IJCNLP 2011

Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2011)

November 8, 2011
Shangri-La Hotel
Chiang Mai, Thailand

AFNLP

IJCNLP 2011

Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2011)

Collocated event at the 5th International Joint Conference on Natural Language Processing

November 8, 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

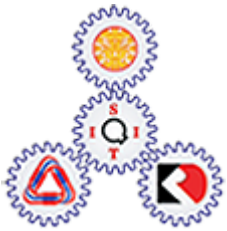
Thailand Convention and Exhibition Bureau (TCEB)

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

Preface

Welcome to the IJCNLP Workshop on South and Southeast Asian Natural Language Processing (WSSANLP). 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 2nd Workshop on South and Southeast Asian NLP. We have received 15 long and short submissions. On the basis of our review process, we have competitively selected 9 papers.

We look forward to an invigorating workshop.

Rajeev Sangal (Chair WSSANLP),
IIIT Hyderabad, India

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

2nd Workshop on South and Southeast Asian Natural Language processing

Workshop Chair:

Rajeev Sangal, IIIT Hyderabad, India

Workshop Organization Co-chair:

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

Invited Speaker:

Pushpak Bhattacharyya, IIT Bombay, India

Organizers:

Aasim Ali, Punjab University College of Information Technology, University of the Punjab, Pakistan
Amitava Das, Jadavpur Univeristy, India
Fahad Iqbal Khan, COMSATS IIT Lahore, Pakistan
M. G. Abbas Malik, King Abdulaziz University, Saudi Arabia
Smriti Singh, Indian Institute of Technology Bombay (IITB), India

Program Committee:

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
Nicola Cancedda, Xerox Research Center Europe (XRCE), France
Eric Castelli, International Research Center MICA, Vietnam
Laurence Danlos, University of Paris 7, France
Georges Fafiotte, GETALP-LIG, Université de Grenoble, France
Zulfiqar HabibCOMSATS Institute of Information Technology, Pakistan
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
Krit KOSAWAT, Human Language Technology Laboratory (HLT) National Electronics and Computer Technology Center (NECTEC), Thailand
Bal Krishna Bal, University of Kathmandu, Nepal
A. Kumaran, Microsoft Research, India
Gurpreet Singh Lehal, Punjabi University Patiala, India
Haizhou Li, Institute for Infocomm Research, Singapore
M. G. Abbas Malik, King Abdulaziz University, Saudi Arabia
Bali Ranaivo-Malançon, Multimedia University, 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
Ruvan Weerasinghe, University of Colombo School of Computing, Sri Lanka

Table of Contents

Program the 2nd Workshop on South and Southeast Asian Natural Language Processing

Tuesday, November 8, 2011

8:30–8:45 Opening Remarks

8:45–10:00 Invited Talk by Pushpak Bhattacharyya

10:00–10:30 Break

WSSANLP Session I

10:30–11:00 *Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati*
Kartik Suba, Dipti Jiandani and Pushpak Bhattacharyya

11:00–11:30 *Improving Persian-English Statistical Machine Translation:Experiments in Domain Adaptation*
Mahsa Mohaghegh, Abdolhossein Sarrafzadeh and Tom Moir

11:30–12:00 *Thai Word Segmentation Verification Tool*
Supon Klaithin, Kanyanut Kriengket, Sitthaa Phaholphinyo and Krit Kosawat

12:00–12:30 *The Semi-Automatic Construction of Part-Of-Speech Taggers for Specific Languages by Statistical Methods*
Tomohiro YAMASAKI, Hiromi WAKAKI and Masaru SUZUKI

12:30–14:00 Lunch Break

WSSANLP Session II

14:00–14:30 *Towards a Malay Derivational Lexicon: Learning Affixes Using Expectation Maximization*
Suriani Sulaiman, Michael Gasser and Sandra Kuebler

14:30–15:00 *Punjabi Language Stemmer for nouns and proper names*
Vishal Gupta and Gurpreet Singh Lehal

15:00–15:30 *Challenges in Urdu Text Tokenization and Sentence Boundary Disambiguation*
Zobia Rehman, Waqas Anwar and Usama Ijaz Bajwa

15:30–16:00 Break

Tuesday, November 8, 2011 (continued)

WSSANLP Session III

16:00–16:30 *Challenges in Developing a Rule based Urdu Stemmer*
Sajjad Ahmad Khan, Waqas Anwar and Usama Ijaz Bajwa

16:30–17:00 *Developing a New System for Arabic Morphological Analysis and Generation*
Mourad Gridach and Noureddine Chenfour

17:00–17:15 Closing Remarks