Coling 2010

23rd International Conference on Computational Linguistics

Proceedings of the

4th Workshop on Cross Lingual Information Access

28 August 2010
Beijing International Convention Center

Produced by
Chinese Information Processing Society of China

To order the CD of Coling 2010 and its Workshop Proceedings, please contact:

Chinese Information Processing Society of China
No.4, Southern Fourth Street
Haidian District, Beijing, 100190
China
Tel: +86-010-62562916
Fax: +86-010-62562916
cips@iscas.ac.cn

Introduction

Welcome to the Coling Workshop on *Cross Lingual Information Access*.

Cross-lingual information access (CLIA) is concerned with technologies and applications that enable people to freely access information expressed in any language which may differ from the query language. As the web has grown to include rich contents in many different languages, and with rapid globalization, there is a growing demand for CLIA. Ordinary netizens who surf the Internet for special information and communicate in social networks, global companies which provide multilingual services to their multinational customers, governments who aim to lower the barriers to international commerce and collaboration and homeland security are in need of CLIA. This has triggered vigorous research and development activity in CLIA. This workshop is the fourth in a series of workshops and aims to address the need of CLIA. The previous three workshops were held during IJCAI 2007 in Hyderabad, IJCNLP 2008 in Hyderabad, and NAACL 2009 in Colorado.

In this workshop, in addition to Cross-lingual Information Retrieval (CLIR), the focus is on multi-lingual information extraction, information integration, summarization and other key technologies that are useful for CLIA. The workshop aims to bring together researchers from a variety of fields such as information retrieval, computational linguistics, machine translation, and practitioners from government and industry to address the issue of information need of multi-lingual societies. This workshop also aims to highlight and emphasize the contributions of Natural Language Processing (NLP) and Computational Linguistics to CLIA, in addition to the previously better represented viewpoint from Information Retrieval.

The workshop received a total of fourteen submissions, out of which the proceedings includes ten papers covering various aspects of this field. There are two papers on corpus acquisition. The papers by Pattabhi *et al.* and Lejune *et al.* focus on acquiring multilingual documents on various topics. There are three papers on bilingual lexicon acquisition. The papers by Okita *et al.* and Chatterjee *et al.* propose methods for word alignment and lexicon extraction from parallel and comparable corpora, while the paper by Rapp *et al.* proposes to learn dictionaries from monolingual corpora containing foreign words. Tang *et al.* do named entity translation for cross language question answering applications by combining a number of different sources, namely, machine translation, online encyclopedia and web documents. Falaise *et al.* use a light ontology to extraxt content from multilingual texts and user requests associated with images. Litvak *et al.* explore the performance of summarization methods across two languages. The paper by Vachchani *et al.* presents studies on pseudo relevance feedback utilizing multiple assisiting languages. Hajlaoui *et al.* discuss multilinguization and personalization in natural language based systems.

Besides these contributed papers, the workshop features two invited talks. Professor Pushpak Bhattacharya will speak on word sense disambiguation and information retrieval. Dr Tetsuya Sakai will speak on multilinguality at NTCIR.

With this gamut of topics, we look forward to a lively exchange of ideas in the workshop.

We take this opportunity to thank all the members of the Program Committee for their timely and insightful reviews, to the two invited speakers for kindly agreeing to speak at the workshop, the authors

who submitted their work to this workshop and all the participants of this workshop.

Organizers, 4th Workshop on Cross Lingual Internet Access at COLING 2010, Beijing

Min Zhang, Sudeshna Sarkar, Raghavendra Udupa and Adam Lopez

Organizers:

Min Zhang, Institute for Infocomm Research(Singapore)
Sudeshna Sarkar, Indian Institute of Technology Kharagpur(India)
Raghavendra Udupa, Microsoft Research(India)
Adam Lopez, The University of Edinburgh(United Kingdom)

Program Committee:

