Coling 2010

23rd International Conference on Computational Linguistics

Proceedings of the

1st Workshop on South and southeast Asian Natural Language Processing

24 August 2010
Beijing International Convention Center

Produced by
Chinese Information Processing Society of China
All rights reserved for Coling 2010 CD production.

To order the CD of Coling 2010 and its Workshop Proceedings, please contact:

Chinese Information Processing Society of China
No.4, Southern Fourth Street
Haidian District, Beijing, 100190
China
Tel: +86-010-62562916
Fax: +86-010-62562916
`cips@iscas.ac.cn`

Preface

Welcome to the Coling 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.

There 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 Asia and Southeast Asia represent 34.94 percent of the total population of the world. 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.

A characteristic of these languages is that they are under-resourced. But 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 Source (Lewis, 2009) languages for higher level processing. Table 1: Population and Number of Living Languages of The task of morphology, however, is South and Southeast Asia 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 you this volume of proceedings of 1st Workshop on South and Southeast Asian NLP. We have received total 18 long and short paper submissions. On the basis of our review process, we have competitively selected 13 papers, but unfortunately 6 papers were withdrawn from the workshop due to double submission and authors chose to present their paper in other events. We hope that we will be able to make this workshop so successful that people would like to present their papers in this workshop in future.

M. G. Abbas Malik, *Chair of Organizing Committee WSSANLP*

Workshop Chair:

Aravin K. Joshi, *Chair of the Worskhop*, University of Pennsylvania, (USA)

Organizers:

M. G. Abbas Malik, (*Chair of Organizing Committee*), GETALP-LIG, University of Grenoble (France)
Aasim Ali, CRULP, National University of Computer and Emerging Sciences (Pakistan)
Asif Ekbal, Dept. of Computational Linguistics, University of Heidelberg (Germany)
Dulip Herath, University of Colombo School of Computing (Sri Lanka)
Hong-Thi Nguyen, GETALP-LIG, University of Grenoble (France)
Muhammad Humayoun, LAMA, Universit de Savoie (France)
Menaka Sankaralingam AUKBC Research Centre, Chennai (India)
Monojit Choudhury, Researcher, Microsoft Research (India)
Sadaf Abdul Rauf, Universit du Maine (France)
Smriti Singh, Indian Institute of Technology Bombay (India)

Program Committee:

Ranaivo-Malanon Bali, Multimedia University (Malaysia)
Sivaji Bandyopadhyay, Jadavpur University (India)
Vincent Berment, GETALP-LIG / INALCO (France)
Laurent Besacier, GETALP-LIG, Universit de Grenoble (France)
Pushpak Bhattacharyya, IIT Bombay (India)
Christian Boitet, GETALP-LIG, Universit de Grenoble (France)
Nicola Cancedda, Xerox Research Center Europe (France)
Eric Castelli, International Research Center MICA (Vietnam)
Luong Chi Mai, Institute of IT, Vietnamese Academy of Science and Technology (Vietnam)
Laurence Danlos, University Paris 7 (France)
Georges Fafiotte, GETALP-LIG, Universit de Grenoble (France)
John A. Goldsmith, University of Chicago (USA)
Grard Huet, INRIA (France)
San San Hnin Tun, Cornell University (USA)
Sarmad Hussain, National University (Pakistan)
Abid Khan, University of Peshawar (Pakistan)
Wunna Ko Ko, Northern Illinois University (USA)
Bal Krishna Bal, University of Kathmandu (Nepal)
A. Kumaran, Microsoft Research (India)
Gurpreet Singh Lehel, Punjabi University Patiala (India)
Haizhou Li, Institute for Infocomm Research (Singapore)
Alec Marantz, New York University (USA)

Christian Monson, OHSU (USA)
Annie Montaut, INALCO Paris (France)
Hammam Riza, Agency for the Assessment and Application of Technology (Indonesia)
Rajeev Sangal, IIIT Hyderabad (India)
Anne Schiller, Xerox Research Center Europe (France)
L. Sobha, AU-KBC Research Centre (India)
Chan Somnoble, Royal University of Phnom Penh (Cambodia)
Virach Sornlertlamvanich, TCL, National Institute of Information and Communication Technology (Thailand)
Ruvan Weerasinghe, University of Colombo School of Computing (Sri Lanka)
Khaver Zia, Beacon House National University (Pakistan)

Invited Speaker:

Rajeev Sangal, IIIT Hyderabad (India)

Table of Contents

Boosting N-gram Coverage for Unsegmented Languages Using Multiple Text Segmentation Approach
Solomon Teferra Abate, Laurent Besacier and Sopheap Seng 1

Thai Sentence-Breaking for Large-Scale SMT
Glenn Slayden, Mei-Yuh Hwang and Lee Schwartz 8

Clause Identification and Classification in Bengali
Aniruddha Ghosh, Amitava Das and Sivaji Bandyopadhyay 17

A Paradigm-Based Finite State Morphological Analyzer for Marathi
Mugdha Bapat, Harshada Gune and Pushpak Bhattacharyya 26

Web Based Manipuri Corpus for Multiword NER and Reduplicated MWEs Identification using SVM
Thoudam Doren Singh and Sivaji Bandyopadhyay 35

A Word Segmentation System for Handling Space Omission Problem in Urdu Script
Gurpreet Lehal 43

Hybrid Stemmer for Gujarati
Pratikkumar Patel, Kashyap Popat and Pushpak Bhattacharyya 51

Conference Program

Tuesday, August 24, 2010

16:00–16:10	Opening Remarks
16:10–16:40	Invited Talk by Dr. Rajeev Sangal
16:40–17:00	*Boosting N-gram Coverage for Unsegmented Languages Using Multiple Text Segmentation Approach* Solomon Teferra Abate, Laurent Besacier and Sopheap Seng
17:00–17:20	*Thai Sentence-Breaking for Large-Scale SMT* Glenn Slayden, Mei-Yuh Hwang and Lee Schwartz
17:20–17:40	*Clause Identification and Classification in Bengali* Aniruddha Ghosh, Amitava Das and Sivaji Bandyopadhyay
17:40–17:50	break
17:50–18:10	*A Paradigm-Based Finite State Morphological Analyzer for Marathi* Mugdha Bapat, Harshada Gune and Pushpak Bhattacharyya
18:10–18:30	*Web Based Manipuri Corpus for Multiword NER and Reduplicated MWEs Identification using SVM* Thoudam Doren Singh and Sivaji Bandyopadhyay
18:30–18:40	*A Word Segmentation System for Handling Space Omission Problem in Urdu Script* Gurpreet Lehal
18:40–18:50	*Hybrid Stemmer for Gujarati* Pratikkumar Patel, Kashyap Popat and Pushpak Bhattacharyya
18:50–19:00	Closing Remarks