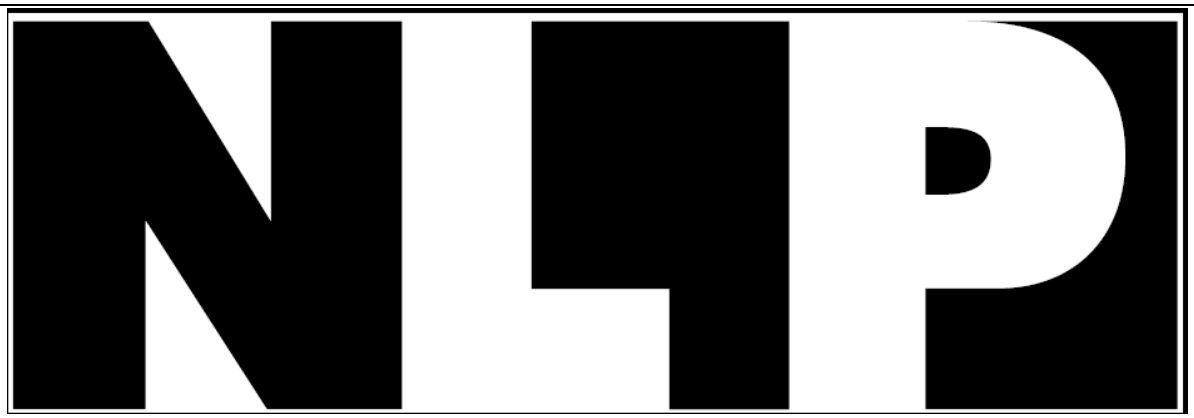
Sixth International Joint Conference on
Natural Language Processing

Proceedings of the Fourth Workshop on South and Southeast Asian Natural Language Processing WSSANLP - 2013

We wish to thank our sponsors and supporters!

Platinum Sponsors



www.anlp.jp

Silver Sponsors

www.google.com

Bronze Sponsors

www.rakuten.com

Supporters

NAGOYA CONVENTION & VISITORS BUREAU

Nagoya Convention & Visitors Bureau

We wish to thank our organizers!

Organizers

Asian Federation of Natural Language Processing (AFNLP)

Toyohashi University of Technology

ISBN 978-4-9907348-8-6

Preface

Welcome to the 4th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP - 2013), a collocated event at the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013) , 14 - 18 October, 2013. 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 4th Workshop on South and Southeast Asian Natural Language Processing. We have received total 15 submission in the categories of long paper and short paper. On the basis of our review process, we have competitively selected 9 long (regular) papers for oral presentations and 3 short papers for poster presentations.

We look forward to an invigorating workshop.

Pushpak Bhattacharyya (Chair WSSANLP-2013),
Indian Institute of Technology Bombay, India

M.G. Abbas Malik (Chair of Organizing Committee WSSANLP-2013),
Faculty of Computing and Information Technology (North Jeddah Branch),
King Abdulaziz University, Saudi Arabia

The Fourth Workshop on South and Southeast Asian Natural Language processing WSSANLP-2013

Workshop Chair:

Pushpak Bhattacharyya, Indian Institute of Technoloyg Bombay, India

Workshop Organization Co-chair:

M. G. Abbas Malik, King Abdulaziz University, Saudi Arabia

Invited Speaker:

Dekai Wu, HKUST Human Language Technology Center

Organizers:

Aasim Ali, Punjab University College of Information Technology, University of the Punjab, Pakistan
Amitava Das, Jadavpur Univeristy, India

Program Committee:

Sadaf Abdul Rauf, Fatima Jinnah Women University, Pakistan
Naveed Afzal, King Abdulaziz University, Saudi Arabia
Aasim Ali, University of the Punjab, Pakistan
M. Waqas Anwar, COMSATS Institute of Information Technology, Pakistan
Bal Krishna Bal, Kathmandu University, Nepal
Sivaji Bandyopadhyay, Jadavpur University, India
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
Amitava Das, Norwegian University of Science and Technology, Norway
Alain Desoulieres, INALCO Paris, France
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
Imtiaz Hussain Khan, King Abdulaziz University, Saudi Arabia
A. Kumaran, Microsoft Research, India
Haizhou Li, Institute for Infocomm Research, Singapore
M. G. Abbas Malik, King Abdulaziz University - North Jeddah Branch, Saudi Arabia

Violaine Prince, University of Montpellier 2, France
Bali Ranaivo-Malançon, Universiti Malaysia Sarawak, Malaysia
Hammam Riza, Agency for the Assessment and Application of Technology (BPPT), Indonesia
L. Sobha, AU-KBC Research Centre, Chennai, India
Virach Sornlertlamvanich, TCL, National Institute of Information and Communication Technology, Thailand
Sriram Venkatapathy, Xerox Research Center Europe, France
Eric Wehrli, University of Geneva, Switzerland

Table of Contents

Workshop Program

Friday October 18, 2013

(9:30 - 9:40) Openning Session

(9:40 - 10:35) Invited Talk

\+ by Prof. Dekai Wu, HKUST Human Language Technology Center

(10:35 - 10:50) Coffee Break

Session Regular Papers 1: (10:50 - 12:30) WSSANLP Session 1

10:50 *Fast Bootstrapping of Grapheme to Phoneme System for Under-resourced Languages - Application to the Iban Language*
Sarah Samson Juan and Laurent Besacier

11:15 *LexToPlus: A Thai Lexeme Tokenization and Normalization Tool*
Choochart Haruechaiyasak and Alisa Kongthon

11:40 *A Three-Layer Architecture for Automatic Post Editing System Using Rule-Based Paradigm*
Mahsa Mohaghegh, Abdolhossein Sarrafzadeh and Mehdi Mohammadi

12:05 *Statistical Stemming for Kannada*
Suma Bhat

(12:30 - 13:30) Lunch break

Friday October 18, 2013 (continued)

Session Regular Papers 2: (13:30 - 14:45) WSSANLP Session 2

13:30 *On Application of Conditional Random Field in Stemming of Bengali Natural Language Text*
Sandipan Sarkar and Sivaji Bandyopadhyay

13:55 *Urdu Hindi Machine Transliteration using SMT*
M. G. Abbas Malik, Christian Boitet, Laurent Besacier and Pushpak Bhattcharyya

14:20 *Urdu Spell Checking: Reverse Edit Distance Approach*
Saadat Iqbal, Muhammad Waqas Anwar, Usama Ijaz Bajwa and Zobia Rehman

(14:45 - 15:00) Coffee Break

Session Poster Papers: (15:00 - 15:50) WSSANLP Session 3

15:00 *Information Mining from Islamic Scriptures*
Abdul Rauf Saeed and Syed Waqar Jaffry

15:25 *English to Urdu Hierarchical Phrase-based Statistical Machine Translation*
Nadeem Khan, Muhammad Waqas Anwar, Usama Ijaz Bajwa and Nadir Durrani

Session Regular Papers 3: (15:50 - 16:40) WSSANLP Session 4

15:50 *Cliticization and Endoclitics Generation of Pashto Language*
Azizud Din

16:15 *Malayalam Clause Boundary Identifier: Annotation and Evaluation*
Sobha Lalitha Devi and Lakshmi S

Friday October 18, 2013 (continued)

(16:40 - 17:00) Closing Remarks