COLING 2012

24th International Conference on Computational Linguistics

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP)

Workshop chairs:
Virach Sornlertlamvanich and Abbas Malik

08 December 2012
Mumbai, India

Diamond sponsors

Tata Consultancy Services
Linguistic Data Consortium for Indian Languages (LDC-IL)

Gold Sponsors

Microsoft Research
Beijing Baidu Netcon Science Technology Co. Ltd.

Silver sponsors

IBM, India Private Limited
Crimson Interactive Pvt. Ltd.
Yahoo
Easy Transcription & Software Pvt. Ltd.

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP)
Virach Sornlertlamvanich and Abbas Malik (eds.)
Revised preprint edition, 2012

Published by The COLING 2012 Organizing Committee
Indian Institute of Technology Bombay,
Powai,
Mumbai-400076
India
Phone: 91-22-25764729
Fax: 91-22-2572 0022
Email: pb@cse.iitb.ac.in

Also available online in the ACL Anthology at http://aclweb.org

Introduction

Welcome to the 3rd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP - 2012), a collocated event at COLING 2012, 8 - 15 December, 2012. South Asia comprises of the countries, Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan and Sri Lanka. Southeast Asia, on the other hand, consists of Brunei, Burma, Cambodia, East Timor, Indonesia, Laos, Malaysia, Philippines, Singapore, Thailand and Vietnam.

This area is the home to thousands of languages that belong to different language families like Indo-Aryan, Indo-Iranian, Dravidian, Sino-Tibetan, Austro-Asiatic, Kradai, Hmong-Mien, etc. In terms of population, South Asian and Southeast Asia represent 35 percent of the total population of the world which means as much as 2.5 billion speakers. Some of the languages of these regions have a large number of native speakers: Hindi (5th largest according to number of its native speakers), Bengali (6th), Punjabi (12th), Tamil(18th), Urdu (20th), etc.

As internet and electronic devices including PCs and hand held devices including mobile phones have spread far and wide in the region, it has become imperative to develop language technology for these languages. It is important for economic development as well as for social and individual progress.

A characteristic of these languages is that they are under-resourced. The words of these languages show rich variations in morphology. Moreover they are often heavily agglutinated and synthetic, making segmentation an important issue. The intellectual motivation for this workshop comes from the need to explore ways of harnessing the morphology of these languages for higher level processing. The task of morphology, however, in South and Southeast Asian Languages is intimately linked with segmentation for these languages.

The goal of WSSANLP is:

- Providing a platform to linguistic and NLP communities for sharing and discussing ideas and work on South and Southeast Asian languages and combining efforts.

- Development of useful and high quality computational resources for under resourced South and Southeast Asian languages.

We are delighted to present to you this volume of proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing. We have received total 39 submission in the categories of long paper, short paper and demonstration. On the basis of our review process, we have competitively selected 9 long papers for oral presentations, 12 short papers for poster presentations and 2 demonstrations.

We look forward to an invigorating workshop.

Virach Sornlertlamvanich (Chair WSSANLP),
National Electronics and Computer Technology Center (NECTEC), Thailand

M.G. Abbas Malik (Chair of Organizing Committee WSSANLP),
Faculty of Computing and Information Technology (North Branch),
King Abdulaziz University, Saudi Arabia

Workshop Chair:

Virach Sornlertlamvanich (National Electronics and Computer Technology Center (NECTEC), Thailand

Workshop Organization Co-chair:

M. G. Abbas Malik (Faculty of Computing and Information Technology (North Branch), King Abdulaziz University, Saudi Arabia

Invited Speaker:

Christian Boitet (GETALP - LIG, University of Grenoble, France)

Organizers:

Aasim Ali (Punjab University College of Information Technology, University of the Punjab, Pakistan)
Amitava Das (Jadavpur Univeristy, India)
Smriti Singh (Indian Institute of Technology Bombay (IITB), India)

Program Committee:

