NAACL HLT 2009

The Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3)

Proceedings of the Workshop

June 4, 2009
Boulder, Colorado

Production and Manufacturing by
Omnipress Inc.
2600 Anderson Street
Madison, WI 53707
USA

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-932432-33-6

Introduction

The development of digital and online information repositories is creating many opportunities and also new challenges in information retrieval. The availability of online documents in many different languages makes it possible for users around the world to directly access previously unimagined sources of information. However in conventional information retrieval systems the user must enter a search query in the language of the documents in order to retrieve it. This requires that users can express their queries in those languages in which the information is available and can understand the documents returned by the retrieval process. This restriction clearly limits the amount and type of information that an individual user really has access to.

Cross Lingual Information Access is concerned with technologies that let users express their query intheir native language, and irrespective of the language in which the information is available, present theinformation in the user-preferred language or set of languages, in a manner that satisfies the user's informationneeds. The additional processing may take the form of machine translation of snippets, summarization and subsequent translation of summaries and/or information extraction.

In recent times, research in Cross Lingual Information Access has been vigorously pursued through several international fora, such as, the Cross-Language Evaluation Forum (CLEF), NTCIR Asian Language Retrieval, Question-answering Workshop and such other fora. A workshop geared towards cross language information retrieval in Indian languages (FIRE) was organized in December 2008. In addition to CLIR, significant results have been obtained in multilingual summarization workshops and cross-language named entity extraction challenges by the ACL (Association for Computational Linguistics) and the Geographic Information retrieval (GeoCLEF) track of CLEF.

The previous two issues of this workshop were held in January 2007, during IJCAI 2007 in Hyderabad, India (http://search.iiit.ac.in/CLIA2007/) and subsequently during IJCNLP 2008 in Hyderabad, India (http://search.iiit.ac.in/CLIA2008/). Both the previous workshops attracted an encouraging number of submissions, and a large number of registered participants.

This third international workshop on Cross Lingual Information Access aims to bring together various trends in multi-source, cross and multilingual information retrievaland access, and provide a venue for researchers and practitioners from academia, government, and industry to interact and share a broad spectrum of ideas, views and applications. The present workshop includes an invited keynote talk, presentations of technical papers selected after peer review followed by a panel discussion.

The workshop starts with an invited keynote talk titled Cross-Language Information Access: Looking Backward, Looking Forward by Douglas W. Oard. The talk starts with a brief recapitulation of two earlier generations of automated support for cross-language information access, the first from roughly 1964 to 1985, and the second from roughly 1989 to the present. With that as background, the talk takes stock of where we are, and where we see unmet needs that call for capabilities beyond what can currently be accomplished. It will be concluded with a few observations about how we might expect the role of the research community to evolve as progressively more capable cross-language information access technologies become commercially viable. In the other paper in the first session, Zhuang et al. report a quasi-language-independent subword recognizer trained on multiple languages, to obtain an abstracted representation of speech data in an unknown language. A retrieval model based on finite

state machines for fuzzy matching of speech sound patterns, and further for speech retrieval has been proposed. A pilot study of speech retrieval in unknown languages is presented using English, Spanish and Russian as training languages, and Croatian as the unknown target language.

In the second session, Raj and Maganti present a transliteration based search engine capable of searching 10 multi-script and multi-encoded Indian languages content on the web. Bouma et al. present a method for cross-lingual alignment of template and infobox attributes in Wikipedia. Elena Filatova presents preliminary results on quantifying Wikipedia multilinguality which show that asymmetries in multilingual Wikipedia do not make it an undesirable corpus for NLP applications training.

In the third session, Zubaryeva and Savoy present a new statistical approach to opinion detection and its evaluation on the English, Chinese and Japanese corpora. Katragadda et al. describe a sentence position based summarizer based on a sentence position policy, created from the evaluation test bed of recent summarization tasks at Document Understanding Conferences (DUC). Sankar and Sobha propose an efficient text summarization technique that involves two basic operations, finding coherent chunks in the document and ranking the text in the individual coherent chunks and picking the sentences that rank above a given threshold.

