COLING 2014

The 25th International Conference on Computational Linguistics

Proceedings of the Conference the 5th Workshop on South and Southeast Asian NLP WSSANLP - 2014

August 23, 2014
Dublin, Ireland

ISBN 978-1-873769-41-6
Proceedings of the Fifth Workshop on South and Southeast Asian Natural Language Processing
Christian Boitet and M.G. Abbas Malik (eds.)

Preface

Welcome to the 5th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP - 2014), a collocated event at the 25th International Conference on Computational Linguistics (COLING 2014) , 23 - 29 August, 2014.

South and Southeast Asia comprise 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 5th Workshop on South and Southeast Asian Natural Language Processing. We have received total 18 submission in the categories of long paper and short paper. On the basis of our review process, we have competitively selected 7 long (regular) papers for oral presentations and 7 short papers for poster presentations.

We look forward to an invigorating workshop.

Christian Boitet (Chair WSSANLP-2014),
University of Grenoble I, France

M.G. Abbas Malik (Co-Chair WSSANLP-2014),
Faculty of Computing and Information Technology (North Jeddah Branch),
King Abdulaziz University, Saudi Arabia

The Fifth Workshop on South and Southeast Asian Natural Language processing WSSANLP-2014

WSSANLP Organizers

Workshop Chair

Christian Boitet, University of Grenoble I, France

Workshop Co-Chairs

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

Organizing Committee

Amitava Das, University of North Texas, USA

Sadaf Abdul Rauf, Fatima Jinnah Women University, Pakistan

WSSANLP Invited Speaker

Vincent Berment, INaLCO, Paris, France (Lecturer)
LIG/GÉTALP, Grenoble, France (Associated Researcher)
Taranis Software, Paris, France (Director)

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
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
Erik Cambria, National University of Singapore, Singapore
Eric Castelli, International Research Center MICA, Vietnam
Sandipan Dandapat, IIT Guwahati, India
Amitava Das, University of North Texas, USA
Alexander Gelbukh, Center for Computing Research, CIC, Mexico
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
Malhar Kulkarni, Indian Institute of Technology Bombay, India
Amba Kulkarni, Department of Sanskrit Studies, University of Hyderabad, India
A. Kumaran, Microsoft Research, India
Gurpreet Singh Lehal, Punjabi University, Patiala, India
M. G. Abbas Malik, King Abdulaziz University, Saudi Arabia
Bali Ranaivo-Malançon, University Malaysia Sarawak, Malaysia
Fuji Ren, University of Tokushima, Japan
Hammam Riza, Agency for the Assessment and Application of Technology (BPPT), Indonesia
Paolo Rosso, Universitat Politècnica de València, Spain
Huda Sarfraz, Beacon house National University, Pakistan
Dipti Mishra Sharma, IIIT Hyderabad, India
Sunil Sivadas, Institute for Infocomm Research, Singapore
L. Sobha, AU-KBC Research Centre, 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

WSSANLP 2014 Program

Saturday August 23, 2014

(8:45 - 9:00) Openning Session

(9:00 - 10:15) Invited Talk

\+ by Vincent Berment

(10:15 - 10:45) Coffee Break

Session Regular Papers 1: (10:45 - 12:30) WSSANLP Session 1

10:45 *Towards Identifying Hindi/Urdu Noun Templates in Support of a Large-Scale LFG Grammar*
Sebastian Sulger and Ashwini Vaidya

11:10 *Konkanverter - A Finite State Transducer based Statistical Machine Transliteration Engine for Konkani Language*
Vinodh Rajan

11:35 *Integrating Dictionaries into an Unsupervised Model for Myanmar Word Segmentation*
Ye Kyaw Thu, Andrew Finch, Eichiro SUMITA and Yoshinori Sagisaka

12:00 *A Framework for Learning Morphology using Suffix Association Matrix*
Shilpa Desai, Jyoti Pawar and Pushpak Bhattacharyya

(12:30 - 14:00) Lunch break

Saturday August 23, 2014 (continued)

Session Short Papers: (14:00 - 15:15) WSSANLP Session 2

English to Urdu Statistical Machine Translation: Establishing a Baseline
Bushra Jawaid, Amir Kamran and Ondrej Bojar

A hybrid approach for automatic clause boundary identification in Hindi
Rahul Sharma and Soma Paul

RBMT as an alternative to SMT for under-resourced languages
Guillaume de Malézieux, Amélie Bosc and Vincent Berment

Developing an interlingual translation lexicon using WordNets and Grammatical Framework
Shafqat Mumtaz Virk, K.V.S Prasad, Aarne Ranta and Krasimir Angelov

A Dictionary Data Processing Environment and Its Application in Algorithmic Processing of Pali Dictionary Data for Future NLP Tasks
Jürgen Knauth and David Alfter

Constituent structure representation of Pashto Endoclitics
Azizud Din, Bali Ranaivo-Malançon and M. G. Abbas Malik

Real Time Early-stage Influenza Detection with Emotion Factors from Sina Microblog
Xiao Sun, Jiaqi Ye and Fuji Ren

(15:15 - 15:45) Coffee Break

Session Regular Papers 2: (15:45 - 17:00) WSSANLP Session 3

15:45 *Building English-Vietnamese Named Entity Corpus with Aligned Bilingual News Articles*
Quoc Hung Ngo, Dinh Dien and Werner Winiwarter

16:10 *Character-Cluster-Based Segmentation using Monolingual and Bilingual Information for Statistical Machine Translation*
Vipas Sutantayawalee, Peerachet Porkeaw, Thepchai Supnithi, Prachya Boonkwan and Sitthaa Phaholphinyo

16:35 *A rule based approach for automatic clause boundary detection and classification in Hindi*
Rahul Sharma

Saturday August 23, 2014 (continued)

(17:00 - 17:15) Closing Remarks