CLIA 2008

2nd International Workshop on

Cross Lingual Information Access (CLIA)

Addressing the Information Need of Multilingual Societies

The Third International Joint Conference On Natural Language Processing IJCNLP 2008

Proceedings of the Workshop

11 January 2008, Hyderabad, India

Preface

Welcome to the second international workshop on Cross Lingual Information Access (CLIA 2008), with a focus on "Addressing the Information Need of Multilingual Societies".

In this workshop, like in the previous year, our aim was to bring together various trends in cross and multi-lingual information retrieval and access. This year we have accepted eight papers after a careful review process and these accepted papers are included in the proceedings.

The workshop will have four sessions, each focusing on a specific theme: Cross Language Information Retrieval, Translations and Transliterations in CLIR, Information Extraction/Summarization in CLIR contexts, and, finally a session on the overview of the experiences of Indian research groups in the CLEF-2007 competition.

There are three papers in the first session on Cross Language Information Retrieval:
The first paper explores the effects of language relatedness on multilingual Information retrieval. This paper presents a case study with Indo-European and Semitic Languages and addresses some of the challenges posed by Semitic languages IR. The paper on Identifying Similar and Co-referring Documents Across Languages, authors make use of Vector Space Model (VSM) and Named Entities in identifying the co-reference and similarity. In the paper on finding parallel texts on the web using cross-language information retrieval, CLIR techniques are used in combination with structural features to retrieve candidate document pairs from the web. These three papers are part of the session on Cross Language Information Retrieval.

In the second session on Translations and Transliterations in CLIR, we will again have three papers will be presented: The first paper presents results of some experiments in Mining Named Entity Transliteration Pairs from Comparable Corpora, employing English-Tamil named entity parallel comparable corpus texts. The second paper on Domain-Specific Query Translation for Multilingual Information Access using Machine Translation Augmented with Dictionaries Mined from Wikipedia authors demonstrates that effective query translation for CLIA can be achieved in the domain of cultural heritage using a standard MT system, and that domain specific phrase dictionaries that are may be automatically mined from the online Wikipedia. The paper Statistical Transliteration for Cross Language Information Retrieval using HMM alignment model and CRF, presents a technique that combines HMM and CRF for transliteration task in CLIR.

In the third session we have two papers. The first paper is Script Independent Word Spotting in Multilingual Documents, which describes a system that accepts a query in the form of text from the user and returns a ranked list of word images from document image corpus based on similarity with the query word. The second paper is about building a document graph based multi-document summarizer that makes use of a graph model at offline processing time as well as the query time.

Finally, in addition to all the refereed papers, we have six invited presentations by various teams focusing on Indian language CLIR. These presentations are based on the work done by these teams for Ad-hoc task in Cross Language Evaluation Forum (CLEF) in 2007. Teams from IIT Bombay (focusing Marathi, Hindi), IIT Kharagpur (Bengali and Hindi), IIIT Hyderabad (Telugu and Hindi), Microsoft Research India (Tamil, Telugu and Hindi) and Jadhavpur University (Bengali, Telugu and Hindi) will present their work to achieve CLIR for queries in Indian languages and documents in English. In this special session, a team from ISI, Kolkata will make a presentation on FIRE (Forum for Information Retrieval Evaluation), a proposed cross language evaluation forum, specifically for Indian languages. Abstracts of these presentations are also included in these proceedings.

We would like to thank all authors for the hard word that they have put in, in submission, rework and presentation. The workshop would not be possible without them. We would also like to thank the program committee and all the reviewers for their valuable feedback. We hope you would enjoy the workshop.

"We would like to thank Minhaj Babji for all his help in preparing these proceedings as well as supporting the organizing committee during all phases of the workshop."

Vasudeva Varma,
Pushpak Bhattacharya,
Sivaji Bandyopadhyay,
A. Kumaran,
Sudeshna Sarkar.

