person and exactly one unknown document. The number of known documents per problem varies from 1 to 10. All documents within a single problem are in the same language, and best efforts were applied to assure that within-problem documents are matched for genre, register, theme, and date of writing. Moreover, the length of the documents varies from a few hundred to a few thousand words. The participants were asked to develop their software so that they can handle any set of such author verification problems in the specified languages. For each problem, they have to generate a binary answer (“yes”, if the unknown document was written by that author or “no”, if the unknown document was not written by that author). It was also possible to leave some problems unanswered. In addition, the participants could optionally produce a

confidence score, namely a real number in the interval  $[0, 1]$  where 1 means that it is absolutely sure that the unknown document was written by that author and 0 means the opposite.

**Evaluation Corpus.** The corpus we built for the author identification task covers three languages: English, Greek, and Spanish. For each language there is a set of problems, where one problem comprises a set of documents of known authorship by a single author and exactly one document of unknown authorship. All the documents within a problem are in the same language, placed in a separate folder, and the language information was encoded in the problem label (i.e., folder name). The training corpus comprised 10 problems in English, 20 problems in Greek and 5 problems in Spanish. The test corpus was more balanced across languages comprising 30 problems in English, 30 problems in Greek and 25 problems in Spanish. The English part of the corpus<sup>6</sup> consists of extracts from published textbooks on computer science and related disciplines. The Greek part of the corpus comprises newspaper articles published in the Greek weekly newspaper TO BHMA<sup>7</sup> from 1996 to 2012. The Spanish part of the corpus<sup>8</sup> consisted of excerpts from newspaper editorials and short fiction.

**Performance Measures.** The participants of our lab were asked to provide a simple “yes/no” binary answer for each problem of the author identification task. Optionally, in case a software was not confident enough for to decide a problem, it could be left unanswered. To evaluate the output of a software, we used the following measures:

$$\text{Recall} = \frac{\#\text{correct\_answers}}{\#\text{problems}} \qquad \text{Precision} = \frac{\#\text{correct\_answers}}{\#\text{answers}}$$

Note that in case a participant’s software provides answers all problems, these two measures are equal.

The final ranking was computed by combining these measures via  $F_1$  for the whole evaluation corpus comprising all three languages. That way, a method that can only deal with a certain language will be ranked very low. In addition, to evaluate the participants that also submitted a confidence score (a real number in the set  $[0, 1]$ ) we used Receiver-Operating Characteristic (ROC) curves and the area under the curve (AUC) as a single measure. ROC curves provide a more detailed picture over the ability of the author verification methods to assign high confidence scores to their answers. For the calculation of ROC curves, any missing answers were assumed to be wrong answers. Again, softwares that can only handle documents of a certain language will produce low AUC scores. Finally, since we asked for software submissions so that the software is executed at our site, it is possible for the first time to compare the runtime of the different author verification methods.

**Evaluation Results.** In total, 18 participants submitted their software for this task. The final evaluation results and the ranking of the participants according to the overall  $F_1$  score are depicted in Table 4 (left). Results for each of the three examined languages

<sup>6</sup> This part of the corpus was contributed by Patrick Brennan of Juola & Associates.

<sup>7</sup> <http://www.tovima.gr>

<sup>8</sup> Sheila Queralt of Universitat Pompeu Fabra and Angela Melendez of Duquesne University assisted in preparing this part of the corpus.

**Table 4.** Author identification results in terms of  $F_1$  and runtime (left table) as well as AUC for softwares that output confidence scores (right table)

Team	Overall	English	Greek	Spanish	Runtime	Team	Overall	English	Greek	Spanish
Seidman	0.753	0.800	0.833	0.600	1091.3 m	Jankowska	0.777	0.842	0.711	0.804
Halvani	0.718	0.700	0.633	0.840	0.1 m	Seidman	0.735	0.792	0.824	0.583
Layton	0.671	0.767	0.500	0.760	0.2 m	Ghaeini	0.729	0.837	0.527	0.926
Petmanson	0.671	0.667	0.567	0.800	603.6 m	Feng	0.697	0.750	0.580	0.772
Jankowska	0.659	0.733	0.600	0.640	4.0 m	Petmanson	0.651	0.672	0.513	0.788
Vilarino	0.659	0.733	0.667	0.560	93.0 m	Bobicev	0.642	0.585	0.667	0.654
Bobicev	0.655	0.644	0.712	0.600	28.6 m	Grozea	0.552	0.342	0.642	0.689
Feng	0.647	0.700	0.567	0.680	1406.9 m	baseline	0.500	0.500	0.500	0.500
Ledesma	0.612	0.467	0.667	0.720	0.5 m	Kern	0.426	0.384	0.502	0.372
Ghaeini	0.606	0.691	0.461	0.667	2.1 m	Layton	0.388	0.277	0.456	0.429
van Dam	0.600	0.600	0.467	0.760	0.2 m	Sorin	0.082	0.658	–	–
Moreau	0.600	0.767	0.433	0.600	130.0 m					
Jayapal	0.576	0.600	0.633	0.480	0.1 m					
Grozea	0.553	0.400	0.600	0.680	6.8 m					
Vartapetiance	0.541	0.500	0.533	0.600	7.0 m					
Kern	0.529	0.533	0.500	0.560	10.4 m					
baseline	0.500	0.500	0.500	0.500	–					
Veenman	0.417	0.800	–	–	16.0 m					
Sorin	0.331	0.633	–	–	60.7 m					

are provided as well. Moreover, 10 participants also submitted confidence scores together with their binary answers. This allowed us to compute ROC curves and the corresponding AUC values for those participants. The results of this evaluation are shown in Table 4 (right).

As concerns the features to represent the stylistic properties of texts, traditional solutions were followed including mainly character, lexical, and syntactic features. The latter require the use of language-specific NLP tools and considerably increase the runtime cost. The classification methods can be divided into intrinsic and extrinsic ones. Intrinsic methods make their decisions based solely on the set of known and unknown documents per problem. Conversely, extrinsic methods use external resources, such as additional documents of known authorship taken from the training corpus or downloaded from the web, and usually attempt to transform the one-class classification problem to a binary classification problem. The winning submission follows this approach and is based on the impostors method introduced in [27]. Ensemble classification models are very effective in both intrinsic and extrinsic approaches. Most participants attempt to tune the parameters of their systems separately for each language and sometimes they use external corpora in this procedure. Moreover, text length normalization seems to be a significant factor especially for producing a reliable confidence score for each provided answer.

## 4 Author Profiling

Author profiling is about predicting an author's demographics based on her writing. For example, profiling algorithms are used to determine an author's gender, age, native

**Table 5.** Corpus statistics of the evaluation corpus applied for author profiling

Lang	Age	Gender	No. of Authors		Lang	Age	Gender	No. of Authors	
			Training	Test				Training	Test
en	10s	male	8 600	888	es	10s	male	1 250	144
		female	8 600	888			female	1 250	144
	20s	male	(72) 42 828	(32) 4 576		20s	male	21 300	2 304
		female	(25) 42 875	(10) 4 598			female	21 300	2 304
	30s	male	(92) 66 708	(40) 7 184		30s	male	15 400	1 632
		female	66 800	7 224			female	15 400	1 632
$\Sigma$			236,600	25,440	$\Sigma$			75 900	8 160

language, personality type, etc. Author profiling is a problem of growing importance in a variety of areas, such as forensic linguistics and marketing. From the former perspective, the ability to determine the linguistic profile of the author of a suspicious text solely by analyzing the text is useful for suspect verification. Similarly, from a marketing perspective, companies are interested to know what types of people like or dislike their products, based on the analysis of blogs and online product reviews.

The starting point for our research is the seminal work of Argamon et al. [4], who were the first to demonstrate a correlation of word usage and author demographics. Until now, however, research within computational linguistics [3] and social psychology [39] has mainly focused on English text. In our lab, we therefore focus on predicting an author's gender and age based on both English and Spanish text. Moreover, we put particular emphasis on the use of everyday language and analyze how it reflects basic social and personality processes by using text obtained from social media.

**Evaluation Corpus.** To construct a large-scale evaluation corpus, we crawled public social media sites where user posts can be obtained along with labels that indicate author demographics such as gender and age. Table 5 shows the basic statistics of the compiled English and Spanish corpora. The corpora consist of files where each file contains at least one post and at most 1000 words of combined posts of an individual author. In case an author wrote posts amounting to more than that, more than one file for that author were generated. Authors with little data were kept in order to provide a realistic cross-section of authors within our evaluation framework.

The age labels are divided into age groups, following the approach of Schler et al. [51]: 10s (ages 13-17), 20s (ages 23-27) and 30s (ages 33-47). Within each age group, the subcorpora are balanced by gender; however, the subcorpora between age groups were left unbalanced. In addition, we introduced a small number of posts from sexual predators as well as posts with a sexual topic obtained from conversations between adults [23]. The numbers in parentheses found in the table denote the number of such conversations in the respective parts of our corpus.

**Evaluation Results.** In Table 6, the prediction accuracies for gender, age groups, and the combination are shown. Accuracies compute as ratio of the number of correctly predicted authors and total number of authors. To obtain the total score, we compute the average of the accuracies. The overall best performing approach across both languages

**Table 6.** Author profiling results in terms of accuracy on English (left) and Spanish (right) texts

English				Spanish			
Team	Total	Gender	Age	Team	Total	Gender	Age
Meina	0.3894	0.5921	0.6491	Santosh	0.4208	0.6473	0.6430
Pastor L.	0.3813	0.5690	0.6572	Pastor L.	0.4158	0.6299	0.6558
Seifeddine	0.3677	0.5816	0.5897	Cruz	0.3897	0.6165	0.6219
Santosh	0.3508	0.5652	0.6408	Flekova	0.3683	0.6103	0.5966
Yong Lim	0.3488	0.5671	0.6098	Ladra	0.3523	0.6138	0.5727
Ladra	0.3420	0.5608	0.6118	De-Arteaga	0.3145	0.5627	0.5429
Aleman	0.3292	0.5522	0.5923	Kern	0.3134	0.5706	0.5375
Gillam	0.3268	0.5410	0.6031	Yong Lim	0.3120	0.5468	0.5705
Kern	0.3115	0.5267	0.5690	Sapkota	0.2934	0.5116	0.5651
Cruz	0.3114	0.5456	0.5966	Pavan	0.2824	0.5000	0.5643
Pavan	0.2843	0.5000	0.6055	Jankowska	0.2592	0.5846	0.4276
Caurcel Diaz	0.2840	0.5000	0.5679	Meina	0.2549	0.5287	0.4930
H. Farias	0.2816	0.5671	0.5061	Gillam	0.2543	0.4784	0.5377
Jankowska	0.2814	0.5381	0.4738	Moreau	0.2539	0.4967	0.5049
Flekova	0.2785	0.5343	0.5287	Weren	0.2463	0.5362	0.4615
Weren	0.2564	0.5044	0.5099	Cagnina	0.2339	0.5516	0.4148
Sapkota	0.2471	0.4781	0.5415	Caurcel Diaz	0.2000	0.5000	0.4000
De-Arteaga	0.2450	0.4998	0.4885	H. Farias	0.1757	0.4982	0.3554
Moreau	0.2395	0.4941	0.4824	baseline	0.1650	0.5000	0.3333
baseline	0.1650	0.5000	0.3333	Aleman	0.1638	0.5526	0.2915
Gopal Patra	0.1574	0.5683	0.2895	Seifeddine	0.0287	0.5455	0.0512
Cagnina	0.0741	0.5040	0.1234	Gopal Patra	–	–	–

was provided by Lopez-Monroy et al. [33], computed as averaged between both English and Spanish.

With regard to the used features among the different approaches Lopez-Monroy et al. [33] employ second-order representations based on relationships between documents and author profiles, whereas Meina et al. [34] exploit collocations. The latter do not seem to perform as good in Spanish as they do in English, or they are more difficult to be tuned. Almost all approaches rely on writing style features. Nevertheless, a wide variety of performances were obtained, showing that they may not be very easy to handle. Part-of-speech features were employed by five different approaches, including the two best performing ones for English [34] and Spanish [50], whereas the remaining systems are ranked below the median rank. Readability features are also widely used: the approach of Gillam [16] uses them exclusively, which demonstrates the performance of such features in isolation. With the exception of Meina et al. [34]’s approach, all developers that employ n-gram features are also ranked below the median rank. Using sentiment words [37] and emotion words [13, 14] does not seem to improve accuracy in the same way as using slang words [1, 11, 13, 14]; however, these difference may be due to other features used by the same approaches. Finally we note that, with the exception of Lim et al. [32] and Meina et al. [34], all approaches that employ some kind of preprocessing on the corpora perform worse.

## 5 Handling Large-Scale Software Submissions with TIRA

This is the second time our lab accepts software submissions; in total, 58 softwares were submitted for the three tasks combined. This is more than five times as much compared to last year, where eleven softwares were submitted for the aforementioned plagiarism text alignment task [17]. Building on these experiences, we continue the development of the TIRA experimentation platform which serves as a valuable toolbox for organizing and managing our evaluation process.<sup>9</sup> In what follows, we outline the challenges of software submissions, discuss the technological and organizational means to meet them and how they are currently implemented within TIRA. Moreover, we present an analysis of user errors that provides insights into open problems and gives directions for future development.

### 5.1 Challenges of Software Submissions, and Our Solutions

In traditional evaluation labs, the lab organizers prepare and release a data set for a given task, withholding the ground truth data. Participants research and develop algorithms that solve the task and process the data at their site. Their algorithms' output (so-called runs) is submitted to the lab organizers who evaluate them against the ground truth. The only difference of our lab to the traditional process is that, instead of runs, we asked participants to submit their software in order for it to be run at our site and to be preserved in executable state for future evaluations. Accepting software submissions introduces a number of technical and organizational challenges, though. For each of these challenges, we devise tailored solutions:

1. *Environment Diversity.* With run submissions, participants are not limited with regard to their work environments (i.e., operating systems and programming languages). With software submission, lab organizers either need to restrict work environments or be prepared to execute arbitrary software.  
*Our solution:* virtualization; each participant gets full access to a virtual machine and deploys her software so that it can be executed via a pre-defined command.
2. *Executing Untrusted Software.* With software submissions, lab organizers are required to execute participant software at their site. The software often comes in the form of binaries instead of source code; in any case it is virtually impossible to ensure the trustworthiness of submitted software.  
*Our solution:* virtualization; virtual machines encapsulate submitted software.
3. *Data Leakage.* With software submissions, lab organizers may feed private data into the software. However, since the software is untrusted, this data may leak to the public via a number of channels that need to be monitored and secured by lab organizers.  
*Our solution:* sandboxing; before executing software, virtual machines are disconnected from the network, copied, and restored to their previous state afterwards.

---

<sup>9</sup><http://tira.webis.de>



4. *Error Handling.* With run submissions, participants debug their software directly. The only errors that may go unnoticed until after submission are errors in the run format specified by lab organizers. With software submissions, however, lab organizers may experience software errors because of insufficiently tested software or because of phenomena present in the test data that are absent from the training data.  
*Our solution:* unit testing; in case of errors, participants are notified by mail.
5. *Responsibility.* With software submissions, lab organizers assume partial responsibility for the successful evaluation of a participant's software. They must be vigilant about all kinds of errors that may invalidate the output of a submitted software.  
*Our solution:* staged submissions to encourage early bug fixing; TIRA's web front end organizes and visualizes the evaluation process.
6. *Execution Cost.* With run submissions, participants bear the costs of executing their software, since they have to bring their own hardware. With software submissions, lab organizers need to provide sufficient hardware or raise participation fees (e.g., for commercial cloud platforms). Raising fees, however, will hardly be accepted since participants typically already own perfectly suitable hardware.  
*Our solution:* we provide four servers, each hosting up to 20 virtual machines.

The next sub-section details our solutions.

## 5.2 TIRA's Evaluation Toolbox

The goal of TIRA is to automate the evaluation of information retrieval experiments [17, 19]. TIRA's main capability is to integrate an experiment software into a web service and to remote control its execution. It provides a web interface to do so with the click of a button and collects and indexes results and errors for later retrieval (e.g., to construct a leaderboard or to forward error messages to the participants). Building on this basic functionality, we address items 1 to 3 as well as parts of 4 and 5 of the above list of software submission challenges by integrating virtualization, sandboxed execution, unit testing, and staged submissions.

Arbitrary execution environments as well as executing untrusted software can be safely accomplished by virtualization. Upon registration, every participant is given access to a virtual machine running at our site. Access is provided via secure shell as well as virtual network computing (i.e., remote desktop), and administrative rights are provided. This way, participants are able to set themselves up, whereas our only restriction is that their software is executable from the command line and that it has parameters for an input and output directory. To allow for a variety of environments and programming languages, two operating systems are provided, namely Microsoft Windows 7 Enterprise, and Ubuntu Linux 12.04 LTS. Although offering other operating systems would not have been a problem, none were requested.

Although misuse is unexpected, participants still have full administrative control of their virtual machines, so that it is important to take every precaution to prevent confidential data from leaking. For example, running participant software on confidential data may cause it to remain present in the virtual machine after the run is complete (e.g., within temporary files, outputs, logs, or intentionally hidden copies).

Moreover, software running on the virtual machine may attempt to send copies of the data via network to an external host. To prevent such leaks, before running a software, it is moved into a so-called sandbox: (1) a snapshot of the current state of a virtual machine is taken, (2) all connections to external networks are cut, (3) the confidential test data is mounted into the virtual machine, (4) the software is run on the confidential data, (5) the output of the software is copied out of the virtual machine, and finally, (6) the virtual machine's state is reverted to that of its snapshot and all network connections are restored.

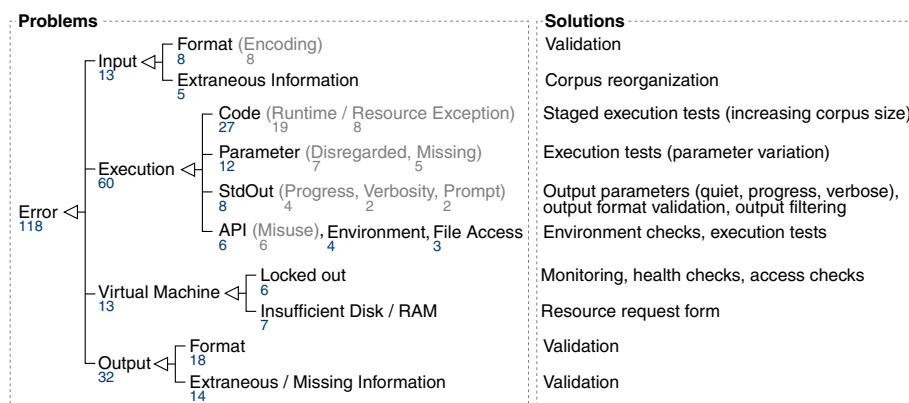
After the deployment of a software onto a virtual machine and after the participant confirms her submission, the virtual machine is moved into the sandbox. Before running the software on test data, we perform a small-scale unit test to ensure the software successfully executes. Then we run it on public data that is also accessible to participants for development purposes. The performances of this run are provided to participants so they can verify that their software behaved as expected during evaluation. Finally, the software is run on the test data and its output is checked for errors. After that, the virtual machine is moved out of the sandbox. In case of errors, participants are notified by mail and invited to re-submit a fixed version of their software.

Finally, from an organizational point of view, we found that staged submissions and engaging participants early to submit their software prototypes allows for early error correction and for getting an estimate of the final number of participants. To incentivize early software submissions, we offered an early bird submission deadline and the opportunity to get a pre-evaluation on a portion of the test data used for the final evaluation; 18 of the 46 participating teams took the opportunity to pre-evaluate their software.

### 5.3 Analysis of User Errors

Our current approach to error handling (item 4 of the above list) is based on a basic unit test that executes a submitted software on a very small sample of the evaluation corpus in order to learn whether it runs through. After that, the entire evaluation corpus is fed into the software. In case of errors, participants are notified by mail. In total, 1493 mails were exchanged within 392 conversations, discussing 118 errors. The number of teams experiencing at least one error is 39 from a total of 46, whereas 26 teams experienced at least two errors and one unlucky team 10. The identification of errors and the subsequent discussions induced a significant amount of manual workload. Sometimes, more than one round-trip was necessary to resolve an error. We analyzed the mails to get a better idea of what kinds of errors occurred and how they can be prevented in the future; Figure 1 organizes the errors into a taxonomy.

In general, input and output errors can be observed in traditional run submissions labs as well, whereas execution errors and virtual machine errors are exclusive to software submission labs. While the former can be easily identified or prevented by providing format validation and simplifying corpus organization, the latter require more intricate solutions or cannot be identified automatically at all. However, since half of all errors are execution errors, the work overhead for lab organizers to have them fixed can be minimized by allowing participants to perform execution tests themselves, for example, using TIRA's web front end. This way, turnaround times are minimized and no mails need be exchanged.



**Fig. 1.** Taxonomy of 118 problems that occurred during our lab along with technical solutions that identify them automatically. The numbers indicate the amount of errors within each category.

## 5.4 Evaluating Submitted Softwares across Years

One of the primary goals of doing software submissions in a lab is to make re-evaluations of the submitted softwares on different data sets possible. Since we are doing software submissions for the second time, this forms an excellent opportunity to demonstrate this possibility by cross-evaluating software from our previous lab on the current evaluation corpora and vice versa. This way, participating in one of our labs corresponds to participating in all of them past, present, and future. Moreover, if a participant submits versions of his software in different years, this will allow to track performance improvements. We evaluated the text alignment softwares submitted to the plagiarism detection task of last year as well as those submitted this year in this way and obtained combined rankings of both years. Discussing the results here is out of scope of this section, however, they can be found in [47].

## 6 Conclusion and Outlook

In conclusion, the creation of standardized evaluation resources for the digital text forensics tasks plagiarism detection, author identification, and author profiling forms the basis for renewed progress to solve these problems. In this regard, our annual lab has made significant headway. With the introduction of software submissions, we hope to go even further by compiling a repository of state-of-the-art implementations of algorithms for these tasks. The research community will benefit from conducting comparative experiments against their own algorithms as well as validating new evaluation corpora by feeding them into existing softwares. More generally, we hope our lab sets a new example of how to accomplish software submissions at large, and that the TIRA experimentation platform and the tools developed for it will be adopted by other researchers. Our future research into evaluation methodology is directed at making software submissions as simple as run submissions, and to further automate the organization of evaluation labs.

**Acknowledgements.** This work was partially supported by the WIQ-EI IRSES project (Grant No. 269180) within the FP7 Marie Curie action.