Eneko Agirre, University of the Basque Country(Spain)
Ai Ti Aw, Institute for Infocomm Research(Singapore)
Sivaji Bandyopadhyay, Jadavpur University(India)
Pushpak Bhattacharyya, IIT Bombay(India)
Nicola Cancedda, Xerox Research Centre(France)
Patrick Saint Dizier, IRIT, Universite Paul Sabatier(France)
Nicola Ferro, University of Padua(Italy)
Guohong Fu, Heilongjiang University(China)
Cyril Goutte, National Research Council of Canada(Canada)
A Kumaran, Microsoft Research(India)
Gareth Jones, Dublin City University(Ireland)
Joemon Jose, University of Glasgow(United Kingdom)
Gina-Anne Levow, National Centre for Text Mining(United Kingdom)
Haizhou Li, Institute for Infocomm Research(Singapore)
Qun Liu, ICT/CAS(China)
Ting Liu, Harbin Institute of Technology(China)
Paul McNamee, Johns Hopkins University(USA)
Yao Meng, Fujitsu R&D Center Co. Ltd.(China)
Mandar Mitra, ISI Kolkata(India)
Doug Oard, University of Maryland, College Park(USA)
Carol Peters, Istituto di Scienza e Tecnologie dell'Informazione(Italy)
Maarten de Rijke, University of Amsterdam(Netherlands)
Paolo Rosso, Technical University of Valencia(Spain)
Hendra Setiawan, University of Maryland(USA)
L Sobha, AU-KBC, Chennai(India)
Rohini Srihari, University at Buffalo, SUNY(USA)
Ralf Steinberger, European Commission Joint Research Centre(Italy)
Le Sun, Institute of Software, CAS(China)
Chew Lim Tan, National University of Singapore(Singapore)
Vasudeva Varma, IIIT Hyderabad(India)
Thuy Vu, Institute for Infocomm Research(Singapore)
Haifeng Wang, Baidu(China)
Yunqing Xia, TsingHua University(China)
Deyi Xiong, Institute for Infocomm Research(Singapore)

Guodong Zhou, SooChow University(China)
Chengqing Zong, Institute of Automation, CAS(China)

Invited Speakers:

Pushpak Bhattacharya, Indian Institute of Technology Bombay(India)
Tsetsuya Sakai, Microsoft Research Asia(China)

Table of Contents

Conference Program

Saturday, August 28, 2010

8:35–8:45 Opening Remarks by Sudeshna Sarkar, Min Zhang, Adam Lopez and Raghavendra Udupa

8:45–9:40 Invited Talk 1:

Word Sense Disambiguation and IR
Pushpak Bhattacharyya

Invited Talk 2

11:00–11:55 *Multiliguality at NTCIR, and moving on ...*
Tetsuya Sakai

9:40–10:05 *Filtering news for epidemic surveillance: towards processing more languages with fewer resources*
Gael Lejeune, Antoine Doucet, Roman Yangarber and Nadine Lucas

10:05–10:30 *How to Get the Same News from Different Language News Papers*
T Pattabhi R K Rao and Sobha Lalitha Devi

10:30–11:00 Morning Break

11:55–12:20 *The Noisier the Better: Identifying Multilingual Word Translations Using a Single Monolingual Corpus*
Reinhard Rapp, Michael Zock, Andrew Trotman and Yue Xu

12:20–13:50 Lunch

13:50–14:15 *Multi-Word Expression-Sensitive Word Alignment*
Tsuyoshi Okita, Alfredo Maldonado Guerra, Yvette Graham and Andy Way

14:15–14:40 *Co-occurrence Graph Based Iterative Bilingual Lexicon Extraction From Comparable Corpora*
Diptesh Chatterjee, Sudeshna Sarkar and Arpit Mishra

14:40–15:05 *A Voting Mechanism for Named Entity Translation in EnglishChinese Question Answering*
Ling-Xiang Tang, Shlomo Geva, Andrew Trotman and Yue Xu

Saturday, August 28, 2010 (continued)

15:05–15:30 *Ontology driven content extraction using interlingual annotation of texts in the OMNIA project*
Achille Falaise, David Rouquet, Didier Schwab, Hervé Blanchon and Christian Boitet

15:30–16:00 Afternoon Break

16:00–16:25 *Towards multi-lingual summarization: A comparative analysis of sentence extraction methods on English and Hebrew corpora*
Marina Litvak, Mark Last, Slava Kisilevich, Daniel Keim, Hagay Lipman and Assaf Ben Gur

16:25–16:50 *More Languages, More MAP?: A Study of Multiple Assisting Languages in Multilingual PRF*
Vishal Vachhani, Manoj Chinnakotla, Mitesh Khapra and Pushpak Bhattacharyya

16:50–17:15 *Multilinguization and Personalization of NL-based Systems*
Najeh Hajlaoui and Christian Boitet

17:15–17:25 Closing