IJCNLP 2011

Proceedings of the 5th Workshop on Cross Lingual Information Access

November 13, 2011
Shangri-La Hotel
Chiang Mai, Thailand

AFNLP

IJCNLP 2011

the 5th International Joint Conference on Natural Language Processing

Proceedings of the 5th Workshop on Cross Lingual Information Access

November 13, 2011
Chiang Mai, Thailand

We wish to thank our sponsors

Gold Sponsors

www.google.com

www.baidu.com

The Office of Naval Research (ONR)

The Asian Office of Aerospace Research and Development (AOARD)

Department of Systems Engineering and Engineering Managment, The Chinese University of Hong Kong

Silver Sponsors

Microsoft Corporation

Bronze Sponsors

Chinese and Oriental Languages Information Processing Society (COLIPS)

Supporter

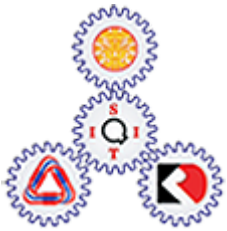
Thailand Convention and Exhibition Bureau (TCEB)

We wish to thank our sponsors

Organizers

Asian Federation of Natural Language Processing (AFNLP)

National Electronics and Computer Technology Center (NECTEC), Thailand



Sirindhorn International Institute of Technology (SIIT), Thailand

Rajamangala University of Technology Lanna (RMUTL), Thailand

Maejo University, Thailand

Chiang Mai University (CMU), Thailand

Introduction

Welcome to the IJCNLP-2011 Workshop on *Cross Lingual Information Access:Computational Linguistics and the Information Need of Multilingual Societies*.

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 (CLIA) is concerned with any technologies and applications that enable people to freely access information that is expressed in any languages. With the rapid development of globalization and digital online information in Internet, huge demand for cross lingual information access has emerged from ordinary netizens (polyglots or monoglots) who are surfing the Internet for special information (e.g. travelling, product description), and communicating in soaring social networks (e.g. Facebook, Youtube, Twitter, Myspace), to global companies which provide multilingual services to their multinational customers, and governments who aim to lower the barriers to international commerce and collaboration, and homeland security. This huge demand has triggered vigorous research and development in CLIA.

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, cross language information retrieval in Indian languages (FIRE) and such other fora. 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.

This workshop is a continuous effort to address the need of cross-lingual information access on top of its previous four issues which were held during IJCAI 2007 in Hyderabad, IJCNLP 2008 in Hyderabad, NAACL 2009 in Colorado, and COLING 2010 in Beijing. It aims to bring together researchers from a variety of fields such as information retrieval, computational linguistics, machine translation, and digital library, and practitioners from government and industry to address the issues of information need of multilingual society.

This fifth international workshop on Cross Lingual Information Access aims to bring together various trends in multi-source, cross and multilingual information retrieval and 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. This workshop also aims to highlight and emphasize the contributions of Natural Language Processing (NLP) and Computational Linguistics to CLIA. The present workshop includes an invited keynote talk followed by presentations of technical papers selected after peer review.

The workshop starts with an invited keynote talk*Web-based Machine Translation* given by Haifeng Wang.

The technical paper presentations will start from the second session of the workshop. The paper by Knoth *et al* addresses the issue of explicit semantic analysis for cross-lingual link discovery. This paper explores how to automatically generate cross-language links between resources in large document

collections.The paper presents new methods that are applicable to any multilingual document collection. They reported a comparative study on the Wikipedia corpus and provide new insights into the evaluation of link discovery systems. In the work of Siva Reddy and Serge Sharoff, they propose cross language PoS taggers for Indian Languages. They show how to build a cross-language PoS tagger for Kannada exploiting the resources of Telugu. In addition they also build large corpora and a morphological analyser for Kannada. They showed that a cross-language taggers are as efficient as mono-lingual taggers. The work by Duo Ding introduces an ongoing work of leveraging a cross-lingual topic model (CLTM) to integrate the multilingual search results. The CLTM detects the underlying topics of different language results and uses the topic distribution of each result to cluster them into topic-based classes. In CLTM, they unify distributions in topic level by direct translation, thus distinguishing from other multi-lingual topic models, which mainly concern the parallelism at document or sentence level. They suggested that CLTM clustering method is effective and outperforms few other existing document clustering techniques. Manaal *et al* propose a soundex-based translation correction in Urdu-English cross-language information retrieval. They discuss the challenges associated with the resource-poor language like Urdu and show the effectiveness of the proposed approach on the benchmark dataset. Li *et al* adopted the contextualized hidden Markov model (CHMM) framework for unsupervised Russian PoS tagging. They propose a backoff smoothing method that incorporates left, right, and unambiguous context into the transition probability estimation during the expectation-maximization process. They show that the resulting model achieves overall and disambiguation accuracies comparable to a CHMM using the classic backoff smoothing method for HMM-based PoS tagging. Johannes Knopp addresses extending a multilingual lexical resource by bootstrapping named entity classification using Wikipedia category system. Their approach is able to classify more than two million named entities and improves the quality of an existing NER resource.