## References

- [1] Aleman, Y., Loya, N., Vilarino Ayala, D., Pinto, D.: Two Methodologies Applied to the Author Profiling Task—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [2] Argamon, S., Juola, P.: Overview of the International Authorship Identification Competition at PAN-2011. In: Proc. of CLEF 2011 (2011)
- [3] Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, Genre, and Writing Style in Formal Written Texts. *TEXT* 23, 321–346 (2003)
- [4] Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically Profiling the Author of an Anonymous Text. *Commun. ACM* 52(2), 119–123 (2009)
- [5] Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: EvaluatIR: An Online Tool for Evaluating and Comparing IR Systems. In: Proc. of SIGIR 2009 (2009)
- [6] Blockeel, H., Vanschoren, J.: Experiment Databases: Towards an Improved Experimental Methodology in Machine Learning. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenić, D., Skowron, A. (eds.) *PKDD 2007. LNCS (LNAI)*, vol. 4702, pp. 6–17. Springer, Heidelberg (2007)
- [7] Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating Gender on Twitter. In: Proc. EMNLP 2011 (2011)
- [8] Clough, P., Stevenson, M.: Developing a Corpus of Plagiarised Short Answers. *Lang. Resour. Eval.* 45, 5–24 (2011)
- [9] Clough, P., Gaizauskas, R., Piao, S.S.L., Wilks, Y.: METER: MEasuring TExt Reuse. In: Proc. ACL 2002 (2002)
- [10] De Roure, D., Goble, C., Stevens, R.: The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. *Future Gener. Comp. Sy.* 25, 561–567 (2009)
- [11] Caurcel Diaz, A.A., Gomez Hidalgo, J.M.: Experiments with SMS Translation and Stochastic Gradient Descent in Spanish Text Author Profiling—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [12] Downie, J.S.: The Music Information Retrieval Evaluation Exchange (2005–2007): A Window into Music Information Retrieval Research. *Acoust. Sc. and Tech.* 29(4), 247–255 (2008)
- [13] Hernandez Farias, D.I., Guzman-Cabrera, R., Reyes, A., Rocha, M.A.: Semantic-based Features for Author Profiling Identification: First Insights—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [14] Flekova, L., Gurevych, I.: Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [15] Forner, P., Navigli, R., Tufis, D. (eds.): *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers* (2013)
- [16] Gillam, L.: Readability for author profiling?—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [17] Gollub, T., Burrows, S., Stein, B.: First Experiences with TIRA for Reproducible Evaluation in Information Retrieval. In: Proc. of OSIR at SIGIR 2012 (August 2012)
- [18] Gollub, T., Stein, B., Burrows, S.: Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Proc. of SIGIR 2012 (2012)

- [19] Gollub, T., Stein, B., Burrows, S., Hoppe, D.: TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In: Proc. of TIR at DEXA 2012. IEEE (2012)
- [20] Goswami, S., Sarkar, S., Rustagi, M.: Stylometric Analysis of Bloggers' Age and Gender. In: Proc. of ICWSM 2009 (2009)
- [21] Haggag, O., El-Beltagy, S.: Plagiarism Candidate Retrieval Using Selective Query Formulation and Discriminative Query Scoring—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [22] Holmes, J., Meyerhoff, M.: *The Handbook of Language and Gender*. Blackwell Handbooks in Linguistics. Wiley (2003)
- [23] Inches, G., Crestani, F.: Overview of the International Sexual Predator Identification Competition at PAN-2012. In: Proc. of CLEF 2012 (2012)
- [24] Juola, P.: Authorship Attribution. *Found. and Trends in IR* 1, 234–334 (2008)
- [25] Juola, P.: Ad-hoc Authorship Attribution Competition. In: Proc. of ALLC 2004 (2004)
- [26] Juola, P.: An Overview of the Traditional Authorship Attribution Subtask. In: Proc. of CLEF 2012 (2012)
- [27] Koppel, M., Winter, Y.: Determining if Two Documents are by the Same Author. *Journal of the American Society for Information Science and Technology* (to appear)
- [28] Koppel, M., Argamon, S., Shimoni, A.R.: Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing* 17(4), 401–412 (2002)
- [29] Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research* 8, 1261–1276 (2007)
- [30] Koppel, M., Schler, J., Argamon, S.: Authorship Attribution in the Wild. *Language Resources and Evaluation* 45, 83–94 (2011)
- [31] Kong, L., Qi, H., Du, C., Wang, M., Han, Z.: Approaches for Source Retrieval and Text Alignment of Plagiarism Detection—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [32] Lim, W.Y., Goh, J., Thing, V.L.L.: Content-centric age and gender profiling—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [33] Pastor Lopez-Monroy, A., Montes-Y-Gomez, M., Jair Escalante, H., Villasenor-Pineda, L., Villatoro-Tello, E.: INAOE's participation at PAN' 13: Author Profiling task—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [34] Meina, M., Brodzinska, K., Celmer, B., Czokow, M., Patera, M., Pezacki, J., Wilk, M.: Ensemble-based Classification for Author Profiling using Various Features—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [35] Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: “How Old Do You Think I Am?”; A Study of Language and Age in Twitter. In: Proc. of ICWSM 2013 (2013)
- [36] Nguyen, D., Smith, N.A., Rosé, C.P.: Author Age Prediction from Text Using Linear Regression. In: Proc. of LaTeCH at ACL-HLT
- [37] Gopal Patra, B., Banerjee, S., Das, D., Saikh, T., Bandyopadhyay, S.: Automatic Author Profiling Based on Linguistic and Stylistic Features—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [38] Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting Age and Gender in Online Social Networks. In: Proc. of SMUC 2011 (2011)
- [39] Pennebaker, J.W.: *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury, USA (2013)
- [40] Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology* 54(1), 547–577 (2003)
- [41] Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Proc. of PAN at SEPLN 2009 (2009)
- [42] Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In: Proc. of CLEF 2010 (2010)

- [43] Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An Evaluation Framework for Plagiarism Detection. In: Proc. of COLING 2010 (2010)
- [44] Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd International Competition on Plagiarism Detection. In: Proc. of CLEF 2011 (2011)
- [45] Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., Stein, B.: Overview of the 4th International Competition on Plagiarism Detection. In: Proc. of CLEF 2012 (2012)
- [46] Potthast, M., Hagen, M., Stein, B., Graßegger, J., Michel, M., Tippmann, M., Welsch, C.: ChatNoir: A Search Engine for the ClueWeb09 Corpus. In: Proc. of SIGIR 2012 (2012)
- [47] Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.: Overview of the 5th International Competition on Plagiarism Detection. In: Proc. of CLEF 2013 (2013)
- [48] Potthast, M., Hagen, M., Völske, M., Stein, B.: Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: Proc. of ACL 2013. ACM (to appear, August 2013b)
- [49] Rodríguez Torrejón, D.A., Martín Ramos, J.M.: Text Alignment Module in CoReMo 2.1 Plagiarism Detector—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [50] Santosh, K., Bansal, R., Shekhar, M., Varma, V.: Author Profiling: Predicting Age and Gender from Blogs—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [51] Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of Age and Gender on Blogging. In: Proc. of CAAW 2006 (2006)
- [52] Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology* 60, 538–556 (2009)
- [53] Stamatatos, E.: Plagiarism Detection Using Stopword N-grams. *Journal of the American Society for Information Science and Technology* 62(12), 2512–2527 (2011)
- [54] Stein, B., Meyer zu Eißel, S., Potthast, M.: Strategies for Retrieving Plagiarized Documents. In: Proc. of SIGIR 2007 (2007)
- [55] Suchomel, Š., Kasprzak, J., Brandejs, M.: Diverse Queries and Feature Type Selection for Plagiarism Discovery—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [56] Williams, K., Chen, H., Chowdhury, S.R., Giles, C.L.: Unsupervised Ranking for Plagiarism Source Retrieval—Notebook for PAN at CLEF 2013. In: Forner, et al. (eds.) [15]
- [57] Wojnarski, M., Stawicki, S., Wojnarowski, P.: TunedIT.org: System for Automated Evaluation of Algorithms in Repeatable Experiments. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) *RSCTC 2010. LNCS*, vol. 6086, pp. 20–29. Springer, Heidelberg (2010)
- [58] Zhang, C., Zhang, P.: Predicting Gender from Blog Posts. Technical report, University of Massachusetts Amherst, USA (2010)

# QA4MRE 2011-2013: Overview of Question Answering for Machine Reading Evaluation

Anselmo Peñas<sup>1</sup>, Eduard Hovy<sup>2</sup>, Pamela Forner<sup>3</sup>, Álvaro Rodrigo<sup>1</sup>,  
Richard Sutcliffe<sup>4</sup>, and Roser Morante<sup>5</sup>

<sup>1</sup> NLP&IR Group, UNED, Spain  
{anselmo, alvarory}@lsi.uned.es

<sup>2</sup> Carnegie Mellon University, USA  
hovy@cmu.edu

<sup>3</sup> CELCT, Italy

forner@celct.it

<sup>4</sup> School of CSEE, University of Essex, UK  
rsutcl@essex.ac.uk

<sup>5</sup> CLiPS, University of Antwerp, Belgium  
roser.morante@ua.ac.be

**Abstract.** This paper describes the methodology for testing the performance of Machine Reading systems through Question Answering and Reading Comprehension Tests. This was the attempt of the QA4MRE challenge which was run as a Lab at CLEF 2011–2013. The traditional QA task was replaced by a new Machine Reading task, whose intention was to ask questions that required a deep knowledge of individual short texts and in which systems were required to choose one answer, by analysing the corresponding test document in conjunction with background text collections provided by the organization. Four different tasks have been organized during these years: Main Task, Processing Modality and Negation for Machine Reading, Machine Reading of Biomedical Texts about Alzheimer's disease, and Entrance Exams. This paper describes their motivation, their goals, their methodology for preparing the data sets, their background collections, their metrics used for the evaluation, and the lessons learned along these three years.

## 1 Introduction

The general goal of the Question Answering for Machine Reading Evaluation (QA4MRE) is to assess the ability of systems in two reading abilities: to answer questions about a text under reading, and to acquire knowledge from reading, especially the knowledge involved in the textual inferences that bridge the gap between texts, questions and answers.

The evaluation of these abilities can be approached in two principal different ways: the first one is to define a formal language (e.g., relational database), ask the systems to translate texts into the formal language representation (i.e., Information and

Relation Extraction), and then evaluate systems by using structured queries formulated in the formal language.

The second main approach is agnostic with regard to any particular representation: systems' input queries about the text are natural language questions. This is related to how Question Answering (QA) is being articulated during the last decade. In QA4MRE we follow this approach but with a significant change with respect to previous QA campaigns over unstructured text.

## 1.1 From QA to Reading Comprehension Tests

By 2005 we realized that there was an upper bound of 60% of accuracy in system performance, despite more than 80% of the questions being answered by at least one participant. We understood that we had a problem of error propagation in the traditional QA pipeline (Question Analysis, Retrieval, Answer Extraction, Answer Selection/Validation). Thus, in 2006 we proposed a task called Answer Validation Exercise (AVE) [6]. The aim was to produce a change in QA architectures to give more responsibility to the validation step. In AVE we assumed there was a previous step of hypothesis generation and the hard work had to be done in the validation step. This is a kind of classification task that could take advantage of Machine Learning. The same idea is behind the architecture of IBM's Watson (DeepQA project) that successfully participated at Jeopardy [1].

After the three editions of AVE we tried to transfer our conclusions to the main QA task at CLEF 2009 and 2010 [9]. The first step was to introduce the option of leaving questions unanswered. This is an easy way of testing systems' confidence: if a system is not sure about its answers, it can decide to let unanswered a question instead of risking giving an incorrect answer. This is related to the development of validation technologies. Then, we needed a measure able to reward systems that reduce the number of questions answered incorrectly without affecting system accuracy, by leaving unanswered the questions they estimated they couldn't answer. The measure was an extension of accuracy called  $c@1$  [5], tested during 2009 and 2010 QA campaigns at CLEF, and used also in the current evaluation.

However, this change wasn't enough. Almost all systems continued relying on IR engines to retrieve relevant passages and then trying to extract the exact answer from them. This is not the change in the architecture we expected, and again, results didn't go beyond the 60% pipeline upper bound. Finally, we understood that the change in the architecture requires to put more effort on the development of answer validation/selection technologies. For this reason, in the current formulation of the task, the step of retrieval is put aside for a while, focusing on the development of technologies able to work with a single document, and to answer questions about it.

In the new setting, we started again de-compounding the problem into hypothesis generation and validation. Thus, in QA4MRE we test systems only for the validation step. Together with the questions, the organization provides a set of candidate answers. This gives the evaluation the format of traditional Multiple Choice Reading Comprehension tests.



This development parallels the introduction in 2009 of the Machine Reading Program (MRP) by DARPA in North America. The goals of the program were to develop systems that perform deep reading of small numbers of texts in given domains and to answer questions about them. Analogously to QA4MRE, the MRP program involves batteries of questions for the evaluation of system understanding. However, testing queries were structured according to target ontologies, forcing participant teams to focus on the problem of document transformation into the formal representation defined by these target ontologies. Thus the Machine Reading challenge had to pass through the Information Extraction paradigm.

In QA4MRE we think it is important to leave the door open to find synergies with emerging research areas such as those related to Distributional Semantics, Knowledge Acquisition, and Ontology Induction. For this reason, we are agnostic with respect to the query language and the machine internal representation. Thus, questions and answers are posed in natural language.

## 1.2 Hypotheses, Research Questions and Specific Goals

Summing up, these are the hypotheses we make:

- Progress on Question Answering requires new architectures based on Hypothesis Generation and Answer Validation.
- There is a gap between texts, questions, and answers that requires, among other things, background knowledge and textual inference.
- Knowledge Bases of factual relations are not enough as sources of knowledge. Language interpretation requires other kinds of knowledge attached to language in different layers, from paraphrases to common sense general axioms.

Then, several research questions arise, including:

- What is the role of knowledge in bridging the gap between Texts, Questions, and Answers? To what extent can this knowledge be automatically derived from large text collections?
- What kind of synergies can be found between the use of relational knowledge bases, distributional semantics, and propositional semantics?
- Are systems able to consider extra-propositional aspects of meaning like modality and negation?
- How can one determine systems' levels of inference?
- What benchmarks best measure future progress in the field?
- How to evaluate systems ability to ensure that an answer is correct or even, to decide that there are no correct answers among candidates?

The evaluation campaigns aimed at giving, at least, partial answer to those questions by means of developing an evaluation methodology with 100% reusable benchmarks able to measure progress in the future (in several languages). Once this task is accomplished, the task now is to determine the current state of the art, and envisage next steps in the research agenda.

### 1.3 Roadmap

In 2011 we defined the following principles and roadmap:

1. Focus on validation: Questions have attached a set of candidate answers.
  - a. Step 1. All questions have one and only one correct candidate answer.
  - b. Step 2. Introduce questions that require inference.
  - c. Step 3. Introduce questions with no correct candidate answer.
  - d. Step 4. Introduce questions that require textual inference after reading a large set of documents related to the test.
2. Introduce hypothesis generation: Organization provides reference collections of documents related to the tests.
  - a. Step 5. Questions about a single document, but no candidate answers are provided.
  - b. Step 6. Full setting of QA where systems have to generate hypotheses considering the reference collection and provide the answer together with the set of documents that support the answer.

After three years, we have addressed most of the first phase (Steps 1–4), but the question now is if systems have achieved performance levels that ensure a qualitative difference if we try phase 2.

## 2 The Task

The QA4MRE task focuses on the reading of single documents and the identification of the answers to a set of questions. Questions are in the form of multiple choice, each having several options, and only one correct answer. The detection of correct answers might eventually require various kinds of inference and the consideration of previously acquired background knowledge from reference document collections. Although the additional knowledge obtained through the background collection may be used to assist with answering the questions, the principal answer is to be found among the facts contained in the test documents given. Thus, reading comprehension tests do not require only *semantic understanding* but they assume a *reasoning process* which involves using implications and presuppositions, retrieving the stored information, performing inferences to make implicit information explicit. Many different forms of *knowledge* take part in this process: linguistic, procedural, world-and-common-sense knowledge. All these forms coalesce during processing and it is sometimes difficult to clearly distinguish and reconstruct them in a system that needs additional knowledge and inference rules in order to understand the text and to give sensible answers.

By giving only a single document per test, systems are required to understand every statement and to form connections across statements in case the answer is spread over more than one sentence. Systems are requested to (i) understand the test questions, (ii) analyse the relation among entities contained in questions and entities expressed by the candidate answers, (iii) understand the information contained in the documents, (iv) extract useful pieces of knowledge from the background collections, (v) and select the correct answer from the five alternatives proposed.

From 2011 until 2013, four tasks have been organized in QA4MRE. These tasks are described in detail in the CLEF Online Working Notes.

## **2.1 Main Task**

The main task has been available in several languages (including Arabic, Bulgarian, English, German, Italian, Romanian, and Spanish). Test sets were divided into topics (AIDS, Climate Change, Music, Society and Alzheimer's disease). For each topic a background collection was provided, together with a set of testing documents for which questions were formulated, and candidate answers offered [7, 8, 10].

The resulting benchmark contains parallel tests into several languages (documents, questions and candidate answers are translations), and comparable documents as background reference collections.

Questions were made by task organizers to test a pre-selected set of question types and different levels of inference. In many cases, selecting the correct answer requires to gather previous information from the reference collection.

## **2.2 Machine Reading on Biomedical Texts about Alzheimer's disease**

This pilot task explored the ability of a system to answer questions using scientific language. The test posed questions in the Biomedical domain with a special focus on one disease, namely Alzheimer's. Texts were taken from PubMed Central related to Alzheimer's and from 66,222 Medline abstracts [4, 12].

Here, the specific domain enabled us to explore Machine Reading linked to controlled vocabularies, entity types, and a predefined set of relations among these entity types. Thus, the task aimed at finding contact points with approaches based on Information Extraction.

## **2.3 Japanese University Entrance Exams**

In all previous tasks, questions were posed by organizers with the aim of evaluating automatic systems under different reading abilities, types of questions, inference degree, etc. However, these questions were developed for the task and, thus, they can be arguably artificial.

In the challenge of "Entrance Exams", the goal is to test systems in a real scenario, like in a Turing test. Thus, systems were evaluated under the same conditions humans are evaluated to enter the University of Tokyo. For this purpose, some exercises about Reading Comprehension were extracted from actual exams [13].

This exercise was organized in coordination with the "Entrance Exams" task at NTCIR. Exams were created by the Japanese National Center for University Admissions Tests and the "Entrance Exams" corpus was provided by NII's Todai Robot Project and NTCIR.

## **2.4 Processing Modality and Negation for Machine Reading**

This task was aimed at evaluating whether systems were able to understand extra-propositional aspects of meaning like modality and negation [2, 3]. Modality is a

grammatical category that expresses aspects related to the attitude of the speaker towards his/her statements, including certainty, factuality, and evidentially. Negation is a grammatical category that allows changing the truth value of a proposition. Modality and negation interact to express extra-propositional aspects of meaning. This task exploited the same topics and background collections of the Main Task. However, test documents were specifically selected to ensure the properties required for the questions. Participating systems had to decide whether given events in the texts were Asserted, Negated, or Speculated. The task was offered in English only in 2011 and 2012. In 2013 we integrated modality and negation into the Main Task by including some questions that required this kind of processing in order to answer correctly.

### 3 The Background Collections

Human language text does not include all the information we want to transmit. This is because we omit information we know the reader will obtain from the context and their own language of the world. However, this fact represents a big issue for systems aimed at managing the knowledge contained in tests

Therefore, the use of Background Knowledge represents a very important element of the evaluation setting. Since no text is ever complete, the goal of reference/background collections is to contextualize the reading of a single document within its general topic, allowing systems to construct models of knowledge and inference as needed to overcome gaps, omissions, assumptions, and otherwise incomplete information in the given texts and questions. Such models can be constructed before the actual test or at run-time, at the discretion of the system.

We define *background knowledge* in terms of the relation between the testing questions (and answers) and the background collection. To determine the potential kinds of uses of prior knowledge, we distinguish at least four main types of background knowledge (although in fact it's a continuum):

1. Very specific facts related to the document under study. For example, the relevant relation between two concrete people involved in a specific event.
2. General facts not specific to any particular event. For example, geographical knowledge, main players in international affairs, movie stars, world wars. Also acronyms, transformations between quantities and measures, etc.
3. General abstractions that humans use to interpret language, to generate hypotheses or to fill missing or implicit information. For example, abstractions such as the result of observing the same event with different players (e.g., petroleum companies drill wells, quarterbacks throw passes, etc.)
4. Linguistic knowledge. For example, synonyms, hypernyms, transformations such as active/passive or nominalizations. Also transformations from words to numbers, meronymy, and metonymy.

Obviously this is not an exhaustive list. For example, we do not include ontological relations that enable temporal and spatial reasoning, or reasoning on quantities, which are also all relevant. Nonetheless, we believe that the collections allow systems to

extract, formalize, and apply during QA processing a lot of the kinds of information that people call ‘commonsense and world knowledge’.

It is important to develop a good methodology for building background collections for the evaluation task. Ideally, the background collection should cover completely the corresponding topic. This is feasible sometimes and unrealistic at others. For example, in the case of the pilot on Biomedical documents about Alzheimer's disease, a set of experts built a query (a set of conjunctions and disjunctions over 18 terms) that approximates very much the retrieval of all relevant documents (more than 66,000) without introducing much noise. However, this is not so easy in more open domains (e.g., Climate Change) or cases with non-specialized sources of information. In these cases, we crawl the web using, for each language and topic, a list of keywords and a list of sources. Keywords are translated into English and then translated into the other languages. Documents may be crawled from a variety of sources: newspapers, blogs, Wikipedia, journals, magazines, etc. The web sources are obviously language dependent, and each language also requires a list of possible web sites with documents related to the topic.

We realized after the first campaign that, since we organizers knew the test set, we used that information to select the keywords, and ensure the coverage of the questions. The effect is not only that background collections didn't cover completely the topic, but also that the collections have some bias with respect to the real distribution of concepts.

For this reason, the assumption that the ideal background collection should include all relevant documents for the topic (and only them) was made explicit, and as organizers we bear it in mind. Thus, we face the same problem as traditional Information Retrieval: we want all relevant documents (and only them), and we use queries (keywords) to retrieve them

The first strategy with the aim of ensuring the coverage of the topic as much as possible is to make the topic specific enough (e.g., AIDS medicaments rather than AIDS). The second strategy is to try to cover (at least partially) each of the possible principal ‘dimensions/aspects’ of that topic. How? First, by detecting a good central overview text, such as a Wikipedia article that defines the topic, ‘suggests’ its principal aspects, and provides links to additional good material. Then, organizers enumerate these dimensions and prepare a set of queries for each dimension. They document this process with three benefits: (i) to know what organizers and participants can expect or not from the collection; (ii) to give another dimension of re-usability; and (iii) to explore how Machine Reading will connect to Information Retrieval in the future.

## 4 Test Collections

The methodology developed for creating test collections translated into several languages consists of the following steps:

1. Four English documents are selected for each of the four topics (Aids, Alzheimer's, Climate Change, Music and Society). They are selected from

various sources and comprised the test documents against which questions were asked. Documents are chosen from copyright-free sources or by kind permission to the owners (as for example in 2013 with documents of the Editor in Chief, Editor and Oxford University Press).

2. In order to have a set of identical questions for the languages involved, test documents are translated by expert translators recruited from the Translation for Progress<sup>1</sup> platform for all languages.
3. To ensure that translations are faithful to the original document in both meaning and style and of good quality, all the documents are manually checked and corrected when necessary. We wanted to avoid a situation where portions of the original English text were left out of the translation in a particular target language, or perhaps modified or interpreted in a particular manner which would have made the question impossible to answer in that language.
4. Fifteen multiple-choice questions are then devised for each test document (the ‘Main’ questions). A question always had five candidate answers from which to choose, with one clearly correct answer and four clearly incorrect answers. The last edition included in all cases the fifth candidate answer “None of the above”, and six of the fifteen questions were composed so as to have no answer in the text. The correct response to each of these six questions was thus “None of the above”.
5. In addition to the fifteen Main questions, the 2013 edition included also one or more Auxiliary questions. Each Auxiliary question was a simplified version of an existing Main question. The format of these questions was identical to that of Main questions, i.e. a question followed by five multiple-choice answers. In most cases, the Auxiliary question required less inference to answer. The idea was that if a system was able to answer the Auxiliary question but not the corresponding Main question, the problem could be its ability to perform the missing inference.
6. Once the questions had been composed in the language of the original author, each was then translated into English. The English versions of the questions and candidate answers are carefully checked by a referee to verify that they are clear, that the intended answer is clearly correct, that the intended answer is in the test document, and that the other candidate answers are clearly incorrect. Questions are modified accordingly.

---

<sup>1</sup> <http://www.translationsforprogress.org/main.php> A Translation Exchange site linking volunteer translators (e.g., linguistics students or professionals in foreign languages interested in building experience as translators can link up with low-budget organizations who are in need of translation work, but without the budget to pay for it. There are currently over 1450 registered volunteer translator members (for 13 language combinations) and over 160 organization members. Translation for Progress database is open for viewing for the general public, but if you wish to post your profile or contact a volunteer translator, a registration is required.

