WSSANLP 2016

6th Workshop on South and Southeast Asian Natural Language Processing

Proceedings of the Conference

December 11-16, 2016
Osaka, Japan

ISBN978-4-87974-705-1

Preface

Welcome to the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP - 2016), a collocated event at the 26th International Conference on Computational Linguistics (COLING 2016) , December 11 - 16, 2016 at Osaka International Convention Center, Osaka, Japan.

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), and Urdu (20th).

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 6th Workshop on South and Southeast Asian Natural Language Processing. We have received total 37 submissions in the categories of long paper and short paper. On the basis of our review process, we have competitively selected 19 full papers and 3 short papers.

We look forward to an invigorating workshop.

Dekai Wu (Chair WSSANLP-2016),
Hong Kong University of Science and Technology, Hong Kong

Pushpak Bhattacharyya (Co-Chair WSSANLP-2016),
Indian Institute of Technology Patna, India

Workshop Chair

Dekai Wu, Hong Kong University of Science and Technology, Hong Kong

Workshop Co-Chair

Pushpak Bhattacharyya, Indian Institute of Technology Patna, India

Key Note Speaker

Alain Désoulières, INALCO - CERLOM, France

Organisers

M. G. Abbas Malik, Auckland University of Technology, Auckland, New Zealand (chair)
Sadaf Abdul Rauf, Fatima Jinnah Women University, Islamabad, Pakistan
Mahsa Mohaghegh, Unitec Institute of Technology, Auckland, New Zealand

Programme Committee

Sadaf Abdul Rauf, Fatima Jinnah Women University, Pakistan
Naveed Afzal, Cardiovascular Biomarkers Laboratory, Mayo Clinic, USA
Tafseer Ahmed, DHA Suffa University, Pakistan
Aasim Ali, University of the Punjab, Pakistan
Jalal S. Alowibdi, University of Jeddah, Saudi Arabia
Saleh Alshomrani, University of Jeddah, Saudi Arabia
Amer Alzaidi, University of Jeddah, Saudi Arabia
M. Waqas Anwar, COMSATS Institute of Technology Abbottabad, Pakistan
Bal Krishna Bal, Kathmandu University, Nepal
Sivaji Bandyopadhyay, Jadavpur University, India
Vincent Berment, GETALP-LIG and INALCO, France
Laurent Besacier, University of Grenoble, France
Pushpak Bhattacharyya, Indian Institute of Technology Patna, India
Hervé Blanchon, University of Grenoble, France
Christian Boitet, University of Grenoble, France
Miriam Butt, University of Konstanz, Germany
Eric Castelli, International Research Center MICA, Vietnam
Amitava Das, Indian Institute of Information Technology, Sri City, India
Alain Desoulieres, INALCO - CERLOM, France
Alexander Gelbukh, Center for Computing Research, CIC, Mexico
Choochart Haruechaiyasak, National Electronics and Computer Technology Center (NECTEC), Thailand

Sarmad Hussain, University of Engineering and Technology Lahore, Pakistan

Aravind K. Joshi, University of Pennsylvania, USA

Amba Kulkarni, University of Hyderabad, India

Gurpreet Singh Lehal, Punjabi University, Patiala, India

Haizhou Li, Institute for Infocomm Research, Singapore

M. G. Abbas Malik, Auckland University of Technology, New Zealand

Mahsa Mohaghegh, Unitec Institute of Technology, New Zealand

Ajit Narayanan, Auckland University of Technology, New Zealand

K. V. S. Prasad, Chalmers University of Technology, Sweden

Bali Ranaivo-Malançon, University of Malaysia Sarawak, Malaysia

Paolo Rosso, Universitat Politècnica de València, Spain

Huda Sarfraz, Beacon house National University, Pakistan

Hossein Sarrafzadeh, High Technology Transdisciplinary Research Network, Unitec Auckland, New Zealand

L. Sobha, AU-KBC Research Centre, India

Virach Sornlertlamvanich, TCL, National Institute of Information and Communication Technology, Thailand

Ruvan Weerasinghe, University of Colombo School of Computing, Sri Lanka

Table of Contents

Full Papers

Short Papers

Conference Program

Sunday, December 11, 2016

WSSANLP 2016 Openning

9:00–9:10 *Openning Remarks*

9:10–10:00 *Key Note by Alain Désoulières, INALCO, CERLOM, France*

10:00–10:20 Coffee and Tea Break

10:20–12:00 WSSANLP Session 1: Oral Presentations

Session Chair: Dekai Wu

10:20–10:40 *Compound Type Identification in Sanskrit: What Roles do the Corpus and Grammar Play?*
Amrith Krishna, Pavankumar Satuluri, Shubham Sharma, Apurv Kumar and Pawan Goyal