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

We thank Haifeng Wang 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.

Organizing Committee
The 5th International Workshop on Cross Lingual Information Access
IJCNLP 2011
November 13, 2011.

Organizers:

Asif Ekbal, IIT Patna, India (Co-chair)
Deyi Xiong, Institute for InfoComm Research, Singapore (Co-chair)
Prasenjit Majumder, DAIICT, India
Mitesh Khapra, IIT Bombay

Program Committee:

Eneko Agirre, University of the Basque Country
Rafael Banchs, Institute for Infocomm Research
Sivaji Bandyopadhyay, Jadavpur University
Pushpak Bhattacharya, IIT Bombay
Nicola Cancedda, Xerox Research Center
Somnath Chandra, MIT, Govt. of India
Wenliang Chen, Institute for Infocomm Research
Patrick Saint Dizier, IRIT, Universite Paul Sabatier
Xiangyu Duan, Institute for Infocomm Research
Nicola Ferro, University of Padua
Cyril Goutte, National Research Council of Canada
Gareth Jones, Dublin City University
Joemon Jose, University of Glasgow
A Kumaran, Microsoft Research of India
Jun Lang, Institute for Infocomm Research
Swaran Lata, MIT, Govt. of India
Gina-Anne Levow, National Centre for Text Mining (UK)
Qun Liu, Institute of Computing Technology, CAS
Yang Liu, Institute of Computing Technology, CAS
Mandar Mitra, ISI Kolkata
Doug Ouard, University of Maryland, College Park
Carol Peters, Istituto di Scienza e Tecnologie dell'Informazione and CLEF campaign
Paolo Rosso, Technical University of Valencia
Sudeshna Sarkar, IIT Kharagpur
Hendra Setiawan, University of Maryland
L Sobha, AU-KBC, Chennai
Rohini Srihari, University at Buffalo, SUNY
Ralf Steinberger, European Commission - Joint Research Centre, Italy
Le Sun, Institute of Software, CAS
Vasudeva Varma, IIIT Hyderabad
Thuy Vu, Institute for Infocomm Research
Haifeng Wang, Baidu
Yunqing Xia, Tsinghua University, China
Min Zhang, Institute for Infocomm Research
Guodong Zhou, SooChow Universit
Chengqing Zong, Institute of Automation, CAS
Raghavendra Udupa, Microsoft Research

Invited Speaker:

Haifeng Wang, Baidu

Table of Contents

Conference Program

Saturday, November 13, 2011

8:35–8:45 Opening Remarks

8:45–10:00 Keynote Speech

Web-based Machine Translation
Haifeng Wang

10:00–10:30 Break

10:30–11:10 *Using Explicit Semantic Analysis for Cross-Lingual Link Discovery*
Petr Knoth, Lukas Zilka and Zdenek Zdrahal

11:10–11:50 *Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources*
Siva Reddy and Serge Sharoff

11:50–14:00 Lunch

14:00–14:30 *Integrate Multilingual Web Search Results using Cross-Lingual Topic Models*
Duo Ding

14:30–15:00 *Soundex-based Translation Correction in Urdu–English Cross-Language Information Retrieval*
Manaal Faruqui, Prasenjit Majumder and Sebastian Pado

15:00–15:30 *Unsupervised Russian POS Tagging with Appropriate Context*
Li Yang, Erik Peterson, John Chen, Yana Petrova and Rohini Srihari

15:30–16:00 Break

16:00–16:40 *Extending a multilingual Lexical Resource by bootstrapping Named Entity Classification using Wikipedia's Category System*
Johannes Knopp

16:40–17:00 Closing