7. The English versions are then used to translate each question into each of the languages of the task. The same process is used to translate each candidate answer (five per query).
8. The result of this process is a set with 240 Main questions and, in 2013, 44 Auxiliary questions in different languages, each with five multiple-choice answers. The final step is to check that the answer to each question was in fact present in the test document for all the languages of the task.

#### 4.1 Questions

Questions covered five different question types: purpose, method, causal, factoid, and which-is-true. Factoid questions were divided into the following sub-types: Location, Number, Person, List, Time and Unknown. Examples of the basic question types are given below. We took care to spread the question types evenly for a given test document, aiming for two questions per type. Example questions:

PURPOSE: What is the aim of protecting protein deposits in the brain?

METHOD: How can the impact of Arctic drillings be reduced?

CAUSAL: Name one reason why electronic dance music owes a debt to Kraftwerk.

FACTOID (number): What is the approximate number of TB patients?

WHICH-IS-TRUE: Which problem is similar in nature to global warming?

For all questions, the direct answer was contained in the test document; however answering the questions typically required some background knowledge and some form of inference. The required knowledge could be linguistic or could involve basic world knowledge. Linguistic knowledge concerns, for example, the ability to perform co-reference resolution or detect paraphrases on the lexical or syntactic level. World knowledge has to be inferred from the background collection. For instance, the text might mention *Barack Obama* while the question might refer to *the first African American President*. The fact that Barack Obama is the first African American President needs to be learnt from the background collection in order to be able to answer the question.

Typical types of world knowledge involve, for instance, knowledge about the basic referents in a text, e.g., being aware that *Yucca Mountain* is in Nevada. Another type of world knowledge involves knowledge of “life scripts” such as “visiting a restaurant”. Finally, the inference required can also be complex, involving several steps. For example, answering a question might require combining knowledge from the background collection with knowledge from the test document itself. For instance, the question “Who is the wife of the person who won the Nobel Peace Prize in 1992?” contains two facts P and Q, where P=“wife of Y=?” and Q=“winner of Nobel Peace Prize in 1992=Y”. The latter information can be gleaned from the background collection whereas the former is contained within the test document itself.

For each test document, we aimed for a combination of simple, medium, and difficult questions. At most six questions per document did not require knowledge from

the background collection. Two of these were simple questions, i.e., the answer and the fact questioned could be found in the same sentence in the test document. Four questions were of intermediate difficulty in that the answer and the fact questioned were not in the same sentence and could, in fact, be several sentences apart. Finally, the remaining four questions did require utilizing information from the background collection. While not all question types require inference based on the background collection, all of them required some form of textual and linguistic knowledge, such as the ability to detect paraphrases, as we made an effort to re-formulate questions in such a way that the answers could not be found by simple word overlap detection. For each question, we kept track of the inference required to answer it. This made it easier to ensure that that inference could in fact be drawn on the basis of the background collection, i.e., that the background collection did indeed contain the relevant fact. It also makes it possible to carry out further analyses regarding which questions or types of questions were difficult for the systems and why.

When creating the questions, we took care not to introduce any artificial patterns that would help finding the correct answer. Thus we ensured that all answer choices for a question were approximately the same length and consistent with respect to formulation and content, that all of the wrong answers were plausible, and that the placement of the correct answers was random and balanced.

## 5 Evaluation

One of goals of QA4MRE is to promote a change in QA architectures giving more importance to the validation step over the IR component in order to improve results. This is why we consider the possibility of leaving questions unanswered. The idea is that systems might reduce the amount of incorrect answers while keeping the proportion of correct ones, by leaving some questions unanswered.

Then, given a question with its corresponding candidate answers, a participant system can return two kinds of responses:

- An answer selected from the set of candidate ones for that question
- A *NoA* answer. This response should be given if the system considers it is not able to find enough evidences about the correctness of candidate answers and it prefers not to answer the question instead of giving an incorrect answer. Moreover, the system can return as a hypothetical answer the candidate one that it would have been selected, which allows to give some feedback about its validation performance.

The assessments of system's responses are given automatically by comparing them against the gold standard collection. Therefore, no manual assessment was required, which reduces the effort of the evaluation once the collections have been created and makes easier the future development of systems. Each system's response receives one and only one of the following three possible assessments:



- *Right* if the system has selected the correct answer among the set of candidate ones of the given question;
- *Wrong* if the system has selected one of the wrong answers;
- *NoA* if the system has decided not to answer the question. Where the system returned a hypothetical answer, this answer was assessed as *NoA\_R* in the case of it being correct or *NoA\_W* if it was wrong.

After previous years' experience, we realized that advancing the state of the art requires systems ability to decide whether all candidate answers were incorrect or not. In this way, systems able to take this decision should be rewarded over systems that just rank answers.

This is why we introduce in 2013 an explicit assessment focus on testing the ability to reject candidate answers when they are incorrect. We implemented this change by introducing in our tests a portion of questions (39%) where none of the options are correct and including a new last option in all questions: "None of the above answers is correct" (NCA).

It is important to remark that a *NoA* answer is different to a "None of the answers above is correct" (NCA) answer. The former means that the system does **not return any candidate answer** because it is not confident about giving the correct answer, while the latter means that the system rejects the other candidate answers **but returns a response** that will be assessed as *Right* or *Wrong*.

Participant systems were evaluated from two different perspectives:

1. A question-answering approach, as in the traditional QA evaluation campaigns, where we just evaluate the ability of systems answering a set of questions and rank systems according to the final value given by a measure.
2. A reading-test evaluation, obtaining figures for each particular reading test and topics. This perspective permits us to evaluate whether a system was able to understand a document and to what degree. More in detail, we evaluate if the system is able to pass each test, in a similar way to humans with RC tests, what requires obtaining more than 0.5 of  $c@1$ . This is a kind of evaluation studied with more detail in the pilot Entrance Exams task.

## 5.1 Evaluation Measure

$c@1$  has been the main evaluation in all the tasks celebrated in this Lab.  $c@1$  was firstly introduced in ResPubliQA 2009 [9] and is fully described in [5]. The formulation of  $c@1$  is given in Formula (1).

$$c@1 = \frac{1}{n} \left( n_R + n_U \frac{n_R}{n} \right) \quad (1)$$

where

$n_R$ : number of questions correctly answered.

$n_U$ : number of questions unanswered.

$n$ : total number of questions

The main feature of  $c@I$  is its consideration of unanswered questions.  $c@I$  acknowledges unanswered questions in the proportion that a system answers questions correctly, which is measured using the traditional *accuracy* (the proportion of questions correctly answered). Thus, a higher *accuracy* over answered questions, which might be associated to a better validation, would give more value to unanswered questions, and therefore, a higher final  $c@I$  value. By selecting this measure we wanted to encourage the development of systems able to check the correctness of their responses because NoA answers add value to the final value, while incorrect answers do not.

As a secondary measure, we also provided scores according to *accuracy* (see Formula (2)), the traditional measure applied to past QA evaluations at CLEF. We define *accuracy* considering both answered and unanswered questions.

$$accuracy = \frac{n_R + n_{UR}}{n} \quad (2)$$

where

$n_R$ : number of questions correctly answered.

$n_{UR}$ : number of unanswered questions whose candidate answer was correct.

$n$ : total number of questions

## 5.2 Question Answering Perspective Evaluation

The Question Answering perspective is focused on measuring systems' performance over a set of questions without considering the ability of a system to pass tests associated with documents. This is an approach similar to the one applied in QA@CLEF campaigns before 2011.

The information considered for each system at this level is:

- Total number of questions *ANSWERED*. This number is divided into:
  - total number of questions *ANSWERED* with a *RIGHT* answer,
  - total number of questions *ANSWERED* with a *WRONG* answer.
- Total number of questions *UNANSWERED* (a *NoA* response was given). This number is divided into:
  - total number of questions *UNANSWERED* with a *RIGHT* candidate answer,
  - total number of questions *UNANSWERED* with a *WRONG* candidate answer,
  - total number of questions *UNANSWERED* with an *EMPTY* candidate answer.

The following scores are calculated from this information:

- An overall  $c@I$  score over the whole collection (the set with 160 questions),
- A  $c@I$  score for each topic (40 questions for each topic),

- An overall *accuracy* score (over the 160 questions of the test collection, considering also the candidate answers given to unanswered questions as it has been explained above),
- The proportion of answers correctly discarded (see Formula (3)) in order to evaluate the validation performance.

$$correctly_{discarded} = \frac{n_{UW} + n_{UE}}{n_{UR} + n_{UW} + n_{UE}} \quad (3)$$

where:

$n_{UR}$ : number of unanswered questions whose candidate answer was correct

$n_{UW}$ : number of unanswered questions whose candidate answer was incorrect

$n_{UE}$ : number of unanswered questions whose candidate answer was empty

### 5.3 Reading Perspective Evaluation

The objective of the reading perspective evaluation is to offer information about the performance of a system “understanding” the meaning of each single document. This understanding is evaluated by means of the proposed multiple-choice tests. Each system has to pass a test about a given document similar to the evaluation of RC of new language learners, what was explored in more detailed in the Entrance Exams subtask.

The evaluation is performed taking as reference the  $c@1$  scores achieved for each test (one document with its ten questions). Then, these  $c@1$  scores can be aggregated at topic and global levels in order to obtain the following values:

- Median, average and standard deviation of  $c@1$  scores at test level, grouped by topic,
- Overall median, average and standard deviation of  $c@1$  values at test level.

The median  $c@1$  is provided under the consideration that it can be sometimes more informative at reading level than average values. This is because median is less affected by outliers than average, and therefore it provides more information about the ability of a system to understand a text.

We consider that a system passes a test according to this evaluation perspective if it achieves a score equal or higher than 0.5.

### 5.4 Random Baseline

This baseline randomly selects an answer from the set of candidate answers. Since there is one correct option among five, the overall result of this random baseline is 0.2 (both for *accuracy* and for  $c@1$ ). Systems applying a reasonable kind of processing and reasoning should be able to outperform this baseline.

## 5.5 NCA Baseline

The introduction of the “None of the above answers is correct” option in meaningful proportion, a 39% of questions, allows defining a baseline baseline for a dummy system that always returns this option. This baseline obtained a  $c@1$  of 0.39.

## 6 Lessons Learned

Reader will find the quantitative evaluation and results of all runs in all tasks inside the Working Notes Overview papers available on-line from CLEF site. Here we enumerate the main conclusions drawn from this experience.

If we look at the average results in the Main Task along the three years (Table 1), they are close to 0.25 (slightly above from random at 0.20). In general, individual systems select an incorrect answer over the correct one in most cases. There is one notable exception, a system able to give more correct answers than incorrect ones, achieving in each edition a value than 0.5 of  $c@1$ .

**Table 1.** Overview of results 2011-2013 Main Task

	2011	2012	2013
<b>Average <math>c@1</math></b>	0.21	0.26	0.24
<b>Best <math>c@1</math></b>	0.57	0.65	0.59
<b>Average % of unanswered questions</b>	38%	17%	9%

Table 1 shows also how the percentage of unanswered questions decreased in each edition, despite the fact that  $c@1$  values remain similar. This means that systems decision about answering a question or leaving it unanswered had little improvement. Therefore, it seems systems are increasing the risk of giving incorrect answers instead of focusing on developing better validation technologies, as it was expected with the proposal of this task. Possibly, the evaluation measure is not penalizing enough the increase in number of incorrect answers.

This reflection links with the main conclusion of Entrance Exams 2013. Entrance Exams is a very difficult scenario, even challenging for humans. Thus, we can learn from the strategies humans follow to select the correct answer. In most cases, the only way to determine the correct option is by discarding the rest of candidate answers. In other words, *there is more value on developing strategies to discard incorrect answers than strategies to select correct ones.*

Coming back to the Main Task 2013, the correct option was NCA (“none of the above is correct”) for 39% of questions. This baseline beats all systems except one, and would have been a good starting point to develop a strategy that decides more carefully on giving an answer only when there is evidence enough.

During the three years of the evaluation, the methodology received several refinements, trying to assess better the level of system performance in deep understanding. One key novelty was the introduction in 2013 of auxiliary questions, reformulating some main questions by reducing the need for inference. This innovation clearly

illustrated which types of reasoning systems were better or worse at. We discover systems find difficulties in questions requiring to connect facts as for example in “Who is the wife of the first president of X?” instead of “Who is the wife of Y?”

Another lesson learned is that most participants reduced the concept of answer validation simply to the task of answer ranking. For this purpose, they develop similarity based approaches that do not decide whether there is a correct answer or not among candidates. Generally, they simply trust the ranking score to exceed a given threshold. So, returning to the question of whether systems achieved enough performance to ensure that there will be a qualitative difference when trying full QA scenario, the answer is: possibly not.

Over the years, it has become clear that groups working on Question Answering are not making use of background knowledge collections very much. At most, systems might locate some possibly relevant material from the background collection through simple matching, and then use associated information to help rank the potential answers. Tying in with the point above on answer ranking above, it indicates the difficulty to introduce inference/reasoning into processing.

Regarding the construction of background collections, we learned it is very difficult to adequately define Background Knowledge, and to specify the types and sources that must be considered to solve the full QA scenario. There are increasingly more sources of linked / relational data that, potentially, can be used. However, language goes beyond a predefined set of relations among entities and values. That was the reason to propose the use of text collections inviting participants to acquire propositional knowledge useful for textual inferences. We have not obtained much of value in this regard.

Despite the difficulty on defining Background Knowledge, we have learned that if we want to use text collections to contextualize system readings, we must be very careful to not introduce any kind of bias. So, the idea of creating a background collection able to contextualize a single text can be formulated as a classical Information Retrieval task: retrieve all relevant documents and only them. Any methodological approach must take this ideal as reference and try to approximate it as much as possible.

We believe that the resources generated so far by QA4MRE will serve to measure progress in this direction, since they form a 100% reusable benchmark in several languages.

## 7 Related Work

Over the last years, the QA Track at CLEF has changed its evaluation methodology in order to promote deeper text understanding. Clearly, the task of retrieving just text excerpts (facts, sentences, paragraphs, or documents) is not enough to develop the technology. Besides QA, other evaluation activities were also performed which required deeper analyses of texts, for example Recognizing Textual Entailment (RTE), Answer Validation (AV), and Knowledge Base Population (KBP).

**Question Answering:** a system receives questions formulated in natural language and returns one or more exact answers to these questions, possibly with the locations from which the answers were drawn as justification. The evaluation of QA systems began at the Text Retrieval Conference (TREC) and was continued at the Cross Language Evaluation Forum (CLEF) in the EU and at the NII-NACSIS Test Collection for IR Systems (NTCIR) in Japan. Most of the questions used in these evaluations ask about facts (e.g., Who is the president of XYZ?) or definitions (e.g., What does XYZ mean?). Since systems could search for answers among several documents (using IR engines), it was generally possible to find in some document a ‘system-friendly’ statement that contained exactly the answer information stated in an easily matched form. This made QA both shallow and relatively easy.

**Recognizing of Textual Entailment (RTE):** a system must decide whether the meaning of a text (the Text T) entails the meaning of another text (the Hypothesis H): whether the meaning of the hypothesis can be inferred from the meaning of the text [14]. RTE systems have been evaluated at the RTE Challenges, whose first competition was proposed in 2005. The RTE Challenges encourage the development of systems that have to treat different semantic phenomena.

**Answer Validation Exercise (AVE)** [6, 15, 16]. A combination of QA and RTE evaluations, Answer Validation (AV) is the task of deciding, given a question and an answer from a QA system, whether the answer is correct or not. AVE was a task focused on the evaluation of AV systems and it was defined as a problem of RTE in order to promote a deeper analysis in QA.

Another application of RTE, similar to AVE, in the context of Information Extraction was performed in a pilot task at the RTE-6 with the aim of studying the impact of RTE systems in Knowledge Base Population (KBP). The objective of this pilot task is to validate the output of participant systems at the KBP slot-filling task that was included in the Text Analysis Conference (TAC). Systems participating at the KBP slot-filling task must extract from documents some values for a set of attributes of a certain entity. Given the output of participant systems at KBP, the RTE KBP validation pilot consists of deciding whether each of the values detected for an entity is correct according to the supporting document. For taking this decision, participant systems at the RTE KBP validation pilot receive a set of T-H pairs, where the hypothesis is built combining an entity, an attribute and a value.

Other efforts closer to our proposal for evaluating systems understanding took place as the ANLP/NAACL 2000 Workshop on Reading Comprehension Tests as evaluation for computer-based language understanding systems. This workshop proposed to evaluate understanding systems by means of Reading Comprehension (RC) tests. The evaluation consisted of a set of texts and a series of questions about each text. Quite interestingly, most of the approaches presented at that workshop showed how to adapt QA systems to such kind of evaluation.

A more complete evaluation methodology of MR systems has been reported in [11], where the authors also proposed to use RC tests. However, the objective of these tests was to extract correct answers from documents, which is similar to QA without an IR engine.

## 8 Conclusions

QA4MRE is characterised by two major innovations. First, there was a transition from traditional Question Answering based on shallow text analysis of large document collections, to a new focus involving deep analysis of individual documents. Over the years, the QA challenges adopted simple questions that required almost no inferences to find the correct answers. These surface-level evaluations promoted QA architectures based on Information Retrieval (IR) techniques, in which the final answers were obtained after focusing on selected portions of retrieved documents and matching sentence fragments or sentence parse trees. No real understanding of documents was achieved, since none was required by the evaluation. Machine Reading, on the other hand, requires the automatic understanding of texts at a deeper level, so this task encourages participants to build a different kind of system.

The second innovation of the task lay in the evaluation. Instead of manually inspecting answers to judge whether they were correct, evaluation was entirely automatic. This was made possible by adopting questionnaires comprising multiple-choice questions whose exact answers could be determined in advance. This strategy also enabled more complex types of question to be asked as well as posing fewer restrictions on the form of the answers.

This new evaluation was well received by the QA community. Significant lessons were learned from it.

**Acknowledgements.** Anselmo Peñas and Álvaro Rodrigo have been partially supported by the Research Network MA2VICMR (S2009/TIC-1542) and READERS project (CHIST-ERA). Eduard Hovy was supported by two DARPA grants in Machine Reading.

## References

1. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefter, N., Welty, C.: Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31(3) (2010)
2. Morante, R., Daelemans, W.: Annotating Modality and Negation for a Machine Reading Evaluation. *CLEF 2011 Labs and Workshop Notebook Papers* (2011)
3. Morante, R., Daelemans, W.: Annotating Modality and Negation for a Machine Reading Evaluation. *CLEF 2012 Evaluation Labs and Workshop Online Working Notes* (2012)
4. Morante, R., Krallinger, M., Valencia, A., Daelemans, W.: Machine Reading of Biomedical Texts about Alzheimer's Disease. *CLEF 2012 Evaluation Labs and Workshop Online Working Notes* (2012)
5. Peñas, A., Rodrigo, Á.: A Simple Measure to Assess Non-response. In: *Proceedings of 49th Annual Meeting of the Association for Computational Linguistics-Human Language Technologies (ACL-HLT 2011)*, Portland, Oregon, USA (2011)

6. Peñas, A., Rodrigo, Á., Sama, V., Verdejo, F.: Overview of the Answer Validation Exercise 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 257–264. Springer, Heidelberg (2007)
7. Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Forascu, C., Sporleder, C.: Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation. Working Notes, CLEF 2011 (2011)
8. Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P.: Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation. Working Notes, CLEF 2012 (2012)
9. Peñas, A., et al.: Overview of resPubliQA 2009: Question answering evaluation over european legislation. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 174–196. Springer, Heidelberg (2010)
10. Sutcliffe, R., Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Forascu, C., Benajiba, Y., Osenova, P.: Overview of QA4MRE Main Task at CLEF 2013. Working Notes, CLEF 2013 (2013)
11. Wellner, B., Ferro, L., Greiff, W., Hirschman, L.: Reading Comprehension Tests for Computer-based Understanding Evaluation. *Natural Language Engineering* 12(4), 305–334 (2006)
12. Morante, R., Krallinger, M., Valencia, A., Daelemans, W.: Machine Reading of Biomedical Texts about Alzheimer’s Disease 2013. Working Notes, CLEF 2013 (2013)
13. Peñas, A., Miyao, Y., Hovy, E., Forner, P., Kando, N.: Overview of QA4MRE 2013 Entrance Exams Task. Working Notes, CLEF 2013 (2013)
14. Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d’Alché-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 177–190. Springer, Heidelberg (2006)
15. Peñas, A., Rodrigo, Á., Verdejo, F.: Overview of the Answer Validation Exercise 2007. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 237–248. Springer, Heidelberg (2008)
16. Rodrigo, Á., Peñas, A., Verdejo, F.: Overview of the Answer Validation Exercise 2008. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 296–313. Springer, Heidelberg (2009)



# Multilingual Question Answering over Linked Data (QALD-3): Lab Overview

Philipp Cimiano<sup>1</sup>, Vanessa Lopez<sup>2</sup>, Christina Unger<sup>1</sup>, Elena Cabrio<sup>3</sup>,  
Axel-Cyrille Ngonga Ngomo<sup>4</sup>, and Sebastian Walter<sup>1</sup>

<sup>1</sup> CITEC, Universität Bielefeld, Germany  
{cimiano,cunger}@cit-ec.uni-bielefeld.de,  
swalter@techfak.uni-bielefeld.de

<sup>2</sup> IBM Research, Dublin, Ireland  
vanlopez@ie.ibm.com

<sup>3</sup> INRIA Sophia-Antipolis, France  
elena.cabrio@inria.fr

<sup>4</sup> Universität Leipzig, Germany  
ngonga@informatik.uni-leipzig.de

**Abstract.** The third edition of the open challenge on Question Answering over Linked Data (QALD-3) has been conducted as a half-day lab at CLEF 2013. Differently from previous editions of the challenge, has put a strong emphasis on multilinguality, offering two tasks: one on multilingual question answering and one on ontology lexicalization. While no submissions were received for the latter, the former attracted six teams who submitted their systems' results on the provided datasets. This paper provides an overview of QALD-3, discussing the approaches proposed by the participating systems as well as the obtained results.

## 1 Introduction

While more and more semantic data is published on the web, the question of how typical web users can access this body of knowledge becomes of crucial importance. Over the past years, there is a growing amount of research on interaction paradigms that allow end users to profit from the expressive power of Semantic Web standards while at the same time hiding their complexity behind an intuitive and easy-to-use interface; for an overview see [11]. Especially natural language interfaces have received wide attention, as they allow users to express arbitrarily complex information needs in an intuitive fashion and, at least in principle, in their own language. The key challenge lies in translating the users' information needs into a form such that they can be evaluated using standard Semantic Web query processing and inferencing techniques. To this end, systems have to deal with a heterogeneous, distributed and very large set of highly interconnected data. The availability of such an amount of open and structured data has no precedents in computer science and approaches that can deal with the specific character of linked data are urgently needed. In addition, multilinguality has become an issue of major interest for the Semantic Web community,

as both the number of actors creating and publishing data in languages other than English, as well as the amount of users that access this data and speak native languages other than English is growing substantially. In order to achieve the goal that users from all countries have access to the same information, there is an impending need for systems that can help in overcoming language barriers by facilitating multilingual access to semantic data originally produced for a different culture and language.

The main objective of the open challenges on *question answering over linked data*<sup>1</sup> (QALD) is to provide an up-to-date, demanding benchmark that establishes a standard against which question answering systems over structured data can be evaluated and compared. QALD-3 is the third instalment of the QALD open challenge, organized as a half-day lab at CLEF 2013.

The rest of the paper describes the previous editions of the challenge (Section 2), details the main novelties and the experimental setting of QALD-3 (Section 3) and the results obtained by the participating systems (Section 4). Section 5 then draws some conclusions about the current edition and proposes ideas for next editions of the challenge.

## 2 Previous QALD Challenges

The QALD challenges aim to bring together researchers and developers from different communities, including NLP, Semantic Web, human-computer interaction, and databases. The first edition, QALD-1, was organised in the context of the workshop *Question Answering Over Linked Data* at ESWC 2011. The second edition, QALD-2, was run in the context of the workshop *Interacting With Linked Data* at ESWC 2012 and broadened the scope to also include other paradigms for interacting with linked data as well as encourage communication across interaction paradigms.