(Editors CLIA 2008 Workshop)

Committees

Organizing Committee

Vasudeva Varma, IIIT Hyderabad, India
Pushpak Bhattacharya, IIT Bombay, India

Sudeshna Sarkar, IIT Kharagpur, India

A Kumaran Microsoft Research, India

Sivaji Bandyopadhyay, Jadavpur University, Kolkata, India

Program Committee

Asanee Kawtrakul, Kasetsart University, Bangkok, Thailand

Carol Peters, Istituto di Scienza e Tecnologie dell'Informazione and CLEF campaign, Italy

Gilles Serasset, GETALP-LIG, Grenoble, France

Kumaran A, Microsoft Research, Bangalore, India

Lucy Vanderwende, Microsoft Research, USA

Mandar Mitra, ISI Kolkata, India

Paolo Rosso, Universidad Politecnica de Valencia (UPV), Spain

Patrick Saint Dizier, IRIT, Universite Paul Sabatier, Toulouse, France

Paul McNamee, Johns Hopkins University, USA

Petri Myllymaki, University of Helsinki, Finland

Pushpak Bhattacharya, IIT Bombay, India

Ralf Steinberger, European Commission - Joint Research Centre, Italy

Sivaji Bandyopadhyay, Jadavpur University, Kolkata, India

Sobha L, AU-KBC, Chennai, India

Sudeshna Sarkar, IIT Kharagpur, India

Vasudeva Varma, IIIT Hyderabad, India

Workshop Program
11 January 2008, Hyderabad, India

08:45-09:00 Workshop Introduction and Opening Remarks

09:00-10:30 Session-1
Cross Language Information Retrieval

The Effects of Language Relatedness on Multilingual Information Retrieval: A Case Study With Indo-European and Semitic Languages
Peter Chew and Ahmed Abdelali.

Identifying Similar and Co-referring Documents Across Languages
Pattabhi R K Rao T and Sobha L.

Finding parallel texts on the web using cross-language information retrieval
Achim Ruopp and Fei Xia.

10:30 - 11:00 Tea Break

11:00 - 12:30 Session II
Translation and Transliteration in CLIR

Some Experiments in Mining Named Entity Transliteration Pairs from Comparable Corpora
K Saravanan and A Kumaran.

Domain-Specific Query Translation for Multilingual Information Access using Machine Translation Augmented With Dictionaries Mined from Wikipedia
Gareth Jones, Fabio Fantino, Eamonn Newman and Ying Zhang.

Statistical Transliteration for Cross Language Information Retrieval using HMM alignment model and CRF
Prasad Pingali, Suryaganesh, Sreeharsha Yella and Vasudeva Varma.

12:30 - 14:00 Lunch Break

14:00 - 15:00 Session III

Cross Language Information Access and Evaluation

Script Independent Word Spotting in Multilingual Documents
Anurag Bhardwaj, Damien Jose and Venu Govindaraju.

A Document Graph Based Query Focused Multi-Document Summarizer
Sibabrata Paladhi and Sivaji Bandyopadhyay.

15:00 - 15:30 Tea Break

15:30 - 17:30 Session IV

CLIR in Indian Languages - Invited Talks

Hindi and Marathi to English Cross Language Information Retrieval
Manoj Kumar Chinnakotla, Sagar Ranadive, Om P. Damani and Pushpak Bhattacharyya

Bengali and Hindi to English CLIR Evaluation
Debasis Mandal, Sandipan Dandapat, Mayank Gupta, Pratyush Banerjee, Sudeshna Sarkar

Bengali, Hindi and Telugu to English Ad-hoc Bilingual task
Sivaji Bandyopadhyay, Tapabrata Mondal, Sudip Kumar Naskar, Asif Ekbal, Rejwanul Haque, Srinivasa Rao Godavarthy

Cross-Lingual Information Retrieval System for Indian Languages
Jagadeesh Jagarlamudi and A Kumaran

Hindi and Telugu to English CLIR using Query Expansion
Prasad Pingali, Vasudeva Varma

FIRE: Forum for Information Retrieval Evaluation
Mandar Mitra and Prosenjit Majumdar.

17:30 - 17:45 Conclusions and Closing Remarks

Table of Contents