In the fourth and final session of the workshop, Mukund and Srihari propose a bootstrapped model that involves four levels of text processing for Urdu and show that increasing the training data for POS learning by applying bootstrapping techniques improves NE tagging results. The workshop concludes with a panel discussion.

We thank Douglas W. Oard for the invited keynote talk, all the members of the Program Committee for their excellent and insightful reviews, the authors who submitted contributions for the workshop and the participants for making the workshop a success. We also express our thanks to Asif, Partha and Babji who helped us in organizing and preparing the proceedings as well as maintaining the workshop webpage.

Organizing Committee
The Third International Workshop on Cross Lingual Information Access
NAACL-HLT 2009
June 4, 2009.

Organizers:

Sivaji Bandyopadhyay, Jadavpur University
Pushpak Bhattacharyya, IIT Bombay
Vasudeva Varma, IIIT Hyderabad
Sudeshna Sarkar, IIT Kharagpur
A Kumaran, Microsoft Research India
Raghavendra Udupa, Microsoft Research India

Program Committee:

A Kumaran, Microsoft Research India
Asif Ekbal, Jadavpur University
Carol Peters, ISTI and CLEF Campaign
Gregory Grefenstette, Exalead, France
Mandar Mitra, ISI KOlkata
Paolo Rosso, University Politecnica de Valencia
Patrick Saint Dizier, IRIT, Universite Paul Sabatier
Paul McNamee, John Hopkins University
Pushpak Bhattacharyya, IIT Bombay
Raghavendra Udupa, Microsoft Research India
Ralf Steinberger, European Commission-Joint Research Centre
Sivaji Bandyopadhyay, Jadavpur University
Sobha, L, AU-KBC
Sudeshan Sarkar, IIT Kharagpur
Vasudeva Varma, IIIT Hyderabad

Invited Speaker:

Douglas W. Oard, College of Information Studies and Institute for Advanced Computer Studies, University of Maryland

Table of Contents

Conference Program

Thursday, June 4, 2009

8:30–9:15 Coffee Service

9:15–10:30 Session 1

9:15–9:30 Inauguration

9:30–10:00 *Cross-Language Information Access: Looking Backward, Looking Forward*
Douglas W. Oard

10:00–10:30 *Speech Retrieval in Unknown Languages: a Pilot Study*
Xiaodan Zhuang, Jui Ting Huang and Mark Hasegawa-Johnson

10:30–11:00 Morning Break

11:00–12:30 Session 2

11:00–11:30 *Transliteration based Search Engine for Multilingual Information Access*
Anand Arokia Raj and Harikrishna Maganti

11:30–12:00 *Cross-lingual Alignment and Completion of Wikipedia Templates*
Gosse Bouma, Sergio Duarte and Zahurul Islam

12:00–12:30 *Directions for Exploiting Asymmetries in Multilingual Wikipedia*
Elena Filatova

12:30–14:00 Lunch Break

14:00-15:30 Session 3

14:00–14:30 *Investigation in Statistical Language-Independent Approaches for Opinion Detection in English, Chinese and Japanese*
Olena Zubaryeva and Jacques Savoy

14:30–15:00 *Sentence Position revisited:*
A robust light-weight Update Summarization 'baseline' Algorithm
Rahul Katragadda, Prasad Pingali and Vasudeva Varma

Thursday, June 4, 2009 (continued)

15:00–15:30 *An Approach to Text Summarization.*
Sankar K and Sobha L

15:30–16:00 Afternoon Break

16:00–17:30 Session 4

16:00–16:30 *NE Tagging for Urdu based on Bootstrap POS Learning*
Smruthi Mukund and Rohini K. Srihari

16:30-17:30 Pannel Discussion