In the context of QALD-1, two datasets were made available—DBpedia and an RDF export of the MusicBrainz database—together with a set of 50 training and 50 test questions each. These questions were created by a student assistant with no background in question answering in order to avoid a bias towards a particular approach. The questions were designed to present potential user questions and to include a wide range of challenges such as lexical ambiguities and complex syntactical structures. All training questions were annotated with corresponding SPARQL queries. For QALD-2, both question sets were combined to build a new training set, and a newly created test set was provided, leading to 100 training and 100 test questions for DBpedia, and 100 training and 50 test questions for MusicBrainz. In addition, a few out-of-scope questions were added to each question set, i.e., questions to which the datasets do not contain the answer, in order to test the ability of participating systems to judge whether a failure to provide an answer lies in the dataset or in the system itself. Further, we provided a small set of questions that could only be answered by combining information from both datasets, DBpedia and MusicBrainz, thus testing a system's ability

<sup>1</sup> <http://www.sc.cit-ec.uni-bielefeld.de/qald>

to combine several linked information sources when searching for an answer. All QALD-2 questions were additionally annotated with keywords in order to encourage keyword-based approaches to take part in the challenge.

For a detailed description of the challenge as well as the participating systems and their results, see [10].

### 3 QALD-3

Capitalizing on the positive feedback which QALD has received from the Semantic Web and NLP communities, in the third edition of the challenge we decided to make a step forward by introducing new elements. To this end, QALD-3 proposed two separate tasks: *multilingual question answering*, that keeps the basic structure of the previous challenges unchanged but introduces multilingualism as the major innovation, and *ontology lexicalization*, aimed at all methods that (semi-)automatically create lexicalizations of ontology concepts. In the following, we present more details about the proposed tasks and the resources we made available to the participants.

#### 3.1 Task 1: Multilingual Question Answering

Task 1 aims at all question answering systems that mediate between a user, expressing his or her information need in natural language, and semantic data. Given a RDF dataset and a natural language question or set of keywords in one of six languages (English, Spanish, German, Italian, French, Dutch), the participating systems had to return either the correct answers, or a SPARQL query that retrieves these answers. In order to evaluate and compare participating systems, three RDF datasets were provided:

- English DBpedia 3.8<sup>2</sup> (including links, most importantly to YAGO categories<sup>3</sup> and MusicBrainz<sup>4</sup>), a community effort to extract structured information from Wikipedia and to make this information available as RDF data
- Spanish DBpedia<sup>5</sup>, containing information from Wikipedia extracted in Spanish (containing almost 100 million RDF triples)
- MusicBrainz, a collaborative effort to create an open content music database. The dataset provided for the challenge is an RDF export containing all classes (artists, albums and tracks) and the most important properties of the MusicBrainz database

These datasets could either be downloaded or accessed through a provided SPARQL endpoint.

---

<sup>2</sup> <http://dbpedia.org>

<sup>3</sup> <http://www.mpi-inf.mpg.de/yago-naga/yago/>

<sup>4</sup> [musicbrainz.org](http://musicbrainz.org)

<sup>5</sup> <http://es.dbpedia.org>

To get acquainted with the datasets and possible questions, a set of 100 training questions for each dataset (i.e. English DBpedia, Spanish DBpedia and MusicBrainz) was provided. Later, systems were evaluated on 100 different test questions. Both training and test questions were mainly adopted from the QALD-2 challenge, slightly modified in order to account for changes in the DBpedia dataset and in order to include feedback obtained from participants of the first two challenges. As major innovation, all questions and keywords were translated into six different languages: English, Spanish, German, Italian, French, and Dutch. Here are some English example questions from the training sets:

– *DBpedia:*

- 5 How many monarchical countries are there in Europe?
- 58 Who produced the most films?
- 74 Which capitals in Europe were host cities of the summer Olympic games?
- 85 In which films did Julia Roberts as well as Richard Gere play?

– *Spanish DBpedia:*

- 2 Who was the son of Alfonso López Pumarejo married to?
- 4 In which city did Eva Perón die?
- 20 What is the area code of Barcelona?
- 40 How many films did Pedro Almodóvar produce?

– *MusicBrainz:*

- 2 Which groups was David Bowie a member of?
- 44 How many versions of the song Smells Like Teen Spirit are there?
- 79 Who did the vocals on the album Sabotage?
- 89 When were The Vertigos founded?

All training and test questions were manually annotated with keywords, corresponding SPARQL queries and with answers retrieved from the provided SPARQL endpoint. Annotations were provided in an XML format. Each of the questions specifies an ID for the question together with a range of other attributes explained below, the natural language string of the question in the six languages, keywords in the same languages, a corresponding SPARQL query, as well as the answers this query returns. Along with a unique ID, the following attributes were specified for each question:

- **answertype** gives the answer type, which can be one the following: **resource** (one or many resources, for which the URI is provided), **string** (a string value), **number** (a numerical value such as 47 or 1.8), **date** (a date provided in the format YYYY-MM-DD, e.g. 1983-11-02), **boolean** (either **true** or **false**)
- **aggregation** indicates whether any operations beyond triple pattern matching are required to answer the question (e.g., counting, filters, ordering)
- **onlydbo** is given only for DBpedia questions and reports whether the query relies solely on concepts from the DBpedia ontology

Here is an example from the DBpedia training set:

```

1 <question id="36" answertype="resource"
2           aggregation="false" onlydbo="false">
3
4 <string lang="en">
5 Through which countries does the Yenisei river flow?
6 </string>
7 <string lang="de">
8 Durch welche Länder fließt der Yenisei?
9 </string>
10 <string lang="es">
11 ¿Por qué países fluye el río Yenisei?
12 </string>
13 <string lang="it">
14 Attraverso quali stati scorre il fiume Yenisei?
15 </string>
16 <string lang="fr">
17 Quels sont les pays traversés par l'Ienisseï?
18 </string>
19 <string lang="nl">
20 Door welke landen stroomt de Jenisej?
21 </string>
22
23 <keywords lang=en>
24 Yenisei river, flow through, country
25 </keywords>
26 ...
27
28 <query>
29 PREFIX res: <http://dbpedia.org/resource/>
30 PREFIX dbp: <http://dbpedia.org/property/>
31 SELECT DISTINCT ?uri WHERE {
32   res:Yenisei River dbp:country ?uri .
33 }
34 </query>
35
36 <answers>
37 <answer>
38 <uri>http://dbpedia.org/resource/Mongolia</uri>
39 </answer>
40 <answer>
41 <uri>http://dbpedia.org/resource/Russia</uri>
42 </answer>
43 </answers>
44 </question>

```

As an additional challenge, some of the training and test questions were out of scope, i.e. they cannot be answered with respect to the dataset.

### 3.2 Task 2: Ontology Lexicalization

Multilingual information access can be facilitated by the availability of lexica in different languages, for example allowing for an easy mapping of Spanish, German, and French natural language expressions to English ontology labels. The task consisted in finding English lexicalizations of a set of classes and properties from the DBpedia ontology, for example in a Wikipedia corpus. The training data provided to the participating systems consisted of a set of 10 classes and 30 properties from the DBpedia ontology, as well as a lexicon containing lexicalizations of those classes and properties in *lemon*<sup>6</sup> format. Classes and properties were randomly chosen from the DBpedia ontology (properties with less than 20 entity pairs to properties with over 100,000 entity pairs). Here is an example of expected lexicalizations for the DBpedia class `TradeUnion`:

```

1 :TradeUnion a lemon:LexicalEntry ;
2   lemon:canonicalForm [ lemon:writtenRep "trade union"@en ]
3   ;
4   lemon:sense [ lemon:reference
5                 <http://dbpedia.org/ontology/TradeUnion> ]
6   ;
7   lexinfo:partOfSpeech lexinfo:noun .
8
9 :LaborUnion a lemon:LexicalEntry ;
10  lemon:canonicalForm [ lemon:writtenRep "labor union"@en ]
11  ;
12  lemon:sense [ lemon:reference
13                <http://dbpedia.org/ontology/TradeUnion> ] ;
14  lexinfo:partOfSpeech lexinfo:noun .

```

### 3.3 Evaluation Measures

The results submitted by participating systems were automatically compared to the gold standard results.

**Task 1.** For each question  $q$ , precision, recall and F-measure were computed as follows:

$$\begin{aligned}
 \text{Recall}(q) &= \frac{\text{number of correct system answers for } q}{\text{number of gold standard answers for } q} \\
 \text{Precision}(q) &= \frac{\text{number of correct system answers for } q}{\text{number of system answers for } q}
 \end{aligned}$$

$$\text{F-Measure}(q) = \frac{2 * \text{Precision}(q) \times \text{Recall}(q)}{\text{Precision}(q) + \text{Recall}(q)}$$

---

<sup>6</sup> <http://lemon-model.net>

On the basis of these measures, overall precision and recall values as well as an overall F-measure value were computed as the average mean of the precision, recall and F-measure values for all questions. In the results reported in Section 4 below, precision, recall and F-measure values refer to the averaged values.

**Task 2.** For each property, the uploaded lexical entries were evaluated automatically by comparing them to the manually created lexical entries along two dimensions: i) lexical precision, lexical recall and lexical F-measure, and ii) lexical accuracy. The first dimension evaluates how many of the gold standard entries for a property were submitted by the participants, and how many of the automatically generated entries are among the gold standard entries (precision), where two entries count as the same lexicalization if their lemma, part of speech and sense coincide. Thus lexical precision  $P_{lex}$  and recall  $R_{lex}$  for a property  $p$  are defined as follows:

$$P_{lex}(p) = \frac{|entries_{auto}(p) \cap entries_{gold}(p)|}{|entries_{auto}(p)|}$$

$$R_{lex}(p) = \frac{|entries_{auto}(p) \cap entries_{gold}(p)|}{|entries_{gold}(p)|}$$

where  $entries_{auto}(p)$  is the set of entries for the property  $p$  in the automatically constructed lexicon, while  $entries_{gold}(p)$  is the set of entries for the property  $p$  in the manually constructed gold lexicon. The F-measure  $F_{lex}(p)$  is then defined as the harmonic mean of  $P_{lex}(p)$  and  $R_{lex}(p)$ , as usual.

The second dimension, lexical accuracy, is necessary in order to evaluate whether the specified subcategorization frame and its arguments are correct, and whether these syntactic arguments have been mapped correctly to the semantic arguments (domain and range) of the property in question. The accuracy of an automatically generated lexical entry  $l_{auto}$  for a property  $p$  w.r.t. the corresponding gold standard entry  $l_{gold}$  is therefore defined as:

$$A_p(l_{auto}) = (frameEq(l_{auto}, l_{gold}) + \frac{|args(l_{auto}) \cap args(l_{gold})|}{|args(l_{gold})|} + \frac{\sum_{a \in args(l_{auto})} map(a)}{|args(l_{auto})|}) / 3$$

Where  $frameEq(l_1, l_2)$  is 1 if the subcategorization frame of  $l_1$  is the same as the subcategorization frame of  $l_2$ , and 0 otherwise, where  $args(l)$  returns the syntactic arguments of  $l$ 's frame, and where

$$map(a) = \begin{cases} 1, & \text{if } a \text{ in } l_{auto} \text{ has been mapped to the same semantic argument} \\ & \text{of } p \text{ as in } l_{gold} \\ 0, & \text{otherwise} \end{cases}$$

When comparing the argument mapping of the automatically generated entry with that of the gold standard entry, only the class of the argument is considered, i.e. *subject* or *object*. This abstracts from the specific type of subject (e.g. *copulative subject*) and object (e.g. *indirect object*, *prepositional object*, etc.) and therefore allows for an evaluation of the argument mappings independently of the correctness of the frame and frame arguments. The lexical accuracy  $A_{lex}(p)$

for a property  $p$  is then computed as the average mean of the accuracy values of each generated lexicalization. All measures are computed for each property and then averaged for all properties.

## 4 Participating Systems, Results and Discussion

Six teams participated in QALD-3, two groups more than in last year's challenge: five teams from Europe (three from France, one from Germany and one from Italy), and one from Asia (China). Participants were allowed to submit runs to one or both of the tasks. Six participants took part in multilingual question answering task, five participants on the DBpedia track only (and all of them on English questions only), and one participant on both DBpedia and MusicBrainz. No runs were submitted for the ontology lexicalization task.

### 4.1 Participating Systems

The participating systems follow different approaches to question answering over linked data. For question interpretation, some rely on linguistic strategies, e.g. the analysis of syntactic patterns, while others implement statistical approaches. In contrast to systems that take the provided natural language question as input, *squall2sparql* takes as input questions in SQUALL, a controlled natural language for English, and *Scalewelis* is based on faceted search instead of question interpretation. In the following, we give some details on the participating systems.

*Intui2* [3] is a prototype system for question answering over linked data that can answer natural language questions with respect to a given RDF dataset by analyzing the questions in terms of the syntactic constituents (*synfragments*) they are composed of. Syntactically, a *synfragment* corresponds to a subtree of the syntactic parse tree of the question, and semantically, it is a minimal span of text that can be interpreted as a concept URI, an RDF triple or a complex RDF query. These *synfragments* are then compositionally combined to an interpretation of the whole input question.

*SWIP* [14] relies on the use of query patterns to address the task of interpreting natural language queries. The query interpretation process consists of two main steps. First, the natural language question is translated into a pivot query, capturing the query focus, a dependency analysis and the extracted relations between substrings of the natural language question. Second, predefined query patterns are mapped to the pivot query, obtaining a list of potential interpretations of the user question, which are then ranked according to their estimated relevance and proposed to the user in form of reformulated natural language questions.

*CASIA* [8] implements a pipeline consisting of question analysis, resource mapping and SPARQL generation. More specifically, the system first transforms and represents natural language questions as a set of *query triples* of the form  $\langle \text{subject}, \text{predicate}, \text{object} \rangle$ , based on a shallow and deep linguistic analysis.



Second, it instantiates these query triples with corresponding resources from DBpedia, resulting in *ontology triples*. Third, based on the ontology triples and question type, SPARQL queries are constructed. Finally, the candidate queries are validated and ranked, and the best query is selected.

*squall2sparql* [5] is a translator from SQUALL, a controlled natural language for English, to SPARQL. Given a SQUALL sentence, the system first translates it into an intermediate logical representation based on Montague grammar. This intermediate representation is then translated into SPARQL by mapping logical constructs to combinations of SPARQL constructs.

*Scalewelis*<sup>7</sup> [7] is a faceted search system that guides the user through the search for an answer. Starting from an initial SPARQL query, facets are created for the first 1,000 results retrieved by that query, consisting of the classes the results belong to as well as properties that relate the results to other entities in the dataset. The user's selection of a facet is then used to refine the query until the answer is found.

The *RTV* system [6] integrates lexical semantic modelling and statistical inferences within a complex architecture that decomposes the natural language interpretation task into a cascade of three different stages: i) the selection of salient information from the question (i.e. predicate, arguments and properties), ii) the location of the salient information in the ontology through joint disambiguation of all candidates, and iii) the compilation of the final query against RDF triples. This architecture exploits a Hidden Markov Model (HMM) to select the proper ontological triples according to the graph nature of RDF. In particular, for each query an HMM model is produced whose Viterbi solution is the comprehensive joint disambiguation across the sentence elements.

## 4.2 Used External Resources and Tools

Table 1 shows the external resources and tools exploited by participating systems. Among the resources, Wikipedia and WordNet are used for semantic knowledge extraction (e.g. for calculating similarity among words in Intui2). Concerning external tools, text processing tools are used for questions preprocessing (i.e. Stanford CoreNLP, MaltParser and Chaos), while information retrieval tools such as Lucene are used to index Wikipedia versions in the RTV system, or to obtain string similarity scores in SWIP.

Two of the participating systems do not rely on linguistic resources at all: *squall2sparql*, where the use of controlled English as input language bypasses most of the problems related to language variability, and *Scalewelis*, which relies on faceted search rather than question interpretation.

## 4.3 Results

Tables 2 and Table 3 report on the results obtained by the participating systems over DBpedia and MusicBrainz datasets, respectively. The column *processed*

<sup>7</sup> <http://lisfs2008.irisa.fr/scalewelis/>

**Table 1.** External resources and tools used by the participating systems

<i>Resources</i>	CASIA	Intui2	SWIP	RTV
WordNet [4]	+	+	-	-
Wikipedia	-	+	-	+
PATTY [12]	+	-	-	-
<i>Tools</i>	CASIA	Intui2	SWIP	RTV
WS4J (WordNet Similarity for Java)	-	+	-	-
Chaos parser [2]	-	-	-	+
MaltParser [13]	-	-	+	-
Stanford CoreNLP [9]	+	+	-	-
Jena ARQ query engine	-	+	-	-
Lucene	-	-	-	+
LARQ (Lucene + ARQ)	-	-	+	-

states for how many of the questions the system provided an answer, *right* specifies how many of these questions were answered with an F-measure of 1, and *partially* specifies how many of the questions were answered with an F-measure strictly between 0 and 1. On the DBpedia dataset, the best F-measure was 0.9 and the lowest was 0.17, the average being 0.4. These results are comparable to the results achieved in earlier challenges, showing that the level of complexity of the questions is still very demanding.

**Table 2.** Results for DBpedia test set

<i>System</i>	<i>Total Processed</i>	<i>Right</i>	<i>Partially</i>	<i>Recall</i>	<i>Precision</i>	<i>F-measure</i>	
squall2sparql	99	99	80	13	0.88	0.93	0.90
CASIA	99	52	29	8	0.36	0.35	0.36
Scalewelis	99	70	32	1	0.33	0.33	0.33
RTV	99	55	30	4	0.34	0.32	0.33
Intui2	99	99	28	4	0.32	0.32	0.32
SWIP	99	21	15	2	0.16	0.17	0.17

**Table 3.** Results for MusicBrainz test set

<i>System</i>	<i>Total Processed</i>	<i>Right</i>	<i>Partially</i>	<i>Recall</i>	<i>Precision</i>	<i>F-measure</i>	
SWIP	50	33	24	2	0.51	0.51	0.51

The following questions on DBpedia were answered by all systems:

<i>ID</i>	<i>Question</i>
21	What is the capital of Canada?
22	Who is the governor of Wyoming?
30	What is the birth name of Angela Merkel?
68	How many employees does Google have?

And the following questions on DBpedia were answered by no systems:

---

*ID Question*

---

14 Give me all members of Prodigy.

16 Does the new Battlestar Galactica series have more episodes than the old one?

92 Show me all songs from Bruce Springsteen released between 1980 and 1990.

96 Give me all B-sides of the Ramones.

---

Of the questions in the test set, 45 queries require to search the answer using other namespaces than the DBpedia ontology (attribute `onlydbo=false`), such as YAGO or FOAF, and 19 queries require aggregation operations (attribute `aggregation=true`), such as comparisons, like in 16 above, superlatives, like in question 15 (What is the longest river?), or filtering, like in 92 above. It is especially in these queries that the systems perform poorly on.

## 5 Conclusion and Perspectives

QALD-3, the third edition of the QALD challenge, has attracted a higher number of participants than previous editions, showing that there is a growing interest among researchers to provide end users with an intuitive and easy-to-use access to the huge amount of data present on the Semantic Web—not only by means of classical question answering but also exploiting other paradigms such as faceted search. Although the main focus of the challenge has been on multilinguality, all participating systems worked on English data only. This shows that the multilingual scenario is not yet broadly addressed, although it is starting to attract attention (for a system that used translated QALD questions for evaluation see [1]). Further, the ontology lexicalization task was addressed only by one system during training phase (and one not participating in the challenge, see [15]) but by no participants during test phase. This hints at a slightly different integration of this task into the challenge, e.g. by providing lexica as additional resources for system\*s\* by inviting participants to share their own lexical resources.

In future challenges, we want to emphasize further aspects of question answering over linked data, such as the need to deal with a variety of interconnected datasets as well as hybrid sources of information (structured RDF data and unstructured text), while keeping the core task of multilingual question answering. Since the MusicBrainz dataset provided in all three QALD challenges was never used as much as DBpedia, we plan to move to a different domain that can arouse a broader interest. In particular, we think that the biomedical domain has the strong potential to attract new participants and to offer new challenges in the field of question answering over linked data.

## References

1. Aggarwal, N., Polajnar, T., Buitelaar, P.: Cross-lingual natural language querying over the web of data. In: Métais, E., Meziane, F., Sarace, M., Sugumaran, V., Vadera, S. (eds.) NLDB 2013. LNCS, vol. 7934, pp. 152–163. Springer, Heidelberg (2013)
2. Basili, R., Zanzotto, F.M.: Parsing engineering and empirical robustness. *Natural Language Engineering* 8, 2002 (2002)
3. Dima, C.: Intui2: A prototype system for question answering over linked data. In: Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF 2013. LNCS. Springer (to appear, 2013)
4. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
5. Ferré, S.: squal2sparql: A translator from controlled English to full SPARQL 1.1. In: Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF 2013. Lecture Notes in Computer Science. LNCS. Springer (to appear, 2013)
6. Giannone, C., Bellomaria, V., Basili, R.: A HMM-based approach to question answering against linked data. In: Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF 2013. LNCS. Springer (to appear, 2013)
7. Guyonvarch, J., Ferré, S.: Scalewelis: A query-based faceted search system on top of SPARQL endpoints. In: Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF 2013. LNCS. Springer (to appear, 2013)
8. He, S., Liu, S., Chen, Y., Zhou, G., Liu, K., Zhao, J.: Casia@QALD-3: A question answering system over linked data. In: Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF 2013. LNCS. Springer (to appear, 2013)
9. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, ACL 2003, vol. 1, pp. 423–430. Association for Computational Linguistics, Stroudsburg (2003)
10. Lopez, V., Unger, C., Cimiano, P., Motta, E.: Evaluation Question Answering Over Linked Data. *Journal of Web Semantics* (in press)
11. Lopez, V., Uren, V.S., Sabou, M., Motta, E.: Is question answering fit for the semantic web?: A survey. *Semantic Web* 2(2), 125–155 (2011)
12. Nakashole, N., Weikum, G., Suchanek, F.: Patty: A taxonomy of relational patterns with semantic types. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, pp. 1135–1145. Association for Computational Linguistics, Stroudsburg (2012)
13. Nivre, J., Hall, J., Nilsson, J.: Maltparser: A data-driven parser-generator for dependency parsing. In: Proc. of LREC 2006, pp. 2216–2219 (2006)
14. Pradel, C., Peyet, G., Haemmerlé, O., Hernandez, N.: Swip at qald-3: results, criticisms and lesson learned. In: Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF 2013. LNCS. Springer (to appear, 2013)
15. Walter, S., Unger, C., Cimiano, P.: A corpus-based approach for the induction of ontology lexica. In: Métais, E., Meziane, F., Sarace, M., Sugumaran, V., Vadera, S. (eds.) NLDB 2013. LNCS, vol. 7934, pp. 102–113. Springer, Heidelberg (2013)

# Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems\*

Enrique Amigó<sup>1</sup>, Jorge Carrillo de Albornoz<sup>1</sup>, Irina Chugur<sup>1</sup>, Adolfo Corujo<sup>2</sup>,  
Julio Gonzalo<sup>1</sup>, Tamara Martín<sup>1</sup>, Edgar Meij<sup>3</sup>,  
Maarten de Rijke<sup>4</sup>, and Damiano Spina<sup>1</sup>

<sup>1</sup> UNED NLP and IR Group  
Juan del Rosal, 16. 28040 Madrid, Spain  
<http://nlp.uned.es>

<sup>2</sup> Lorente and Cuenca  
Lagasca, 88. 28001 Madrid, Spain  
<http://llorenteycuenca.com>

<sup>3</sup> Yahoo! Research  
Diagonal 177, 08018 Barcelona, Spain  
<http://research.yahoo.com/>

<sup>4</sup> ISLA, University of Amsterdam  
Science Park 904, 1098 XH Amsterdam  
<http://isla.science.uva.nl>

**Abstract.** This paper summarizes the goals, organization, and results of the second RepLab competitive evaluation campaign for Online Reputation Management Systems (RepLab 2013). RepLab focused on the process of monitoring the reputation of companies and individuals, and asked participant systems to annotate different types of information on tweets containing the names of several companies: first tweets had to be classified as related or unrelated to the entity; relevant tweets had to be classified according to their polarity for reputation (Does the content of the tweet have positive or negative implications for the reputation of the entity?), clustered in coherent topics, and clusters had to be ranked according to their priority (potential reputation problems had to come first). The gold standard consists of more than 140,000 tweets annotated by a group of trained annotators supervised and monitored by reputation experts.