Naveed Afzal (King Abdulaziz University, Saudi Arabia)
M. Waqas Anwar (COMSATS Institute of Information Technology, Pakistan)
Sivaji Bandyopadhyay (Jadavpur University, India)
Vincent Berment (GETALP-LIG / INALCO, France)
Laurent Besacier (GETALP-LIG, Université de Grenoble, France)
Pushpak Bhattacharyya (IIT Bombay, India)
Hervé Blanchon (GETALP-LIG, Université de Grenoble, France)
Christian Boitet (GETALP-LIG, Université de Grenoble, France)
Miriam Butt (University of Konstanz, Germany)
Eric Castelli (International Research Center MICA, Vietnam)
Amitava Das (Norwegian University of Science and Technology, Norway)
Choochart Haruechaiyasak (NECTEC, Thailand)
Sarmad Hussain
(Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology, Pakistan)
Aravind K. Joshi (University of Pennsylvania, USA)
Abid Khan (University of Peshawar, Pakistan)
A. Kumaran (Microsoft Research, India)
Haizhou Liv (Institute for Infocomm Research, Singapore)
M. G. Abbas Malik (King Abdulaziz University - North Jeddah Branch, Saudi Arabia)
Bali Ranaivo-Malançon (Universiti Malaysia Sarawak, Malaysia)
Hammam Riza (Agency for the Assessment and Application of Technology (BPPT), Indonesia)
Rajeev Sangal (IIIT Hyderabad, India)
L. Sobha (AU-KBC Research Centre, Chennai, India)
Virach Sornlertlamvanich (National Electronics and Computer Technology Center (NECTEC), Thailand)
Sriram Venkatapathy (Xerox Research Center Europe, France)

Table of Contents

3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP)

Program

Saturday, December 8, 2012

WSSANLP Session I

9:30–9:50 Opening Remarks

9:50–10:50 Invited Talk by Christian Boitet

10:50–11:15 *Computational evidence that Hindi and Urdu share a grammar but not the lexicon*
K.V.S Prasad and Shafqat Mumtaz Virk

11:30–12:00 **Tea Break**

WSSANLP Session II

12:00–12:25 *Semantic Relation Extraction from a Cultural Database*
Canasai Kruengkrai, Virach Sornlertlamvanich, Watchira Buranasing and Thatsanee Charoenporn

12:25–12:50 *Bengali Question Classification: Towards Developing QA System*
Somnath Banerjee and Sivaji Bandyopadhyay

12:50–13:15 *Morphological Analyzer for Kokborok*
Khumbar Debbarma, Braja Gopal Patra, Dipankar Das and Sivaji Bandyopadhyay

12:15–13:40 *Comparing Different Criteria for Vietnamese Word Segmentation*
Quy T. Nguyen, Ngan L.T. Nguyen and Yusuke Miyao

13:40–14:30 **Lunch Break**

Saturday, December 8, 2012 (continued)

WSSANLP Session III

14:30–16:30 **Posters and Demonstrations**

Using English Acoustic Models for Hindi Automatic Speech Recognition
Anik Dey, Li Ying and Pascale Fung
Tagger Voting for Urdu
Bushra Jawaid and Ondřej Bojar
BIS Annotation Standards With Reference to Konkani Language
Madhavi Sardesai, Jyoti Pawar, Shantaram Walawalikar and Edna Vaz
Automatic Extraction of Compound Verbs from Bangla Corpora
Sibanshu Mukhopadhayay, Tirthankar Dasgupta, Manjira Sinha and Anupam Basu
Influences of particles on Vietnamese tonal Co-articulation
Thi Lan Nguyen and Do Dat Tran
Toward an amazigh language processing
Fatima Zahra Nejme, Siham Boulaknadel and Driss Aboutajdine
Bidirectional Bengali Script and Meetei Mayek Transliteration of Web Based Manipuri News Corpus
Thoudam Doren Singh
Rule-based Machine Translation between Indonesian and Malaysian
Raymond Hendy Susanto, Septina Dian Larasati and Francis M. Tyers
Building Multilingual Search Index using open source framework
Arjun Atreya, Swapnil Chaudhari, Ganesh Ramakrishnan and Pushpak Bhattacharyya
Automatic Searching for English-Vietnamese Documents on the Internet
Quoc Hung Ngo
Error tracking in search engine development
Swapnil Chaudhari, Arjun Atreya, Ganesh Ramakrishnan and Pushpak Bhattacharyya
An Efficient Database Design for IndoWordNet Development Using Hybrid Approach
Venkatesh Prabhu, Shilpa Desai, Hanumant Redkar, Neha Prabhugaonkar, Apurva Nagvenkar and Ramdas Karmali
IndoWordNet Application Programming Interfaces
Neha Prabhugaonkar, Apurva Nagvenkar and Ramdas Karmali

16:30–17:00 **Tea Break**

WSSANLP Session IV

17:00–17:25 *A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language*
Sajjad Ahmad Khan, Waqas Anwar, Usama Ijaz Bajwa and Xuan Wang

17:25–17:50 *Morpheme Segmentation for Kannada Standing on the Shoulder of Giants*
Suma Bhat

17:50–18:15 *Manipuri Morpheme Identification*
Kishorjit Nongmeikapam, Vidya Raj RK, Nirmal Y and Sivaji B

18:15–18:40 *Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach*
Nidhi Krail, Vishal Gupta

18:40–19:00 Closing Remarks