10:40–11:00 *Comparison of Grapheme-to-Phoneme Conversion Methods on a Myanmar Pronunciation Dictionary*
Ye Kyaw Thu, Win Pa Pa, Yoshinori Sagisaka and Naoto Iwahashi

11:00–11:20 *Character-Aware Neural Networks for Arabic Named Entity Recognition for Social Media*
Mourad Gridach

11:20–11:40 *Development of a Bengali parser by cross-lingual transfer from Hindi*
Ayan Das, Agnivo Saha and Sudeshna Sarkar

11:40–12:00 *Sinhala Short Sentence Similarity Calculation using Corpus-Based and Knowledge-Based Similarity Measures*
Jcs Kadupitiya, Surangika Ranathunga and Gihan Dias

12:00–13:30 Lunch Break

Sunday, December 11, 2016 (continued)

13:30–14:55 WSSANLP Session 2: Poster Presentations

Session Chair: M G Abbas Malik

Full Papers

Enriching Source for English-to-Urdu Machine Translation
Bushra Jawaid, Amir Kamran and Ondřej Bojar

The IMAGACT4ALL Ontology of Animated Images: Implications for Theoretical and Machine Translation of Action Verbs from English-Indian Languages
Pitambar Behera, Sharmin Muzaffar, Atul kr. Ojha and Girish Jha

Crowdsourcing-based Annotation of Emotions in Filipino and English Tweets
Fermin Roberto Lapitan, Riza Theresa Batista-Navarro and Eliezer Albacea

Temporal Information Extraction in Clinical Domain (TIECA)
Shujeevan Kanapathipillai, Viraj Welgama and Ruwan Weerasinghe

Sentiment Analysis of Tweets in Three Indian Languages
Shanta Phani, Shibamouli Lahiri and Arindam Biswas

Dealing with Linguistic Divergences in English-Bhojpuri Machine Translation
Pitambar Behera, Neha Mourya and Vandana Pandey

The development of a web corpus of Hindi language and corpus-based comparative studies to Japanese
Miki Nishioka and Shiro Akasegawa

Automatic Creation of a Sentence Aligned Sinhala-Tamil Parallel Corpus
Riyafa Abdul Hameed, Nadeeshani Pathirennehelage, Anusha Ihalapathirana, Maryam Ziyad Mohamed, Surangika Ranathunga, Sanath Jayasena, Gihan Dias and Sandareka Fernando

Short Papers

Align Me: A framework to generate Parallel Corpus Using OCRs and Bilingual Dictionaries
Priyam Bakliwal, Devadath V V and C V Jawahar

Learning Indonesian-Chinese Lexicon with Bilingual Word Embedding Models and Monolingual Signals
Xinying Qiu and Gangqin Zhu

Creating rich online dictionaries for the Lao–French language pair, reusable for Machine Translation
Vincent Berment

Sunday, December 11, 2016 (continued)

15:00–16:50 WSSANLP Session 3: Oral Presentations

Session Chair: Christian Boitet

15:00–15:10 *Introduction of Language Resources and Evaluation (LRE) Map by Laurent Besacier, member European Language Resources Association (ELRA)*

15:10–15:30 *Clustering-based Phonetic Projection in Mismatched Crowdsourcing Channels for Low-resourced ASR*
Wenda Chen, Mark Hasegawa-Johnson, Nancy Chen, Preethi Jyothi and Lav Varshney

15:30–15:50 *Improving the Morphological Analysis of Classical Sanskrit*
Oliver Hellwig

15:50–16:10 *Query Translation for Cross-Language Information Retrieval using Multilingual Word Clusters*
Paheli Bhattacharya, Pawan Goyal and Sudeshna Sarkar

16:10–16:30 *A study of attention-based neural machine translation model on Indian languages*
Ayan Das, Pranay Yerra, Ken Kumar and Sudeshna Sarkar

16:30–16:50 *Comprehensive Part-Of-Speech Tag Set and SVM based POS Tagger for Sinhala*
Sandareka Fernando, Surangika Ranathunga, Sanath Jayasena and Gihan Dias

WSSANLP 2016 Closing

16:50–17:00 *Closing Remarks*