**Keywords:** RepLab, Reputation Management, Evaluation Methodologies and Metrics, Test Collections, Text Clustering, Sentiment Analysis.

---

\* This research was partially supported by the European Community's FP7 Programme under grant agreement n 288024 (LiMoSINe), ESF grant ELIAS, the Spanish Ministry of Education (FPU grant AP2009-0507 and FPI grant BES-2011-044328), the Spanish Ministry of Science and Innovation (Holopedia Project, TIN2010-21128-C02), and the Regional Government of Madrid under MA2VICMR (S2009/TIC-1542).

## 1 Introduction

In a world of online networked information, where its control has moved to users and consumers, every move of a company and every act of a public figure are subject, at all times, to the scrutiny of a powerful global audience. While traditional reputation analysis is mostly manual, online media allow to process, understand and aggregate large streams of facts and opinions about a company or individual. In this context, Natural Language Processing plays a key, enabling role and we are already witnessing an unprecedented demand for text mining software for ORM. Although opinion mining has made significant advances in the last few years, most of the work has been focused on products. However, mining and interpreting opinions about companies and individuals is, in general, a much harder and less understood problem, since unlike products or services, opinions about people and organizations cannot be structured around any fixed set of features or aspects, requiring a more complex modeling of these entities.

RepLab is an initiative promoted by the EU project LiMoSINE<sup>1</sup> which aims at enabling research on reputation management as a “living lab”: a series of evaluation campaigns in which task design and evaluation are jointly carried out by researchers and the target user communities (reputation management experts). RepLab 2013, as its first edition in 2012 [2], has been organized as a CLEF lab, and the results of the exercise are discussed at CLEF 2013 in Valencia, Spain, on 23-26th September.

RepLab 2013 has been focused on the task of monitoring the reputation of entities (companies, organizations, celebrities, etc.) on Twitter. The monitoring task for analysts consists of searching the stream of tweets for potential mentions to the entity, filtering those that do refer to the entity, detecting topics (i.e., clustering tweets by subject) and ranking them based on the degree to which they are potential reputation alerts (i.e., issues that may have a substantial impact on the reputation of the entity, and must be handled by reputation management experts).

## 2 Tasks

### 2.1 Task Definition

The RepLab 2013 task is defined, accordingly, as (multilingual) topic detection combined with priority ranking of the topics, as input for reputation monitoring experts. The detection of polarity for reputation (does the tweet have negative/positive implications for the reputation of the entity?) is an essential step to assign priority, and is evaluated as a standalone subtask.

Participants were welcome to present systems that attempt the full monitoring task (filtering + topic detection + topic ranking) or modules that contribute only partially to solve the problem. Subtasks that are explicitly considered in RepLab 2013 are:

---

<sup>1</sup> <http://www.limosine-project.eu>

- Filtering. Systems are asked to determine which tweets are related to the entity and which are not. For instance, distinguishing between tweets that contain the word "Stanford" referring to the University of Stanford and filtering out tweets about Stanford as a place. Manual annotations are provided with two possible values: related/unrelated.
- Polarity for Reputation classification. The goal is to decide if the tweet content has positive or negative implications for the company's reputation. Manual annotations are: positive/negative/neutral.
- Topic Detection: Systems are asked to cluster related tweets about the entity by topic with the objective of grouping together tweets referring to the same subject/event/conversation.
- Priority assignment. The full task involves detecting the relative priority of topics. So as to be able to evaluate priority independently from the clustering task, we will evaluate the subtask of predicting the priority of the cluster a tweet belongs to.

A substantial difference between RepLab 2013 and its first edition in 2012 is that, in 2013, the training and test entities are the same, and therefore conventional Machine Learning techniques are readily applicable. RepLab 2013 models a scenario where reputation experts are constantly tracking and annotating information about a client (entity), and therefore it is likely to have manual annotations for data related to the entity of interest. RepLab 2012, on the other hand, modeled the scenario of a web application that can be used by anyone, at any time, using any entity name as keyword. In that case, training material was referred to entities other than those in the training set.

In RepLab 2013 it was possible to present systems that address only filtering, only polarity identification, only topic detection or only priority assignment. Another difference with 2012 is that in its second edition, the RepLab organization provided baseline components for all of the four subtasks. This way any participant was able to participate in the full task regardless of where his particular contribution lied.

Some relevant details on the polarity for reputation and topic detection tasks follow. *Polarity for reputation* is substantially different from standard sentiment analysis: First, when analyzing polarity for reputation, both facts and opinions have to be considered. For instance, "Barclays plans additional job cuts in the next two years" is a fact with negative implications for reputation. Therefore, systems will not be explicitly asked to classify tweets as factual vs. opinionated: the goal is to find polarity for reputation, that is what implications a piece of information might have on the reputation of a given entity, regardless of whether the content is opinionated or not. Second, negative sentiments do not always imply negative polarity for reputation and vice versa. For instance, "R.I.P. Michael Jackson. We'll miss you" has a negative associated sentiment (sadness, deep sorrow), but a positive implication for the reputation of Michael Jackson. And the other way around, a tweet such as "I LIKE IT..... NEXT...MITT ROMNEY...Man sentenced for hiding millions in Swiss bank account", has a positive

sentiment (joy about a sentence) but has a negative implication for the reputation of Mitt Romney.

As for the *topic detection + topic ranking* process, a three-valued classification was applied to assess the priority of each entity-related topic: alert (the topic deserves immediate attention of reputation managers), mildly relevant (the topic contributes to the reputation of the entity but does not require immediate attention) and unimportant (the topic can be neglected from a reputation management perspective). Some of the factors that play a role in the priority assessments are:

- Polarity. Topics with polarity (and, in particular, with negative polarity, where action is needed) usually have more priority.
- Centrality. A high priority topic is very likely to have the company as the main focus of the content.
- User’s authority. A topic promoted by an influential (for example, in terms of the number of followers or the expertise) user has better chances of receiving high priority.

## 2.2 Baselines

The baseline approach consists of tagging tweets (in the test set) with the same tags of the closer tweet in the (entity) training set according to the Jaccard word distance. The baseline is, therefore, a simple version of Memory-Based learning. We have selected this approach for several reasons: (i) It is easy to understand; (ii) It can be applied to every subtask in RepLab 2013; (iii) It keeps the coherence between tasks: if a tweet is annotated as non-related, it will not receive any priority or topic tag. (iv) it exploits the training data set per entity.

## 2.3 Evaluation Measures

All subtasks consist of tagging single tweets according to their relatedness, priority, polarity or topic. However, each one corresponds with a particular artificial intelligence problem: binary classification (relatedness), three-level classification (polarity and priority), clustering (topic detection), and their concatenation (full task). A common feature for all tasks is that the classes, levels or clusters can be unbalanced. This entails challenges for the evaluation methodology definition. First, in classification tasks, a non informative system (i.e. all tweets to the same class) can achieve high scores without providing useful information. Second, in three-level classification tasks, a system could sort tweets correctly without a perfect correspondence between predicted and true tags. Third, an unbalanced cluster distribution across entities produces an important trade-off between precision/recall oriented evaluation metrics (precision or cluster entropy versus recall or class entropy) and that makes the measure combination function crucial for system ranking.

In evaluation, there is a hidden trade-off between interpretability and strictness. For instance, the Accuracy measure is easy to interpret: it simply reports



how frequently the system makes the correct decision. However, it is also easy to be cheated under unbalanced test sets. For instance, returning all tweets in the same class, cluster or level, may have high accuracy if the set is unbalanced. Other measures based on information theory are more strict when penalizing non informative outputs, but at the cost of interpretability. In this evaluation campaign we employ Accuracy as a high interpretable measure, and the combination of Reliability and Sensitivity (R&S) as a strict and theory grounded measure [4].

Basically, R&S assume that any organization task consists of a bag of relationships between documents. In our tasks, two documents are related if they have different priority, polarity or relatedness level, or when they appear in the same cluster. In brief, R&S computes the precision and recall of relationships produced by the systems with respect to the goldstandard. In order to avoid the quadratic effect of document pairwise, R&S is computed for each document relationships and averaged in a second step. Reliability and Sensitivity are computed as, being  $\mathcal{I}$  the set of tweets considered in the evaluation:

$$R(system) = Avg_{i \in \mathcal{I}} R(i) \quad S(system) = Avg_{i \in \mathcal{I}} S(i)$$

$$R(i) = P_{j \in \mathcal{I}} (rel_{gold}(i, j) = rel_{sys}(i, j) | rel_{sys}(i, j))$$

$$S(i) = P_{j \in \mathcal{I}} (rel_{gold}(i, j) = rel_{sys}(i, j) | rel_{gold}(i, j))$$

where  $rel_{gold}(i, j)$  represents that  $i$  has a higher or lower polarity, priority or relatedness than  $j$ , or that  $i$  and  $j$  belong to the same cluster.  $rel_{sys}(i, j)$  is analogous but applied to the system output.

R&S have three main strengths. First, they can be applied to ranking, filtering, organization by levels and grouping tasks. This matches all the RepLab 2013 tasks. In addition, it gives the possibility to evaluate the full task as a whole. Second, it covers simultaneously the desirable formal properties satisfied by other measures in each particular task [4]. Third, according to experimental results that we corroborate with RepLab 2013 data, R&S are strict with respect to other measures: a high score according to R&S ensures a high score according to any traditional measure. In other words, a low score according to one particular traditional measure produces a low R&S score, even when the system is rewarded by other measures.

R and S are combined with the F measure, i.e. a weighted harmonic mean of R and S. This combining function is grounded on the measure theory, and satisfies a set of desirable constraints. One of the most useful is that a low score according to any of both measures strongly penalizes the combined score. However, specially in clustering tasks, the F measure is seriously affected by the relative weight of partial measures (the  $\alpha$  parameter). In order to solve this we complement the evaluation results with the Unanimous Improvement Ratio, which has been proved to be the only weighting independent combining criterion [3]. UIR is computed over the test cases (entities in RepLab) in which all measures corroborates a difference between runs. Being  $S_1$  and  $S_2$  two runs and

$N_{>\forall}(S_1, S_2)$  the amount of test cases for which  $S_1$  improves  $S_2$  for all measures:

$$UIR(S_1, S_2) = \frac{N_{>\forall}(S_1, S_2) - N_{>\forall}(S_2, S_1)}{\text{Amount of cases}}$$

### 3 Dataset

RepLab 2013 uses Twitter data in English and Spanish. The balance between both languages depends on the availability of data for each of the entities included in the dataset. The collection comprises tweets about 61 entities from four domains: automotive, banking, universities and music. The domain selection was done to offer a variety of scenarios for reputation studies. To this aim we included entities whose reputation largely relies on their products (automotive), entities for which transparency and ethical side of their activity are the most decisive reputation factors (banking), entities for which the reputation of which depends on a very broad and intangible set of products (universities) and, finally, entities where the reputation is based almost equally on their products and personal qualities (music bands and artists). Table 1 summarizes the description of the corpus, as well as the number of tweets for both training and test sets, and the distribution by language.

Crawling was performed from 1 June, 2012 until 31 Dec, 2012 using each entity’s canonical name as query. For each entity, at least 2,200 tweets were collected: the first 700 were reserved for the training set and the last 1,500 for the test collection. This distribution was set in this way to obtain a temporal separation (ideally of several months) between the training and test data. The corpus also comprises additional background tweets for each entity (up to 50,000, with a large variability across entities). These are the remaining tweets situated between the training (earlier tweets) and test material (the latest tweets) in the timeline.

**Table 1.** RepLab 2013 dataset

	All	Automotive	Banking	Universities	Music/Artist
Entities	61	20	11	10	20
Training No. Tweets	45,679	15,123	7,774	6,960	15,822
Test No. Tweets	96,848	31,785	16,621	14,944	33,498
Total No. Tweets	142,527	46,908	24,395	21,904	49,320
No. Tweets EN	113,544	38,614	16,305	20,342	38,283
No. Tweets ES	28,983	8,294	8,090	1,562	11,037

These data sets were manually labelled by thirteen annotators who were trained, guided and constantly monitored by experts in ORM. Each tweet is annotated as follows:

- RELATED/UNRELATED: the tweet is/is not about the entity.

**Table 2.** RepLab 2013 dataset for the Filtering Task

	<b>All Automotive Banking Universities Music/Artist</b>				
Training No. Related	34,882	11,356	5,753	3,412	14,361
Training No. Unrelated	10,797	3,767	2,021	3,548	1,461
Test No. Related	75,470	24,415	12,053	7,715	31,287
Test No. Unrelated	21,378	7,370	4,568	7,229	2,211
Total No. Related	110,352	35,771	17,806	11,127	45,648
Total No. Unrelated	32,175	11,137	6,589	10,777	3,672

**Table 3.** RepLab 2013 dataset for the Polarity Task

	<b>All Automotive Banking Universities Music/Artist</b>				
Training No. Positive	19,718	5,749	2,195	2,286	9,488
Training No. Neutral	9,753	4,616	767	894	3,476
Training No. Negative	5,409	991	2,791	232	1,395
Test No. Positive	43,724	24,415	12,053	7,715	31,287
Test No. Neutral	20,740	9,512	1,407	2,443	7,378
Test No. Negative	11,006	2,101	4,994	820	3,091
Total No. Positive	63,442	12,802	5,652	4,452	20,818
Total No. Neutral	30,493	14,128	2,174	3,337	10,854
Total No. Negative	16,415	3,092	7,785	1,052	4,486

- POSITIVE/NEUTRAL/NEGATIVE: the information contained in the tweet has positive, neutral or negative implications for the entity’s reputation.
- Identifier of the topic cluster the tweet has been assigned to.
- ALERT/MILDLY IMPORTANT/UNIMPORTANT: the priority of the topic cluster the tweet belongs to.

Table 2 shows statistics about the filtering subtask. The collection contains 110,352 tweets related with the entities, out of which 34,882 are in the training set and 75,470 are in the test set. The 32,175 unrelated tweets of the dataset are distributed as follows: 10,797 tweets in the training set and 21,378 in the test set. The table also shows the distributions by domain.

Table 3 shows the distribution of polarity classes in the RepLab 2013 dataset. The RepLab 2013 dataset contains 63,442 tweets classified as positive by the annotator, 30,493 classified as neutral and 16,415 classified as negative. The distribution in the training set is 19,718 tweets classified as positive, 9,753 as neutral and 5,409 as negative, while the test set contains 63,442 positive tweets, 30,493 neutral tweets and 16,415 negatives.

**Table 4.** RepLab 2013 dataset for the Topic Detection Task

	<b>All Automotive Banking Universities Music/Artist</b>				
Training No. Topics	3,813	1,389	831	503	1,090
Training Average No. Tweets per Topic	14.40	12.36	11.35	17.57	16.53
Test No. Topics	5,757	1,959	1,121	1,035	1,642
Test Average No. Tweets per Topic	21.14	18.42	18.95	21.78	24.74
Total No. Topics	9,570	3,348	1,952	1,538	2,732
Total Average No. Tweets per Topic	17.77	15.39	15.15	19.67	20.64

**Table 5.** RepLab 2013 dataset for the Priority Detection Task

	<b>All Automotive Banking Universities Music/Artist</b>				
Training No. Alert	1,540	226	841	88	385
Training No. Mildly_Important	17,961	5,388	2,509	1,949	8,115
Training No. Unimportant	15,379	5,742	2,403	1,375	5,859
Test No. Alert	3,240	483	2,195	102	460
Test No. Mildly_Important	38,617	10,967	5,429	4,441	17,780
Test No. Unimportant	33,613	12,965	4,429	3,172	13,047
Total No. Alert	4,780	709	3,036	190	845
Total No. Mildly_Important	56,578	16,355	7,938	6,390	25,895
Total No. Unimportant	48,992	18,707	6,832	4,547	18,906

Table 4 displays the number of topics per set as well as the average number of tweets per topic, which is 17.77 for the whole collection but goes from 14.40 in the training set to 21.14 in the test set. The training set contains 3,813 different topics, the test set 5,757 different topics, for a total of 9,570 different topics in the RepLab 2013 dataset.

Finally, Table 5 summarizes the distributions of tweets in priority classes. The less representative class is *alert*, with 4,780 tweets classified as a possible reputation alert in the whole corpus. *Mildly\_Important* has 56,578 tweets and *Unimportant* receives 48,992 tweets.

In order to determine inter-annotator agreement we perform two different experiments. First, 14 entities (4 automotive, 3 banking, 3 universities, 4 music) have been labeled by two annotators. This subset contains 31,381 tweets that represent 22% of the RepLab 2013 dataset covering all domains. Second, three annotators labeled 3 entities of the automotive domain. Table 6 shows the results of the first experiment of agreement using percentage of agreement and Kappa

**Table 6.** RepLab 2013 agreement: analysis of 14 entities labeled by two annotators

	% Agreement	Cohen $\kappa$	Fleiss $\kappa$	$F_1(\mathbf{R}, \mathbf{S})$
Training Filtering	94.80	70.01	68.84	-
Training Polarity	68.27	41.04	38.93	-
Training Topic Detection	-	-	-	49.59
Training Priority Detection	58.41	23.96	15.96	-
Test Filtering	96.46	68.00	67.86	-
Test Polarity	68.81	42.26	39.92	-
Test Topic Detection	-	-	-	48.07
Test Priority Detection	60.88	29.29	20.91	-
Total Filtering	95.94	66.69	66.35	-
Total Polarity	68.59	41.93	39.79	-
Total Topic Detection	-	-	-	-
Total Priority Detection	60.07	28.04	20.24	-

metrics (both, Cohen and Fleiss) for filtering, polarity and priority detection tasks, and F measure of Reliability and Sensitivity for topic detection task. As can be observed, the percentage of agreement for the filtering subtask is near 100%, while taking in to account the class distribution with the kappa metrics the inter agreement between annotator decreases. The values obtained for reputational polarity in terms of percentage of agreement are quite similar to other studies over sentiment analysis task. As in the filtering subtask, the value obtained with kappa in the reputational polarity subtask decrease with respect of percentage of agreement. For the topic detection subtask, we do not compute inter agreement between annotator for the whole RepLab 2013 dataset. This is due the organization of the labeling process. The annotators consider the training and test set as two different sets, so cannot group tweets of both set. The agreement for the topic detection task is higher than expected, taking into account the complexity of the subtask.

As expected, the results obtained in the experiment with three annotators are lower. As can be seen in Table 7, the inter agreement for the filtering task is quite similar to that obtained in the experiment with two annotators, while the results for the reputational polarity decrease considerably in all metrics. Concerning the topic detection subtask, the table shows the average of F measure over all combinations between annotators. Notably, this task is the one with a lower decrease with respect to the experiment with two annotators, even if this subtask depends on the organization behavior of the annotators. Similarly to the previous experiments of two annotators, as the training and test are considered as two sets by the annotator, the topic detection inter agreement for the whole RepLab 2013 dataset is not computed. Finally, the values obtained for the priority task for three annotators decrease more than for topic detection comparing with the previous experiment, but are still similar.

**Table 7.** RepLab 2013 agreement analysis of 3 entities labeled by three annotators

	% Agreement Fleiss $\kappa$ Average( $F_1(\mathbf{R}, \mathbf{S})$ )		
Training Filtering	92.46	56.63	-
Training Polarity	48.81	36.75	-
Training Topic Detection	-	-	48.11
Training Priority Detection	46.89	27.23	-
Test Filtering	91.54	59.60	-
Test Polarity	51.98	39.11	-
Test Topic Detection	-	-	51.33
Test Priority Detection	53.93	36.04	-
Total Filtering	91.83	59.59	-
Total Polarity	51.03	38.59	-
Total Topic Detection	-	-	-
Total Priority Detection	51.72	33.38	-

## 4 Participation

44 groups signed up for RepLab 2013, although only 15 of them submitted runs to the official evaluation.<sup>2</sup> This year the task was defined in such a way that using the baselines provided by the organizers, every group, besides participating in a concrete subtask, could submit its system to the full task. Nevertheless, only 4 systems explicitly used this possibility.<sup>3</sup> Overall, 5 groups participated in the topic detection subtask, 11 in the reputation polarity classification subtask, 14 in the filtering subtask and 4 in the priority assignment subtask. Below we list the participants and briefly describe the approaches used by each group. Table 8 shows the acronyms and affiliations of the research groups that took part in RepLab 2013.

*CIRGDISCO* participated in the filtering subtask. They exploited “context phrases” found in tweets and Wikipedia disambiguated articles for a particular entity in an SVM classifier that utilizes features extracted from the Wikipedia graph structure, i.e. incoming and outgoing links from and to Wikipedia articles. They used, in addition, features derived from term-specificity and term-collocation features derived from the Wikipedia article of the analysed entity.

*Daedalus* submitted specific runs for the filtering and polarity subtasks, apart from the full task. Their approach to the filtering subtask is based on the use of linguistic processing modules to detect and disambiguate named

<sup>2</sup> One additional group sent their results two days after the deadline, and their runs are reported here as “unofficial.” An asterisk in tables indicates an unofficial result.

<sup>3</sup> Daedalus, GAVKTH, SZTE\_NLP, and UNED ORM.

**Table 8.** List of participants: acronyms and affiliation

Acronym	Affiliation	Country
CIRGDISCO	National University of Ireland, Galway	Ireland
Daedalus	Daedalus, S.A.	Spain
DIUE	Universidade de Évora	Portugal
GAVKTH	Gavagai	Sweden
IE	National University of Singapore	Singapore
LIA	University of Avignon	France
NLP&IR_GROUP_UNED	UNED	Spain
POPSTAR	Universidade Porto	Portugal
REINA	Reina Research Group, University of Salamanca	Spain
SZTE_NLP	University of Szeged	Hungary
UAMCLYR	Universidad Autónoma Metropolitana Cuajimalpa	Mexico
UNED_ORM	UNED	Spain
UNED-READERS*	UNED	Spain
UNEDTECNALIA	Tecnalia Research And Innovation, UNED	Spain
UVA_UNED	University of Amsterdam, UNED	The Netherlands, Spain
volvam	Volvam Analytics and University of Alicante	Ireland, Spain

entities at several levels. The 4 submitted runs are defined by a combination of morphosyntactic-based vs. semantic disambiguation and a case sensitive/insensitive processing of the tweets. On the other hand, the polarity classification uses a lexicon-based approach to sentiment analysis, improved with a full syntactic analysis and detection of negation and polarity modifiers, which also provides the polarity at entity level.

*DIUE* applied a supervised Machine Learning (ML) approach for the polarity classification subtask. The Python NLTK has been used for preprocessing, including file parsing, text analysis and feature extraction. The best run combines bag-of-words with a set of 18 features related to presence of the polarized term, negation before the polarized expression, as well as entity reference based on sentiment lexicons and shallow text analysis.

*GAVKTH* used its commercially available system for the filtering and reputation polarity subtasks. The system, designed for large scale analysis of streaming text and measuring the public attitude towards targets of interest, has been used with no adjustment for the specific subtasks. The basic approach relies on distributional semantics represented in a semantic space by means of a patented implementation of the Random Indexing processing framework.

*LIA* applied a large variety of ML methods mainly based on exploiting tweet contents to filtering, polarity classification, topic detection, and priority assignment. In several experiments some metadata were added and a fewer number of runs incorporated external information by using provided links to Wikipedia and entities' official web sites.

*NLP&IR\_GROUP\_UNED* focused on addressing filtering and reputation polarity classification using an IR method. Viewing these two subtasks as the same problem, i.e. finding the most relevant class to annotate a given tweet, a classical IR approach was applied, using the tweet content as query against an index with the models of the classes used to annotate tweets. The classes were modelled by means of the Kullback Leibler Divergence (KLD), in order to extract their most representative terminology. For topic detection, instead a clustering based technique, this group resorted to Formal Concept Analysis (FCA) to represent the contents in a lattice structure. Topics were extracted from the lattice using a FCA concept, *stability*.

*popstar* participated in the filtering and reputation polarity classification subtasks. For filtering, these researchers explored different learning algorithms considering a variety of features describing the relationship between an entity and a tweet, such as text, keyword similarity scores between entities metadata and tweets, the Freebase entity graph and Wikipedia.

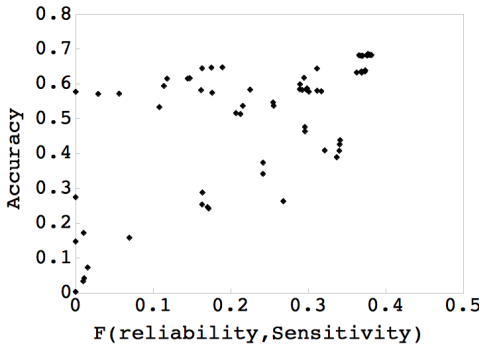
*REINA* used classical systems for the similarity matrix and community detection techniques for topic detection. No distinction was made between languages of the tweets, doing a uniform lexical analysis of all tweets, applying a simple stemmer and removing the words with less than 4 characters. Additionally, the discarded emoticons were considered as well as hashtags and some entities terms. The urls shared by two tweets were deemed as another important feature of the tweets, assuming this is indicative of topic similarity.

*SZTE\_NLP* presented a system to tackle the filtering and reputation polarity classification subtasks using supervised ML techniques. Several Twitter specific text preprocessing and features engineering methods were applied. Besides supervised methods, they also experimented with incorporating clustering information.

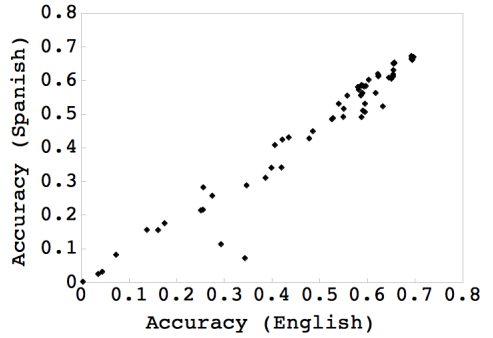
*UAMCLYR* adopted Distributional Term Representations (DTR) to tackle the filtering and reputation polarity classification subtasks. Terms were represented by means of contextual information given by the term co-occurrence statistics. For topic detection and priority assignment, these researchers explored clustering and classification methods as well as term selection techniques working with two settings: single tweets and tweets extended with derived posts.

*UNED ORM* submitted runs to the full task and all the subtasks testing several approaches. First, Instance-based learning using Heterogeneity Based Ranking





**Fig. 1.** Polarity: Accuracy versus R&S



**Fig. 2.** Polarity: Accuracy EN vs ES

to combine seven different similarity measures was applied to all the subtasks. The filtering subtask was also tackled by automatically discovering positive and negative filter keywords, i.e. terms present in a tweet that reliably predict the relatedness or non-relatedness of the message to the analysed entity. The topic detection subtask was attempted with three approaches: agglomerative clustering over Wikified tweets, co-occurrence term clustering and an LDA-based model that uses temporal information. Finally, the polarity subtask was tackled by generating domain specific semantic graphs in order to automatically expand the general purpose lexicon SentiSense.

*UNED-READERS\** applied an unsupervised knowledge-based approach to filter relevant tweets for a given entity. The method exploits a new way of contextualizing entity names from relatively large collections of texts using probabilistic signature models, i.e., discrete probability distributions of words lexically related to the knowledge or topic underlying the set of entities in background text collections. The contextualization is intended to recover relevant information about the entity, particularly, lexically related words, from background knowledge.

*UNEDTECNALIA* submitted a filtering algorithm that takes advantage of the Web of Data in order to create a context for every entity. The semantic context of the analysed entities is generated by querying different data sources (modelled by a set of ontologies) provided by the Linked Open Data Cloud. The extracted context is then compared to the terms contained in the tweet.

*UVA\_UNED*, a collaborative participation of UvA and UNED, focused on applying an active learning approach to the filtering subtask. It consisted of exploiting features based on the detected semantics in the tweet (using Entity Linking with Wikipedia), as well as tweet-inherent features such as hashtags and usernames. The tweets manually inspected during the active learning process were at most 1% of the test data.

*volvam* participated in polarity classification and applied one supervised and two unsupervised approaches, combining ML and lexicon-based techniques with an emotional concept model. These methods had been properly adapted to English and Spanish depending on the resources available for each language. The first, unsupervised, approach made use of fuzzy lexicons in order to catch informal variants that are common in Twitter texts. The supervised method extended the first approach with ML techniques and an emotion concept model, while the last one also employed ML but incorporating the bag-of-concepts approach using SenticNet common-sense affective knowledge.

## 5 Evaluation Results

### 5.1 Polarity

Polarity has been evaluated according to Accuracy and R&S. Only entity-related tweets in the test set have been assessed. In order to keep evaluation independent from the filtering task, we do not penalize polarity annotations made on non-related tweets. That is, only related tweets are considered in the Accuracy and R&S computation. The related tweets without system response are penalized.

**Table 9.** Accuracy, ratio of processed tweets, correlation at entity level, Reliability and Sensitivity for polarity task

RUN	ACC.	PROCESSED TWEET RATIO	CORR. LEVEL	ENT.	R	S	F
SZTE NLP 8	0.69	1.00	0.88		0.48	0.34	0.38
LIA 7	0.65	1.00	0.82		0.50	0.15	0.19
POPSTAR 5	0.64	0.98	0.89		0.43	0.34	0.37
UAMCLYR 2	0.62	1.00	0.82		0.38	0.27	0.29
UNED ORM 2	0.62	1.00	0.70		0.36	0.10	0.15
LIA 3	0.60	1.00	0.64		0.37	0.27	0.29
UNED ORM 1	0.59	1.00	0.87		0.32	0.29	0.30
BASELINE	0.58	1.00	0.87		0.32	0.29	0.30
NLP IR UNED 1	0.58	1.00	0.79		0.33	0.31	0.32
UAMCLYR 05	0.58	1.00	0.78		0.33	0.29	0.30
IE 6	0.58	1.00	0.22		0.94	0.00	0.00
ALL POSITIVE	0.58	1.00	0.00		1.00	0.00	0.00
DIUE 1	0.55	1.00	0.21		0.33	0.22	0.25
VOLVAM 3	0.54	1.00	0.36		0.32	0.23	0.26
IE 5	0.52	1.00	0.18		0.29	0.22	0.21
DAEDALUS 3	0.44	1.00	0.52		0.31	0.40	0.34
VOLVAM 2	0.41	1.00	0.38		0.31	0.39	0.34
GAVKTH 6	0.37	0.98	0.49		0.30	0.21	0.24
ALL NEUTRAL	0.27	1.00	0.00		1.00	0.00	0.00
GAVKTH 2	0.26	0.82	0.21		0.37	0.21	0.27
ALL NEGATIVE	0.15	1.00	0.00		1.00	0.00	0.00

The system results, sorted by accuracy are shown in Table 9. The table includes only the best system, according to R&S or Accuracy, for each team. The second column contains the ratio of tweets for which the output gives results.

The majority class is the dataset is “POSITIVE”. The baseline approach appears in the middle of the ranking. SZTE and POPSTAR teams improve, in general, most systems according to both accuracy and R&S. Note that some systems achieve a low accuracy (under the BASELINE) but with competitive R&S. As R&S only look at the relative ordering between tweets (rather than the actual tags), a possible reason is that, while many tags are not correct, the ordinal polarity relationship between them is correct. Figure 1 illustrates the correspondence between Accuracy and R&S. Note that a high R&S tends to be associated with a high accuracy.

Another important aspect of polarity detection for ORM, is the ability to predict the average polarity of an entity with respect to other entities. To evaluate this ability, we have computed the Pearson correlation between the average estimated and real polarity levels across entities.<sup>4</sup> An interesting result is that some approaches are able to estimate the average polarity reputation for an entity with a 0.9 correlation with the ground truth.

Finally, Figure 2 shows the correlation between Accuracy scores over English versus Spanish tweets. In most cases there is a high correspondence. The accuracy for Spanish seems to be upper bounded by the accuracy over English tweets.

## 5.2 Filtering

In this task, tweets must be classified as related or unrelated to the entity of interest. R&S in filtering tasks (two levels) correspond with the products of precision in both classes and the product or recall scores respectively. Table 10 shows the Accuracy and R&S results for the filtering task. Again, we have included only the best run according to Accuracy or R&S for each team. Most tweets are related (77%). As well as in the polarity tasks, the baseline approach appears in the middle of the ranking for both R&S and Accuracy. Figure 3 shows the correspondence between Accuracy and R&S. As well as in the polarity task, a high R&S ensures a high Accuracy. As well as in polarity task, there are not important differences in system scores when considering the Spanish vs. English tweets. There is a 0.94 Pearson Correlation) between scores over both kind of tweets. In general, the top scores are much higher than in RepLab 2012; this is explained by the fact that in this new dataset the training and test entities are the same.

## 5.3 Priority

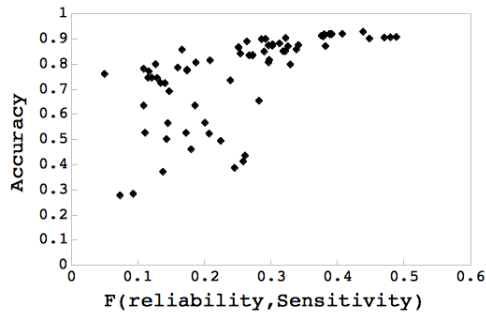
The Priority task consists of classifying tweets into three levels. Reliability represents the ratio of correct priority relationships per tweet, while Sensitivity represents the ratio of captured relationships per tweet. In this case, as well as

---

<sup>4</sup> For the correlation computation, we assign 0,1 and 2 for each class respectively.

**Table 10.** Results for the Filtering Subtask

RUN	R	S	F	ACC
POPSTAR 2	0.73	0.45	0.49	0.91
SZTE NLP 7	0.60	0.44	0.44	0.93
LIA 1	0.66	0.36	0.38	0.87
UAMCLYR 04	0.56	0.4	0.38	0.91
LIA 6	0.62	0.33	0.34	0.88
UNED ORM 2	0.43	0.38	0.34	0.86
BASELINE	0.49	0.32	0.33	0.87
Daedalus 1	0.35	0.45	0.32	0.85
UNED-READERS 2	0.38	0.33	0.28	0.55
CIRG IRDISCO 4	0.34	0.33	0.27	0.84
IE 4	0.45	0.23	0.26	0.44
CIRG IRDISCO 1	0.5	0.24	0.25	0.87
Uva UNED 6	0.68	0.22	0.21	0.82
UNEDTECNALIA 1	0.28	0.29	0.18	0.46
NLP IR GROUP UNED 9	0.29	0.22	0.17	0.78
IE 2	0.46	0.16	0.17	0.53
CIRG IRDISCO 2	0.82	0.16	0.17	0.86
NLP IR GROUP UNED 8	0.31	0.19	0.16	0.79
GAVKTH 1	0.81	0.07	0.05	0.76
ALL RELATED	0	0	0	0.77
ALL UNRELATED	0	0	0	0.23

**Fig. 3.** Accuracy versus R&S in the Filtering Task

in polarity, only the related tweets (according to assessors) are considered in the evaluation process. Table 11 shows the results. Only the best Accuracy and R&S score per team is included. Not all systems have annotated all tweets (see the last column). The best run achieves a high score for both R&S and Accuracy measures. The baseline approach is improved substantially for both measures.

## 5.4 Topic Detection

Topic detection is a clustering task which has been evaluated according to R&S, which correspond with the popular measures Bcubed precision and Recall [1].

**Table 11.** Accuracy, Reliability and Sensitivity Results for the Priority Subtask

RUN	R	S	F	ACC	Amount of processed tweets
LIA 5	0.39	0.32	0.34	0.63	0.97
UNED ORM 1	0.31	0.31	0.3	0.6	1
BASELINE	0.3	0.3	0.3	0.6	1
GAVKTH 2	0.36	0.19	0.25	0.37	0.82
UAMCLYR 2	0.24	0.2	0.2	0.46	1
GAVKTH 7	0.37	0.09	0.13	0.41	0.83
UAMCLYR 3	0.58	0.07	0.09	0.57	1
ALL MILDLY IMPORTANT	0	0	0	0.52	1
ALL UNIMPORTANT	0	0	0	0.44	1
ALL ALERT	0	0	0	0.04	1

**Table 12.** Reliability and Sensitivity in the Topic Detection Task

RUN	S	R	F	Ratio proc. tweets
UNED_ORM_2	0.46	0.32	0.33	0.99
REINA_2	0.32	0.43	0.29	0.79
LIA_3	0.22	0.35	0.25	1.00
UAMCLYR_7	0.35	0.50	0.24	0.97
REINA_1	0.16	0.52	0.23	0.99
BASELINE	0.15	0.22	0.17	1.00
NLP_IR_UNED_1	0.67	0.11	0.17	0.53
ALLINONE	0.07	1.00	0.12	1.00
ALLINALL	1.00	0.04	0.07	1.00

Table 12 displays the results. Only the best F measure is considered for each team. Figure 4 shows that there is an important trade-off between R and S in this task. In these circumstances, the F measure weighted with  $\alpha = 0.5$  rewards the runs located in the diagonal axis. But this choice of  $\alpha$  is, to some extent, arbitrary. For this reason, we check the evaluation results according to UIR (see previous section). UIR is a complementary measure that indicates to what extent run improvements are sensitive to variations in the measure weighting scheme (i.e. in  $\alpha$ ). Table 13 shows for all runs, the other runs which are improved by the first with  $UIR \geq 0, 2$ . This implies that there is a difference higher than 0.2 between the cases in which the first run improves the other for R and S and vice versa. Interestingly, although UAMCLYR\_7 is not the best system in the  $F_{\alpha=0.5}$  ranking, it improves robustly a great amount of runs. Some team runs like LIA are not comparable to each other. Probably, they have different grouping thresholds.

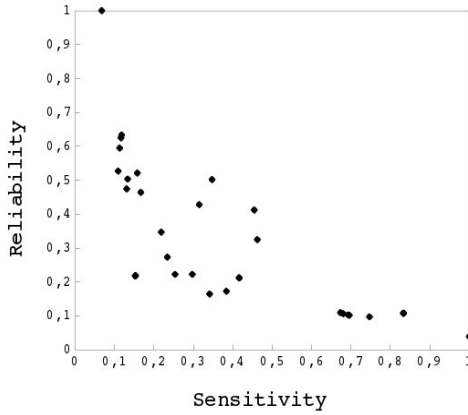


Fig. 4. Reliability vs. Sensitivity in the Topic Detection Task

Table 13. UIR analysis for the Topic Detection Task

RUN	Improves runs UIR $\geq$ 0.2	Amount of improved runs
UAMCLYR_07	UAMCLYR_1,2,3,4,5,6 LIA_2,3,4 REINA_1 BASELINE UNED_ORM_1	12
UNED_ORM_2	LIA_1,2,3,4 BASELINE UNED_ORM_1,3,4,5,6,7 BASELINE	11
REINA_2	LIA_1,2,3,4 BASELINE UAMCLYR_4 UNED_ORM_1,6,7	9
UAMCLYR_8	LIA_2,4 UAMCLYR_1,2,3,4 BASELINE UNED_ORM_1	8
UNED_ORM_4	BASELINE UNED_ORM_1,6 LIA_1,4	5
UNED_ORM_5	BASELINE UNED_ORM_1,6 LIA_1,4	5
REINA_1	UAMCLYR_3,4 BASELINE UNED_ORM_1	4
UNED_ORM_3	BASELINE UNED_ORM_1 LIA_1 UNED_ORM_6	4
UNED_ORM_7	LIA_2,4 BASELINE UNED_ORM_1	4
UAMCLYR_6	BASELINE UAMCLYR_4 UNED_ORM_1	3
UAMCLYR_3	BASELINE UAMCLYR_4 UNED_ORM_1	3
NLP_IR_UNED_10	NLP_IR_UNED_3,4,5	3
UAMCLYR_5	BASELINE UAMCLYR_04 UNED_ORM_1	3
NLP_IR_UNED_8	NLP_IR_UNED_3,4,5	3
NLP_IR_UNED_9	NLP_IR_UNED_3,4,5	3
LIA_2	BASELINE UNED_ORM_1	2
LIA_3	BASELINE UNED_ORM_1	2
LIA_4	BASELINE UNED_ORM_1	2
UNED_ORM_6	BASELINE UNED_ORM_1	2
UAMCLYR_01	UAMCLYR_02	1
UAMCLYR_04	BASELINE	1
NLP_IR_UNED_6	NLP_IR_UNED_4	1
NLP_IR_UNED_7	NLP_IR_UNED_4	1

**Table 14.** Full Task Results

RUN	F measure
UNED_ORM_2	0.19
UNED_ORM_7	0.18
UNED_ORM_4	0.17
UNED_ORM_6	0.17
DAEDALUS_1..8	0.16
UNED_ORM_1	0.16
UNED_ORM_8	0.12
UNED_ORM_3	0.11
UNED_ORM_5	0.11
SZTE_NLP_1..10	0.03

**Table 15.** UIR Analysis for the Full Task

RUN 1	RUN 2	Imp.	Is imp.	UIR
UNED_ORM_2	UNED_ORM_4	24	1	0.38
UNED_ORM_2	UNED_ORM_6	15	0	0.25
UNED_ORM_3	UNED_ORM_5	14	1	0.21
SZTE_7	SZTE_4	44	15	0.47
SZTE_7	SZTE_3	43	15	0.46
SZTE_7	SZTE_6	44	17	0.44
SZTE_7	SZTE_1	42	17	0.41
SZTE_7	SZTE_2	40	15	0.41
SZTE_7	SZTE_5	43	19	0.39
SZTE_7	SZTE_9	40	18	0.36
SZTE_7	SZTE_8	37	17	0.33
SZTE_7	SZTE_10	35	18	0.28
SZTE_10	SZTE_9	37	22	0.25

## 5.5 Full Task

The full task joins filtering, priority and topic detection tasks. The use of R&S allows to apply the same evaluation criterion to all subtasks and therefore, to combine all of them. It is possible to apply R&S directly over the set of relationships (priority, filtering and clustering) but then the most frequent binary relationships dominate the evaluation results (in our case, priority relationships would dominate). We decided to use a weighted harmonic mean (F measure) of the six Reliability and Sensitivity measures corresponding to the three subtasks embedded in the full task. In cases of empty partial outputs, we have completed runs with the baseline approach as specified in the guidelines.

Table 14 shows the team ranking in terms of F. However, this evaluation is highly sensitive to the relative importance of measures in the combining function. For this reason, we have also computed UIR between each pair of runs. Here we consider as an unanimous improvement of system A over system B to those test cases (entities) for which all the six measures are better for A than for B.

Results of the UIR analysis are shown in Table 15. The third and fourth columns represent how many entities one run improves or is improved by the other. It only includes those run pairs for which UIR is bigger than 0.2. As the table shows, actually, runs from different teams are not comparable to each other: improvements in F are dependent on the relative weighting scheme. However, there are a number of significant improvements (in terms of UIR) between runs from the same teams.

## 6 Conclusions

Perhaps the main outcome of RepLab 2013 is its dataset, which comprises more than 142,000 tweets in two languages with four types of high-quality manual annotations, covering all essential aspects of the reputation monitoring process. We expect this dataset to become a useful resource for researchers not only in the field of reputation management, but also for researchers in Information Retrieval and Natural Language Processing in general. Just to give an example, the topics (tweet clusters) together with their relative ranking can be directly mapped into a test collection to evaluate search with diversity algorithms over Twitter.

Comparing with RepLab 2012, availability of training data for the entities in the test set naturally improves system results and also allows for a more straightforward application of machine learning techniques. But the tasks themselves are still far from solved; even with plenty of entity-specific training material the RepLab tasks—polarity, topic detection, and ranking—have proved challenging for state-of-the-art systems.

## References

1. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval* 12(4), 461–486 (2009)
2. Amigó, E., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M.: Overview of RepLab 2012: Evaluating Online Reputation Management Systems. In: *CLEF 2012 Labs and Workshop Notebook Papers* (2012)
3. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks. *Journal of Artificial Intelligence Research* 42(1), 689–718 (2011)
4. Amigó, E., Gonzalo, J., Verdejo, F.: A General Evaluation Measure for Document Organization Tasks. In: *Proceedings of SIGIR 2013* (July 2013)



# Entity Recognition in Parallel Multi-lingual Biomedical Corpora: The CLEF-ER Laboratory Overview

Dietrich Rebholz-Schuhmann<sup>1,2</sup>, Simon Clematide<sup>1</sup>, Fabio Rinaldi<sup>1</sup>,  
Senay Kafkas<sup>2</sup>, Erik M. van Mulligen<sup>3</sup>, Chinh Bui<sup>3</sup>,  
Johannes Hellrich<sup>4</sup>, Ian Lewin<sup>5</sup>, David Milward<sup>5</sup>, Michael Poprat<sup>6</sup>,  
Antonio Jimeno-Yepes<sup>7</sup>, Udo Hahn<sup>4</sup>, and Jan A. Kors<sup>3</sup>

<sup>1</sup> Department of Computational Linguistics, University of Zürich, Ch  
(rebholz,clematide,rinaldi)@ifi.uzh.ch

<sup>2</sup> European Bioinformatics Institute, Wellcome Trust Genome Campus,  
Hinxton, Cambridge, CB10 1SD, U.K.  
kafkas@ebi.ac.uk

<sup>3</sup> Department of Medical Informatics, Erasmus University Medical Center,  
Rotterdam  
(kors,e.vanmulligen)@erasmusmc.nl, bqchinh@gmail.com

<sup>4</sup> Jena University Language & Information Engineering (JULIE) Lab,  
Friedrich-Schiller-Universität Jena, Fürstengraben 30, D-07743 Jena  
(udo.hahn,johannes.hellrich)@uni-jena.de

<sup>5</sup> Linguamatics Ltd, 324 Science Park, Milton Road, Cambridge CB4 0WG  
(ian.lewin,david.milward)@linguamatics.com

<sup>6</sup> Averbis GmbH, Tennenbacher Strasse 11, D-79106 Freiburg  
poprat@averbis.de

<sup>7</sup> National ICT Australia, Victoria Research Laboratory, Melbourne, Australia  
antonio.jimeno@gmail.com

**Abstract.** The identification and normalisation of biomedical entities from the scientific literature has a long tradition and a number of challenges have contributed to the development of reliable solutions. Increasingly patient records are processed to align their content with other biomedical data resources, but this approach requires analysing documents in different languages across Europe [1,2].

The CLEF-ER challenge has been organized by the Mantra project partners to improve entity recognition (ER) in multilingual documents. Several corpora in different languages, i.e. Medline titles, EMEA documents and patent claims, have been prepared to enable ER in parallel documents. The participants have been asked to annotate entity mentions with concept unique identifiers (CUIs) in the documents of their preferred non-English language.

The evaluation determines the number of correctly identified entity mentions against a silver standard (Task A) and the performance measures for the identification of CUIs in the non-English corpora. The participants could make use of the prepared terminological resources for entity normalisation and of the English silver standard corpora (SSCs) as input for concept candidates in the non-English documents.

The participants used different approaches including translation techniques and word or phrase alignments apart from lexical lookup and other text mining techniques. The performances for task A and B was lower for the patent corpus in comparison to Medline titles and EMEA documents. In the patent documents, chemical entities were identified at higher performance, whereas the other two document types cover a higher portion of medical terms. The number of novel terms provided from all corpora is currently under investigation.

Altogether, the CLEF-ER challenge demonstrates the performances of annotation solutions in different languages against an SSC.

## 1 Introduction

Advances in the research community are often driven by specific challenges, which are meant to benchmark the outcomes on a well defined task. Over recent years a number of challenges have been proposed that focus on different tasks for the development of innovative technologies: e.g. different CLEF challenges such as CLEFeHealth and CLEF-IP [3,4], the BioCreAtIve sequel [5,6], the bioNLP Shared Tasks [7], and the CALBC challenge [8,9].

Most challenges propose a gold standard corpus that is then used for the benchmarking of the proposed solutions. In addition, other challenges have been proposed that consider a silver standard corpus instead. This approach allows the processing of large corpora in contrast to the gold standard approaches.

The CLEF-ER challenge is unique in the sense that it combines different expectations and technologies, such as entity recognition in the biomedical domain with multilingual approaches and machine translation.

Furthermore, the CLEF-ER challenge anticipates the processing and management of large resources and will exploit the delivered results for the development of augmented terminological resources.

## 2 Background

The CLEF conference sequel has a long tradition in setting up challenges for the research community. The challenge tasks are concerned with information retrieval, covering different types of electronic data, e.g. images, texts, and their combinations, and also considering different domain knowledges, for example medical and clinical data in comparison to legal texts and patents. All challenges are organised as part of a CLEF laboratory and the overall conference serves the purpose of the exchange of information.

Other challenges in the biomedical research community are also focused to information retrieval, namedly in TREC Genomics [10], but tackle in addition other tasks such as information extraction, entity recognition and fact extraction. The BioCreAtIve challenges are tuned to develop solutions that would help biomedical curators to do their work in finding facts from the literature [11]. The BioNlp Shared Task serves the same purpose and increasingly seeks the

integration between ontological resources and the text mining component. Recently the BioASQ<sup>1</sup> challenge has been introduced, which aims at the tasks of topic identification and question answering in the biomedical domain.

None of the challenges has been organized in a way to feed the results from the challenge into building resources as it is the case for the CLEF-ER challenge and the MANTRA<sup>2</sup> project.

Furthermore, most challenges make use of a gold standard corpus (GSC) to evaluate the contributions from the participants. There is no doubt that a GSC is a precious resource and forms the key means to determine novel standards for a specific task in the research community. On the other side, it has been shown that GSCs are selective in the sense that they limit the evaluation of the specific tasks to a relatively small number of samples as instances representing the standard. By contrast, it is important to develop resources and standards at a scale that are more representative for the underlying tasks and the long-term goals.

The CALBC challenge has been such an initiative that was tackling the annotation of a large-scale corpus in the biomedical domain with a significant number of named entities for the benefits of long-term development of entity recognition solutions. The project partners have prepared a lexical resource, a large-scale annotated corpus, and a triple store containing the facts from the scientific literature covering the information in the annotated corpus.

The MANTRA project and the CLEF-ER challenge extend the work from the CALBC challenge into the development of multilingual resources for the medical domain. With the help of parallel corpora and a multilingual terminological resource, the project partners motivate the participants in the CLEF-ER challenge to contribute annotations in an English and a non-English corpus. The final goal is the annotation of medical entity mentions in the non-English corpus

## 2.1 Overview

This manuscript gives an overview on the setup of the CLEF-ER challenge including the resources that have been developed, the evaluation parameters and the outcomes of the challenge. The next section (“Material and Method”) explains the provided resources, i.e. the terminological resources and the parallel corpora, as well as the evaluation metrics and the generation of the SSCs. Towards the end of the section, an overview on the contributing systems by the participants is given. In the results section, the performances of the systems overall is shown and the performances in dependence of the available corpora, the semantic groups from UMLS, and the different approaches from the participants. In the conclusion section, we will give views on the outcome of the challenge overall.

---

<sup>1</sup> <http://www.bioasq.org/>

<sup>2</sup> <http://www.mantra-project.eu/>

### 3 Material and Method

#### 3.1 Terminologies

The MANTRA Terminological Resources (MTR) [12] used for the CLEF-ER challenge were derived from the Unified Medical Language System (UMLS) Metathesaurus [13]. The UMLS Metathesaurus is an umbrella system combining over 100 biomedical terminologies, e.g. the Medical Subject Headings (MeSH), the Medical Dictionary for Regulatory Activities Terminology (MedDRA, [14]) or the Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT, [15]). The UMLS Metathesaurus contains both hierarchical (e.g. 'isa') and associative (e.g. 'caused by') relations between its entries, called *concepts*. Each concept is identified by a Concept Unique Identifier (CUI) and can have multiple names per language, called *synonyms*. Concepts are organized by semantic types (e.g. 'steroid'), which are themselves organized into semantic groups (e.g. 'chemicals & drugs'). To derive the MTR from the UMLS Metathesaurus we selected a subset containing only entries from selected semantic groups, e.g. anatomy (ANAT). This was done both due to the lower frequency and perceived irrelevance of the other semantic groups. The MTR contain 531,466 concepts with 2,839,277 synonyms (cf. tbl. 1 for details).

The MTR were distributed to the participants as a single file in the OBO format [16], which was selected both due to existing tooling and its readability for humans. The MTR is provided through the submission site of the CLEF-ER challenge<sup>3</sup> and requires a proper UMLS license from the participants.

**Table 1. (Terminological resource):** The English part of the TR contains most terms. Only Spanish is covered in SNOMED-CT. MedDRA terms have been translated in all languages.

Terms	MeSH	SNOMED-CT	MedDRA
en	764,000	1,184,005	56,061
de	77,249	-	50,128
fr	105,758	-	49,586
es	59,678	1,089,723	49,499
nl	40,808	-	50,932

#### 3.2 Selection of Parallel Corpora

Different corpora have been selected and tested as input to the CLEF-ER challenge [12]. The parallel corpora have to be available in different (European) languages, should be available in languages that are shared between the different corpora, should have a reasonable size, and should deal with biomedical topics. The selection of Medline abstracts and EMEA drug labels fulfills the requirements. In addition, patent claims have been selected from patents that cover

<sup>3</sup> <https://sites.google.com/site/mantraeu/terminology>

**Table 2. (Units counts, all corpora):** The number of units is highest in English for Medline. German and French are evenly well covered in all three corpora, and Spanish shows similar coverage, except that Spanish (and Dutch) are not represented for patent texts.

Units	EMEA	Medline	Patent
en	140,552	1,593,546	120,638
de	140,552	719,232	120,637
fr	140,552	572,176	120,636
es	140,552	247,655	
nl	140,552	54,483	

**Table 3. (Submissions to the CLEF-ER challenge):** The Table gives an overview on the submissions to the CLEF-ER challenge. For all corpora and for all languages at least one annotated corpus has been contributed.

Count	Column Labels														Total
	EMEA					Medline					Patent				
Cont.	de	en	es	fr	nl	de	en	es	fr	nl	de	en	fr		
A			3												3
B		1	1				1	1							4
C			2	2				2	2						8
D	1		1	1	1	1		1	1	1	1		1	1	10
E	1			1		1			1		1		1	1	6
F	2		2	2	2	2		2	2	1	2		2	2	19
G						2		2	2						6
<b>Total</b>	<b>4</b>	<b>1</b>	<b>9</b>	<b>6</b>	<b>3</b>	<b>6</b>	<b>1</b>	<b>8</b>	<b>8</b>	<b>2</b>	<b>4</b>		<b>4</b>		<b>56</b>

biomedical topics. In the latter case, the language in the documents different from the scientific language, but the documents form an important part of the biomedical domain.

All corpora have been processed and transformed in a representation that linking the non-English text (called "units") to the English part of the same document. For Medline abstracts a single unit is a Medline<sup>4</sup> title, for the EMEA<sup>5</sup> drug labels individual paragraphs from the documents form a unit each, and for the patent texts the claim section forms a unit. The overall statistics are shown in the table above (cf. tbl. 2).

Beware that the parallel corpora for patent texts provide the complete claim section in three languages, i.e. in en, de and fr, whereas for the EMEA drug labels the complete documents are delivered in five languages (en, de, fr, es and nl). For the Medline titles, the parallel units are mostly in two languages, i.e. in English and in one non-English language again covering de, fr, es and nl. The reason for this lack of congruency is the fact that the non-English Medline titles

<sup>4</sup> <http://www.nlm.nih.gov/pubs/factsheets/medline.html>

<sup>5</sup> <http://www.ema.europa.eu/>

**Table 4. (Generation of the SSC from CLEF-ER submissions):** The Table gives an overview on the submissions to the CLEF-ER challenge. For all corpora and for all languages at least one annotated corpus has been contributed. The voting threshold has been set to 3, which is 50 % of the contributions.

Contributions for monolingual SSC											
	EMEA				Medline				Patent		All
	de	es	fr	nl	de	es	fr	nl	de	fr	
A		1									1
B		1				1					2
C		1	1			1	1				4
D	1	1	1	1	1	1	1	1	1	1	10
E	1		1		1		1		1	1	6
F	1	1	1	1	1	1	1	1	1	1	10
G					1	1	1				3
<b>All</b>	<b>3</b>	<b>5</b>	<b>4</b>	<b>2</b>	<b>4</b>	<b>5</b>	<b>5</b>	<b>2</b>	<b>3</b>	<b>3</b>	<b>36</b>
<b>Proj. Partners</b>	<b>2</b>	<b>2</b>	<b>3</b>	<b>3</b>	<b>2</b>	<b>2</b>	<b>3</b>	<b>3</b>	<b>2</b>	<b>3</b>	<b>25</b>
<b>Total</b>	<b>5</b>	<b>7</b>	<b>7</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>5</b>	<b>5</b>	<b>6</b>	<b>61</b>

stem from documents that have been delivered from non-English journals and the title has been translated into English and not into any other language.

### 3.3 Preparation of the Silver Standard Corpus

Commonly systems are trained with and evaluated against gold standard corpora created by human experts. Due to the human involvement those are both expensive to create and limited in size. MANTRA follows the CALBC approach of using silver standard corpora (SSCs) instead [9], which are created by harmonizing multiple automatically annotated contributions. A voting scheme is used to determine which annotations are included in the SSC, e.g. only those annotated by a majority of systems. An SSC can be used to evaluate the contributions it was created from with standard metrics like f-score, yet this evaluation can only judge the averageness of a contribution and not its objective quality. We also created a variant SSC from de-annotated contributions, i.e. contributions from which those annotations trivially derived from the MTR were removed. This SSC was then used to evaluate the de-annotated contributions, allowing a better judgment of the conformity regarding new terms, which are otherwise obscured by the enormous amount of terms already contained in the MTR.

*Monolingual Mention Evaluation (Evaluation A).* In order to assess the quality of the annotations in all non-English corpora, a mention agreement evaluation against a harmonized Silver Corpus built from the monolingual contributions of the participants and from annotations from project partners was performed. Table 4 shows the number of annotations from the contributors and partners for the centroid-based SSCs [17]. Not all available contributions have been used to

**Table 5. (Overview on the CLEF-ER participants systems):** The description of the systems that have contributed to the CLEF-ER challenge shows high diversity across the approaches used from the participants. Most participants of the challenge made use of external resources either for their terminology or for word or phrase alignments.

	A	B	C	D	E	F	G
Use of Mantra TR	no	yes	no	no	yes	yes	no
Use of Mantra SSC (in English)	yes	yes	yes(?)	no	yes	no	yes
Statistical Machine Translation	yes	no	yes	yes	no	yes	no
Own Dictionary from	yes	no	no	no	no	no	no
Phrasal Alignment	yes	no	no	yes	no	yes	no
Word Alignment / SMT	yes	no	yes	no	no	yes	no
Indexing (corpora), lexical lookup	no	yes	yes	yes	yes	no	yes
NP identification / Chunking	no	yes	yes	no	no	no	no
Multiple assignment of CUIs	yes	yes	no	yes	yes	yes	yes(?)
Use of Entity disambiguation	no	yes	no	no	no	no	no
Evaluation	no	yes	yes	yes	no	yes	no
Languages	en, es	en, es	fr, es	en, de, nl, fr, es	de, fr	en, de, es, fr, nl	en, de, es, fr
New resources	Translated corpus	--	NP taggers in 3 languages	Translated terminological resource	Enriched terminological resource	Enriched terminological resource	--
Other resources	--	UMLS	UMLS, Wikipedia	MeSH, MedDRA, Snomed-CT	BabelNet (WordNet, Wikipedia)	Lingpipe gazetteer, JCoRe NER engine	UMLS
Other tools	TanI Tagger for ER (MEMM based)	--	Stanford parser, Malt parser, MetaMap, Giza++	Google Translate	GERTWOL, OntoGene term matcher	--	--
Synopsis	ER in a translated corpus	Indexing of the terminology, documents as queries	Synopsis- ML co-training approach on pairs of languages	Translation of the terms via Google, indexing of corpora	Translation of terms via BabelNet, lexical lookup in corpora	Phrase-based SMT & NER	ML approach to identify pairs of terms in 2 languages

generate the SSC for the evaluation of the participants, because a contributor with several similar contributions would gain too much votes in favor of his system and the SSC would therefore be biased. The decision, which annotated corpus will be included into the SSC production, has been left with the challenge participant. All monolingual SSCs used a voting threshold of 3. Spanish and French are well-resourced in terms of different annotations. For German and especially Dutch, the number of contributions is less optimal.

*Cross-lingual Concept Evaluation (Evaluation B).* Given the fact that the English terminology covers a lot more concepts and provides more synonyms for them compared to the non-English terminologies, a second evaluation of concept coverage against a harmonized English Silver Standard Corpus built from the Mantra project partners was performed. For each corpus there are 6 different annotations that are harmonized into a centroid-based Silver Standard using a voting threshold of 3. The technical details of the centroid approach for the partner annotations as well as a detailed evaluation of the effect of different voting thresholds can be found in[18]

### 3.4 Participation and Contributions

Seven groups participated into the CLEF-ER challenge and contributed annotated corpora for the evaluation. Table 5 gives an overview on the approach that

has been tested and links the system description to the performance of the tested solutions. As can be seen in tables 3 and 4 the participants contributed different numbers of annotated corpora and in general did not cover all languages. Spanish was the most popular language, i.e. the Spanish corpora have been annotated by the largest number of participants, and the largest number of submissions were linked to Spanish. French was a little bit more popular than German and the least contributions – as expected – were delivered for Dutch. These figures are relevant for the evaluation of the challenge, since a larger number of contributions leads to a larger set of annotated corpora that can be considered for the generation of a SSC in a given language.

Four of seven groups (A, C, D, and F) did apply methods that are linked to statistical machine translation or multi-lingual word alignment. Almost all groups used publicly available resources such as UMLS, Wordnet, Wikipedia and most groups also applied lexical lookup solutions or indexing of the terminological resources. Two groups translated the terms through public resources (i.e. BabelNet, group E) or with the Google translate infrastructure (group D). Altogether, the heterogeneity of the used solutions was high, and it became clear that the CLEF-ER challenge profits from machine translation solutions, although the challenge was announced as an entity recognition task.

Not all submissions were considered to be included for the generation of the SSC, which is based on the annotated corpora by the MANTRA project partners and the CLEF-ER participants (cf. tbl. 4). It is important to avoid that one or several participants dominate the outcome of the SSC by contributing a large number of annotated corpora. Therefore, the participants have been asked to point out one corpus that should server as their contribution to the challenge.

## 4 Resource and Evaluation

### 4.1 Silver Standards, Multilingual Documents

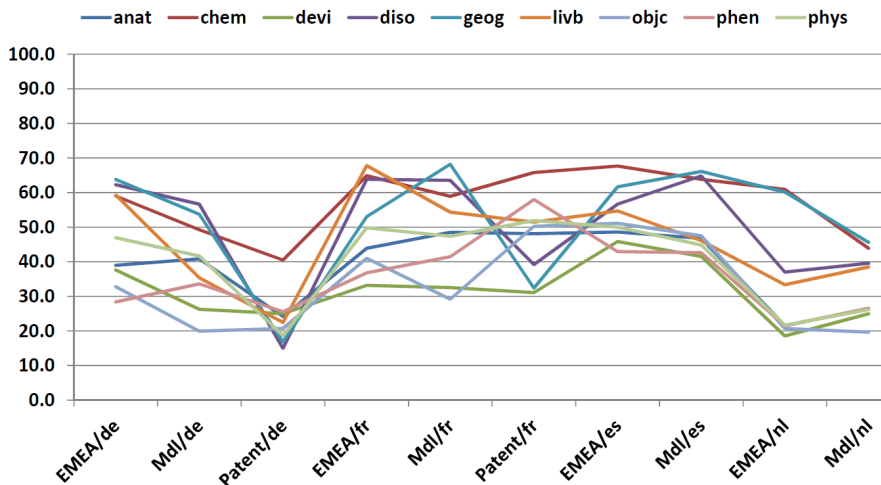
Table 4 gives an overview on the contributions to the monolingual SSCs. For each corpus and for all covered languages, one SSC has been produced from the MANTRA project partners' contributions to enable task A evaluation, i.e. the mention evaluation, and for the task B evaluation, i.e. CUI assignment. Only for the variant of the task B evaluation, where the trivial annotations have been removed (the “deannotated” corpus) the participants' contributions have been added as well.

In total 36 contributions have been received as part of the challenge, and another 25 annotated corpora have been provided from the MANTRA project partners prior to the challenge termination. Two participants contributed 10 annotated corpora, one for each language and for each corpus, and the other participants provided a smaller number of annotated corpora;

### Evaluation of Challenge Contributions

Two different tasks (and evaluations) have been suggested to the participants. In the evaluation A, the entity annotations are compared against an SSC to





**Fig. 1. (Precision, recall and F1-measure for the Evaluation B):** All contributions have been evaluated concerning their assignment of the CUI. The evaluation was performed against the English SSC. The figure shows the average precision, recall and F1-measure of all solutions. Note that the both values for precision and recall are above the F1-measure for the EMEA/es corpus, since the diagram shows average figures for all annotation solutions together.

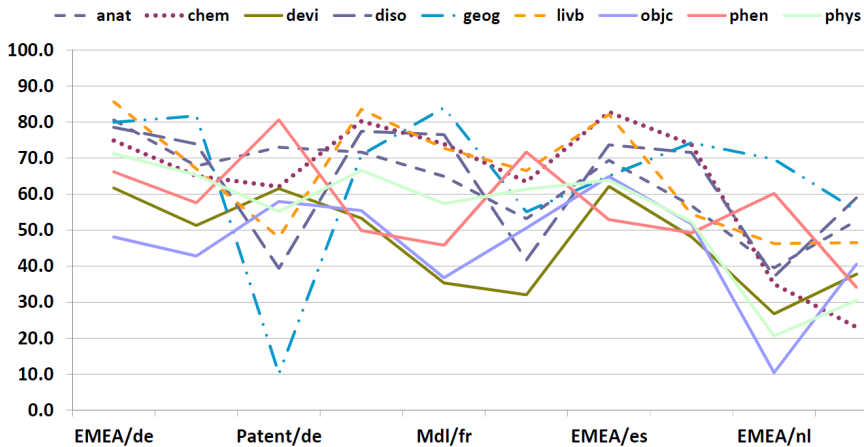
measure the boundary agreement of the participants against the SSC, where the SSC has been produced from annotated contributions from the MANTRA project partners.

In the evaluation B, the CUI assignment in the annotated corpus is evaluated against the prepared English SSC. In this task the participants have to assign the right CUI to a text stretch, which could be the complete unit of the parallel corpus, and the evaluation also does not consider any annotations in the text, but only evaluates against the correct assignment of a CUI to a unit.

Evaluation A and B are complementary in the sense that the boundary annotation (evaluation A) may give the correct mention of an entity, but the entity may still belong to different CUIs, and the correct CUI or mention normalisation may identify the correct concept (or entity), but the assignment to a particular stretch of text is left open.

The first task has been approached in a number of challenges, but not yet in the multi-lingual case covering a large amount of documents. The second task is typical for the biomedical domain and targets the normalisation of entities in non-English documents. This task has not yet been addressed in the multilingual case covering a large amount of parallel documents.

*CUI Assignment (Task B).* The participants had to produce annotations for their preferred corpus in their preferred languages, which should cover at least



**Fig. 2. (Evaluation B for semantic groups):** The average F1-measure across all contributing systems has been calculated per semantic group of the annotations

one non-English language. The annotations had to comprise the assignment of a CUI to the entity mention. As can be seen from the system descriptions (cf. tbl. 5), the participants used different kinds of technologies including the translation of the terminology, the alignment and matching of concept mentions, and the translation of the corpus with the identification of corresponding concepts. The comparison of the CUI assignment in the non-English corpus against the English SSC formed the first evaluation and led to the following results (cf. fig. 1). The F1-measure performance over all contributing systems is better for Medline than for EMEA in all languages except for German, and for all languages the precision is higher in Medline than in EMEA. The F1-measure performance for the German patents (19 %) is a lot lower than for the other two corpora in German, and to a certain extend lower for the annotation of the patents in French in comparison to the other two corpora in French. This result indicates that the identification of entities and on concepts in patent documents is more complex than in the scientific biomedical literature, but the F1-measure for the other corpora ranges between 38 % and 48 %.

Table 6 shows the results for individual participants. The performance of the different solutions shows high heterogeneity, i.e. some entity types are identified well from selected solutions, but not in general across the corpus. As explained before, the annotation of French and Spanish text led to better performances than the annotation of German texts.

*CUI Assignment per Semantic Group (Task B).* The CUIs of the annotations can be categorized according to the semantic group that has been assigned to the CUIs. This grouping can be used to differentiate the performances according to the semantic groups and to give a more detailed analysis on the annotation of

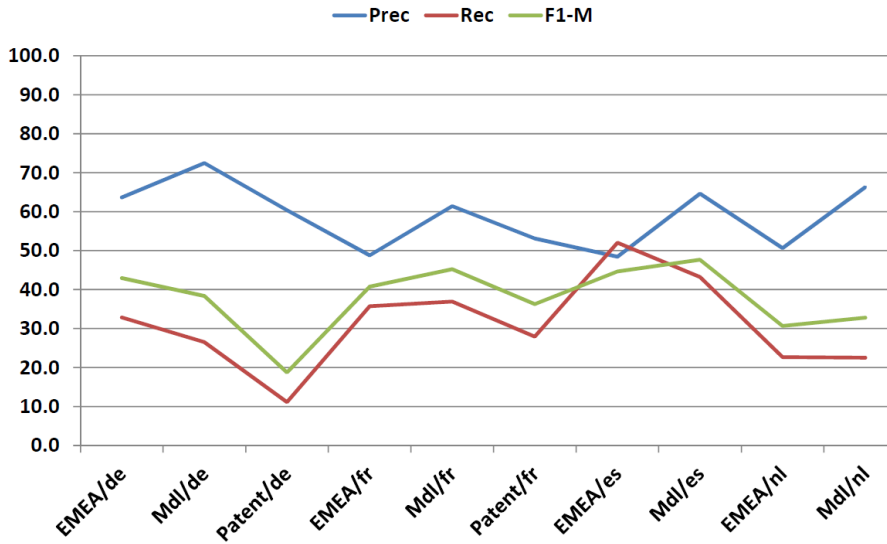
**Table 6. (Evaluation B, F1-Measure, challenge’s participants):** The table to the left shows the individual F1-measure performances of the participants in the evaluation B on the EMEA corpus and on the right for the Medline titles

		ANAT	CHEM	DEVI	DISO	GEOG	LIVB	OBJC	PHEN	PHYS			ANAT	CHEM	DEVI	DISO	GEOG	LIVB	OBJC	PHEN	PHYS
EMEA/de	D1	40.0	72.3	38.2	59.8	77.5	63.5	43.2	37.8	50.3	Med/de	D1	32.9	43.0	26.8	48.4	55.8	42.7	22.1	29.4	42.8
	E1	39.5	53.3	44.3	60.2	56.0	59.5	35.6	26.3	47.8		E1	35.2	46.3	22.1	53.2	54.5	35.4	24.8	29.9	36.8
	F1	38.9	56.8	37.1	65.2	58.5	59.0	27.2	23.8	45.2		F1	36.5	49.9	24.2	58.2	51.7	36.3	18.8	26.9	38.2
	F2	37.4	53.6	30.8	63.7	63.2	55.1	25.1	25.6	44.4		F2	36.7	39.8	19.9	57.0	50.9	37.0	15.5	29.5	38.6
EMEA/es	A1	52.5	48.7	47.4	34.4	78.3	29.5	59.1	60.7	54.8	Med/es	G1	51.7	58.0	32.1	61.2	54.9	29.3	19.0	42.9	46.8
	A2	58.4	51.7	51.1	42.8	80.9	29.8	67.6	62.9	60.2		G2	52.1	58.7	32.5	61.8	54.6	31.2	19.3	42.9	46.9
	B1	52.8	77.4	54.4	70.3	61.7	76.5	48.9	47.6	58.9		B1	50.6	65.2	50.4	77.6	72.1	63.8	59.5	47.8	54.2
	C1	30.5	67.3	8.2	66.6	20.9	61.1	33.8	15.8	35.7		C1	12.5	42.9	0.4	45.1	44.4	11.1	0.9	0.0	0.7
EMEA/fr	C2	31.8	66.7	6.8	66.7	18.3	64.9	35.6	11.5	35.8	Med/fr	C2	11.9	51.7	1.3	51.2	41.8	30.5	2.4	1.6	7.0
	D1	47.2	81.2	56.3	61.6	79.2	70.2	53.0	51.4	56.1		D1	49.2	64.9	50.5	68.1	80.1	57.4	70.3	58.2	56.2
	F1	59.4	77.6	75.1	61.8	77.8	62.5	57.6	49.0	57.7		F1	58.6	68.1	63.2	66.9	76.7	51.9	69.4	55.6	57.8
	F2	56.2	70.8	67.4	48.7	76.0	43.1	53.4	44.7	40.5		F2	56.8	70.5	47.6	63.4	68.5	44.5	46.5	48.7	44.9
EMEA/fr	G1	43.0	55.9	10.1	61.2	43.9	67.3	36.5	10.6	37.3	Med/fr	G1	72.4	76.8	59.3	78.4	79.6	55.7	64.3	68.1	72.4
	C2	36.6	56.2	13.6	61.2	19.6	63.5	37.5	31.8	43.6		G2	62.1	70.6	59.5	68.1	65.7	53.9	66.5	60.9	65.4
	D1	45.6	81.3	58.3	64.8	80.7	71.1	63.6	69.6	64.1		C1	18.5	37.0	0.4	46.8	52.4	37.8	2.6	0.5	5.8
	E1	45.9	60.7	47.3	68.5	50.2	71.2	43.5	42.1	52.1		C2	26.1	45.5	0.7	52.9	56.9	44.2	4.8	1.4	16.6
EMEA/fr	F1	47.5	67.9	45.2	68.1	67.1	71.5	31.4	33.3	50.9	Med/fr	D1	54.0	64.9	62.3	65.3	79.5	53.7	44.7	67.5	62.8
	F2	44.9	67.2	24.4	59.2	56.5	62.2	33.6	33.4	51.0		E1	54.6	66.2	40.6	73.8	74.1	58.7	45.2	42.1	51.5
												F1	57.5	59.0	47.4	72.7	70.6	62.5	36.2	46.7	54.3
												F2	56.1	59.8	16.2	63.1	59.4	58.3	35.8	41.5	52.6
											G1	62.7	71.8	47.7	68.8	77.7	60.7	31.9	68.9	70.3	
											G2	58.7	67.3	44.8	64.9	74.7	58.9	32.0	62.9	64.9	

the different corpora (cf. fig. 2). From this analysis it is possible to derive that chemical entities (‘chem’) and living beings (‘livb’) can be identified at a better rate than the entities from the other groups. In the case of the patent corpus, the identification of the chemical entities can be reached at a rate which is high in comparison to the entities from the other semantic groups. Furthermore, it becomes clear that anatomical entities (‘anat’) and disease & disorder (‘diso’) can be well recognized in Medline abstracts and EMEA drug guidelines in contrast to patents. Overall, the presented results indicate that the identification of the concepts and entities can be achieved at a higher performance level in French and Spanish in contrast to German and Dutch.

*Mention Evaluation (Task A).* The evaluation of the mention annotations has been performed against a SSC that has been generated from the annotated corpora contributed by the MANTRA project partners and the participants of the CLEF-ER challenge. The SSC has been generated as described in section 3.3 and a TP is any mention annotation that nests a centroid in the SSC. This can be interpreted as the identification of a portion of the entity representation that has a high agreement between the different annotation solutions. Every annotated corpus has been evaluated against the appropriate SSC, i.e. the same corpus annotated in the same language. (cf. fig. 3)

The performance evaluation indicates that – with a few exceptions – the annotation of the EMEA documents can be achieved with better results than the annotation of the Medline abstracts, or the patent documents. This result is true for all languages except for Dutch. The mention annotation of the patent documents shows a mixed picture, since in general the performance for the annotation in German and French resembles the performance produced on the other



**Fig. 3. (Evaluation of mentions):** The figure shows the average of the F1-measure across all contributing systems for the mention annotation

two corpora, and comparing the different semantic groups it becomes clear that for selected groups the performance is good (e.g., phenotype – ‘phen’, ‘anat’, ‘livb’ and ‘chem’).

Again, table 7 shows the results for individual participants, but now for the mention annotation. The measured performances are similar to the results from the task B evaluation (cf. tbl. 6). On the other side, the performances on the German corpora has improved for the mention annotation in comparison to the CUI annotations.

*CUI Assignment, Non-trivial Cases (Task A).* Finally we ignored all the trivial assignments of a CUI to the non-English documents, where a ‘trivial’ assignment is determined by the fact that the non-English term is already known in the terminological resources as a synonym to a given English term. This evaluation uses a smaller number of term candidates in the English SSC and focuses the evaluation towards those terms where new term candidates – in comparison to the original terminological resources – can be expected. The performances of the annotation solutions against this set of candidate terms (cf. fig. 4) shows a different picture than the previous analysis (cf. fig. 2). Now the performances of the annotation solutions in French and Spanish are now lower than previously and do not differ much from the annotation solutions in German. It is remarkable that the annotations for the different semantic groups are in a similar range, e.g. for nl, de and es on Medline and EMEA, and it becomes again visible that the

**Table 7. (Evaluation A, F1-Measure, challenge’s participants):** Similar to the the previous table 6, this table shows the F1-measure performances of the individual solutions in the task A evaluation, i.e. annotation of entity mentions in the text

	Contr.	anat	chem	devi	diso	geog	livb	objc	phen	phys		Contr.	anat	chem	devi	diso	geog	livb	objc	phen	phys
EMEA/de	D1	63.2	50.1	52.5	73.1	70.2	81.6	35.9	64.1	63.1	Med/de	D1	67.5	64.6	34.7	70.9	75.4	64.7	43.7	66.5	70.1
	E1	88.0	81.4	69.3	88.1	87.6	87.2	56.4	52.2	68.9		E1	89.1	89.7	85.0	88.4	95.4	83.5	49.7	62.2	79.3
	F1	90.4	83.2	77.9	83.4	79.1	79.2	74.5	59.5	76.4		F1	82.8	79.8	83.2	80.3	86.0	76.1	62.3	58.3	78.6
	F2	80.9	85.0	47.2	69.8	83.1	94.8	25.6	89.2	76.7		F2	56.8	43.1	44.6	71.9	87.4	88.3	44.0	61.2	63.3
EMEA/es	A1	75.4	92.2	74.8	78.5	78.0	96.8	67.5	60.0	75.2	Med/es	G1	56.3	56.6	29.8	66.4	73.4	44.3	28.7	49.1	50.6
	A2	81.1	94.7	88.4	91.5	81.3	97.4	72.3	63.7	83.2		G2	54.6	56.6	30.7	65.9	72.9	45.7	28.7	48.4	50.1
	B1	76.1	82.0	72.6	75.3	66.5	86.9	68.3	60.5	69.2		B1	72.0	78.9	71.4	85.0	78.0	77.9	78.1	60.5	81.0
	C1	45.1	72.7	12.1	75.4	19.6	69.0	51.7	8.7	50.5		C1	24.8	58.6	0.7	56.3	50.8	16.3	0.9	0.2	1.1
EMEA/fr	C2	47.5	73.0	10.2	75.8	20.9	73.2	51.1	5.6	50.9	Med/fr	C2	26.0	70.4	2.6	65.6	49.3	45.7	2.8	1.0	11.9
	D1	62.3	81.3	74.8	66.1	83.4	80.2	55.9	55.5	65.3		D1	73.2	81.5	65.3	80.9	85.6	77.9	79.8	72.2	72.4
	F1	86.7	85.6	83.7	73.4	86.7	70.1	78.6	89.1	75.5		F1	71.3	75.3	77.7	70.7	89.3	48.5	80.7	77.2	71.7
	F2	81.1	81.0	80.6	53.8	84.0	83.6	73.7	80.6	41.5		F2	67.2	80.7	63.0	64.5	77.6	59.1	47.1	67.3	48.9
	C1	70.4	89.5	16.8	86.0	80.9	84.5	42.0	13.2	54.0		G1	59.1	71.1	49.9	74.0	78.9	52.7	59.3	55.4	64.4
	C2	67.2	86.2	36.1	81.4	48.2	79.8	37.2	43.8	55.8		G2	60.8	73.8	54.7	75.7	84.7	57.7	65.4	60.0	66.6
	D1	71.6	64.1	86.5	71.7	80.3	77.0	33.8	53.7	58.6		C1	34.9	77.0	0.5	71.8	83.8	63.3	8.9	0.5	9.5
	E1	79.8	86.6	78.3	82.2	65.7	90.0	61.9	61.0	83.9		C2	49.5	88.2	1.3	81.0	88.7	72.3	18.1	1.9	26.9
	F1	74.0	76.7	76.6	75.1	83.1	86.6	74.6	63.0	69.2		D1	66.9	71.3	54.8	74.3	83.5	66.9	36.5	67.9	64.5
	F2	67.0	78.6	25.4	68.5	67.2	83.7	83.2	64.7	78.4		E1	78.4	89.2	58.3	85.3	89.4	79.6	46.0	52.9	76.7
												F1	83.1	65.6	64.8	81.7	89.4	78.0	56.3	72.6	76.5
												F2	79.6	67.2	13.0	79.1	71.8	76.6	51.0	57.4	77.3
	G1										G1	64.0	66.1	44.3	69.0	83.0	72.0	38.1	56.3	63.3	
	G2										G2	63.4	67.1	45.7	70.1	82.5	73.2	39.4	57.3	64.4	

annotation of EMEA can be achieved at higher performance levels than the annotation of Medline.

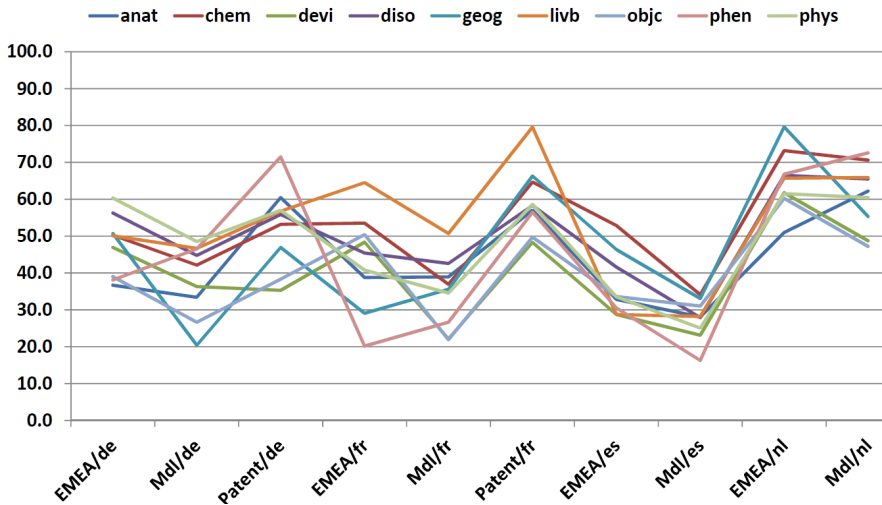
## 5 Conclusions

The CLEF-ER challenge has targeted the task of entity recognition in multi-lingual and parallel documents. The approach is based on the development of an SSC, which would be made available in the English version for the participants of the challenge, and – later on – for the non-English corpora for any further evaluation of the participants’ contributions. At the current state, only preliminary results are available indicating that the task requires the integration of different technologies to achieve ER in multilingual documents. Different approaches have been tested, but further investigation is required to state, which solutions perform best on the given task.

Nonetheless, it becomes clear that evaluation A (“monolingual mention evaluation”) as well as evaluation B (“cross-lingual concept evaluation”) gives us an indication of how well an individual contribution complies with the harmonized contribution where the harmonized contribution (“SSC”) is composed of at least 3 contributions and their agreement induced by the e-centroid method.

On the other side, the analysis shows that the French corpora allow a higher agreement with the SSC than the German and the Spanish corpora. For the Dutch corpora, a high agreement has been achieved through the annotation solutions, but this is biased, since only a very small number of annotated corpora was available.

In the next phase, the contributions from the participants will be analysed for their individual performances on the challenge tasks. Furthermore, the



**Fig. 4. (Evaluation B for semantic groups after term reduction):** Similarly to the previous figure (cf. fig. 2), the average F1-measure of all contributing systems for each semantic group has been calculated, but in contrast to the previous figure the evaluation only considers a subset of all annotations. This subset is specific to novel findings of mentions that are linked to the mention in the parallel English document, but is not confirmed by a synonym in the terminological resource.

MANTRA project partners will mine the contributions for novel terms and will generate a gold standard corpus to evaluate the contributions of the participants on a smaller scale and against the opinion of an expert.

**Acknowledgement.** This work was funded by the European Commission STREP grant number 296410 ("Mantra", FP7-ICT-2011-4.1).

## References

1. Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J.S., Roberts, I., Setzer, A., Tapuria, A., et al.: The CLEF corpus: semantic annotation of clinical text. In: AMIA Annual Symposium Proceedings, vol. 2007, p. 625. American Medical Informatics Association (2007)
2. Lussier, Y.A., Shagina, L., Friedman, C.: Automating icd-9-cm encoding using medical language processing: A feasibility study. In: Proceedings of the AMIA Symposium, p. 1072. American Medical Informatics Association (2000)
3. Catarci, T., Ferro, N., Forner, P., Hiemstra, D., Karlgren, J., Penas, A., Santucci, G., Womser-Hacker, C.: CLEF 2012: information access evaluation meets multilinguality, multimodality, and visual analytics. ACM SIGIR Forum 46, 29–33 (2012)
4. Roda, G., Tait, J., Piroi, F., Zenz, V.: CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 385–409. Springer, Heidelberg (2010)

5. Krallinger, M., Leitner, F., Rodriguez-Penagos, C., Valencia, A.: Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology* 9(suppl. 2), S4 (2008), <http://genomebiology.com/2008/9/S2/S4>
6. Morgan, A., Lu, Z., Wang, X., Cohen, A., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H.H., Torres, R., Krauthammer, M., Lau, W., Liu, H., Hsu, C.N., Schuemie, M., Cohen, K.B., Hirschman, L.: Overview of BioCreative II gene normalization. *Genome Biology* 9(suppl. 2), S3 (2008), <http://genomebiology.com/2008/9/S2/S3>
7. Cohen, K.B., Demner-Fushman, D., Ananiadou, S., Pestian, J., Tsujii, J., Webber, B. (eds.): Proceedings of the BioNLP 2009 Workshop. Association for Computational Linguistics, Boulder (2009), <http://www.aclweb.org/anthology/W09-13>
8. Rebholz-Schuhmann, D., Yepes, A.J., Mulligen, E.M.V., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E., Hahn, U.: CALBC silver standard corpus. *Journal of Bioinformatics and Computational Biology* 8, 163–179 (2010)
9. Rebholz-Schuhmann, D., Jimeno-Yepes, A., Li, C., Kafkas, S., Lewin, I., Kang, N., Corbett, P., Milward, D., Buyko, E., Beisswanger, E., Hornbostel, K., Kouznetsov, A., Witte, R., Laurila, J., Baker, C., Kuo, C.J., Clematide, S., Rinaldi, F., Farkas, R., Móra, G., Hara, K., Furlong, L., Rautschka, M., Lara Neves, M., Pascual-Montano, A., Wei, Q., Collier, N., Mahbub Chowdhury, M.F., Lavelli, A., Berlanga, R., Morante, R., Van Asch, V., Daelemans, W., Marina, J., van Mulligen, E., Kors, J., Hahn, U.: Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *J. Biomedical Semantics* 2(suppl. 5), S11 (2011)
10. Hersh, W., Voorhees, E.: TREC genomics special issue overview. *Inf. Retr. Boston* 12, 1–15 (2009)
11. Lu, Z.: PubMed and beyond: a survey of web tools for searching biomedical literature. *Database (Oxford)*, 2011:baq036 (2011)
12. Rebholz-Schuhmann, D., Clematide, S., Rinaldi, F., Kafkas, S., van Mulligen, E.M., Bui, C., Hellrich, J., Lewin, I., Milward, D., Poprat, M., Jimeno-Yepes, A., Hahn, U., Kors, J.A.: Multilingual semantic resources and parallel corpora in the biomedical domain: the CLEF-ER challenge. In: Proceedings CLEF Conference, vol. 2013 (2013)
13. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32, D267–D270 (2004)
14. Brown, E.G., Wood, L., Wood, S.: The medical dictionary for regulatory activities (MedDRA). *Drug Safety* 20(2), 109–117 (1999)
15. Stearns, M.Q., Price, C., Spackman, K.A., Wang, A.Y.: SNOMED clinical terms: overview of the development process and project status. In: Proceedings of the AMIA Symposium, vol. 662, American Medical Informatics Association (2001)
16. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255 (2007)
17. Lewin, I., Kafkas, S., Rebholz-Schuhmann, D.: Centroids: Gold standards with distributional variations. In: Proceedings of the Language Resources Evaluation Conference, Istanbul, Turkey (2012)
18. Lewin, I., Clematide, S.: Deriving the Mantra Silver Standard. In: Proceedings CLEF Conference, vol. 2013 (2013)

# Author Index

- Amigó, Enrique 333  
Angelini, Marco 29  
Azarbondyad, Hosein 93
- Bellot, Patrice 269  
Bensalem, Imene 53  
Berlanga, Rafael 120, 167  
Beyer, Anna 282  
Bogers, Toine 192  
Bonnet, Pierre 250  
Bui, Chinh 353  
Busse, Matthias 282
- Cabrio, Elena 321  
Caputo, Barbara 250  
Carrillo-de-Albornoz, Jorge 41, 333  
Cazorla, Miguel 250  
Chapman, Wendy W. 212  
Chikhi, Salim 53  
Chugur, Irina 333  
Cimiano, Philipp 321  
Clematide, Simon 353  
Clough, Paul 59  
Corujo, Adolfo 333
- Dehghani, Nazanin 71  
de L. Pertile, Solange 150  
de Rijke, Maarten 333  
Doucet, Antoine 269
- Eibl, Maximilian 13  
Elhadad, Noemie 212  
El-Sonbaty, Yasser 100  
Endrullis, Stefan 1  
Ezzeldin, Ahmed Magdy 100
- Faili, Hesham 93  
Färber, Michael 124  
Ferro, Nicola 29, 192  
Forcada, Rafael 167  
Forner, Pamela 303
- Ganguly, Debasis 108  
Garcia Varea, Ismael 250  
Geva, Shlomo 269
- Goëau, Hervé 250  
Goeuriot, Lorraine 212  
Gollub, Tim 282  
Gonzalo, Julio 333  
Goodale, Paula 59  
Goodwin, Travis 155  
Gschwandtner, Manfred 81  
Gurajada, Sairam 269
- Hahn, Udo 353  
Hall, Mark Michael 17, 192  
Hanbury, Allan 81, 232  
Harabagiu, Sanda M. 155  
Hellrich, Johannes 353  
Hovy, Eduard 303
- Jimeno-Yepes, Antonio 120, 353  
Joly, Alexis 250  
Jones, Gareth J.F. 108, 212
- Kafkas, Senay 353  
Kamps, Jaap 269  
Kazai, Gabriella 269  
Kelly, Liadh 212  
Kholief, Mohamed Hamed 100  
Kim, Se-Jong 179  
Koolen, Marijn 269  
Kors, Jan A. 353  
Kritz, Marlene 81
- Larsen, Birger 1  
Lee, Jong-Hyeok 179  
Leiva, Luis A. 143  
Leveling, Johannes 108, 212  
Lewin, Ian 353  
Lopez, Vanessa 321  
Lupu, Mihai 232
- Malak, Piotr 192  
Martín, Tamara 333  
Martinez, David 212  
Martínez-Gómez, Jesus 250  
Masiero, Ivano 192  
Meij, Edgar 333  
Milward, David 353



- Mirsarraf, Mohammad Reza 71  
 Mishra, Arunav 269  
 Molina, Alejandro 75  
 Morante, Roser 303  
 Moreira, Viviane P. 150  
 Moriceau, Véronique 269  
 Mothe, Josiane 269  
 Mowery, Danielle L. 212  
 Müller, Henning 1, 250  
 Museros, Lledó 167  
  
 Ngonga Ngomo, Axel-Cyrille 321  
  
 Paredes, Roberto 143, 250  
 Pawłowski, Adam 192  
 Peláez-Moreno, Carmen 41  
 Peñas, Anselmo 303  
 Pérez-Catalán, María 120, 167  
 Petras, Vivien 192  
 Piroi, Florina 232  
 Poprat, Michael 353  
 Potthast, Martin 282  
 Pradhan, Sameer 212  
 Preminger, Michael 269  
  
 Rahm, Erhard 1  
 Rangel, Francisco 282  
 Rauber, Andreas 136  
 Rebholz-Schuhmann, Dietrich 120, 353  
 Rettinger, Achim 124  
 Rinaldi, Fabio 353  
 Rodrigo, Álvaro 303  
 Rosso, Paolo 53, 150, 282  
  
 Sachs, Alexander 81  
 Salanterä, Sanna 212  
 Samwald, Matthias 81  
 SanJuan, Eric 75, 269  
 Santucci, Giuseppe 29  
 Savova, Guergana 212  
  
 Savoy, Jacques 192  
 Schenkel, Ralf 269  
 Shakery, Azadeh 93  
 Silvello, Gianmaria 29  
 South, Brett R. 212  
 Spina, Damiano 333  
 Stamatatos, Efstathios 282  
 Stefanov, Veronika 81  
 Stein, Benno 282  
 Suominen, Hanna 212  
 Sutcliffe, Richard 303  
  
 Tadić, Marko 124  
 Tannebaum, Wolfgang 136  
 Tannier, Xavier 269  
 Theobald, Martin 269  
 Thomee, Bart 250  
 Toms, Elaine 17, 192  
 Torres-Moreno, Juan-Manuel 75  
 Trappett, Matthew 269  
 Tsikrika, Theodora 1  
  
 Unger, Christina 321  
  
 Valverde-Albacete, Francisco José 41  
 van Mulligen, Erik M. 353  
 Velupillai, Sumithra 212  
 Villegas, Mauricio 143, 250  
  
 Walter, Sebastian 321  
 Wang, Qiuyue 269  
 Wang, Xiaojie 104  
 Wang, Xuwen 104  
 Wilhelm-Stein, Thomas 13  
  
 Zellhöfer, David 250  
 Zhang, Lei 124  
 Zhang, Qiang 104  
 Zuccon, Guido 212