COLING 2012

24th International Conference on Computational Linguistics

Proceedings of the 10th Workshop on Asian Language Resources

Workshop chairs:
Ruvan Weerasinghe, Sarmad Hussain,
Virach Sornlertlamvanich and Rachel Edita O. Roxas

09 December 2012
Mumbai, India

Diamond sponsors

Tata Consultancy Services
Linguistic Data Consortium for Indian Languages (LDC-IL)

Gold Sponsors

Microsoft Research
Beijing Baidu Netcon Science Technology Co. Ltd.

Silver sponsors

IBM, India Private Limited
Crimson Interactive Pvt. Ltd.
Yahoo
Easy Transcription & Software Pvt. Ltd.

Proceedings of the 10th Workshop on Asian Language Resources
Ruvan Weerasinghe, Sarmad Hussain, Virach Sornlertlamvanich and Rachel Edita O. Roxas (eds.)
Revised preprint edition, 2012

Published by The COLING 2012 Organizing Committee
Indian Institute of Technology Bombay,
Powai,
Mumbai-400076
India
Phone: 91-22-25764729
Fax: 91-22-2572 0022
Email: pb@cse.iitb.ac.in

Also available online in the ACL Anthology at http://aclweb.org

Preface

As research in the field of Natural Language Processing matures across Asia, there is a growing need for developing language resources. However the region is not only short in linguistic resources for the more than 2200 language spoken in the region, there is also lack of experience of the researchers to develop these resources. As the efforts to develop the linguistic resources increases, there is also need to coordinate the efforts to develop common frameworks and processes so that these resources can be used by those dealing with new and under-resourced languages.

The Asian Language Resources Workshop (ALR) is organised under the Asian Federation of Natural Language Processing (AFNLP) to chart and catalogue the status of Asian Language Resources, to investigate and discuss the problems related to the standards and specification of creating and sharing various levels of language resources, to promote a dialogue between developers and users of various language resources in order to address any gaps in language resources and practical applications, and to nurture collaboration in their development, and to provide the opportunity for researchers from Asia to collaborate with researchers in other regions.

We are very pleased to publish this volume that contains the papers presented at the Tenth Workshop on Asian Language Resources (ALR-10) held in conjunction with the 24th International Conference on Computational Linguistics (COLING 2012) from 8th to 15th December 2012 in Mumbai, India. We received a total of 25 submissions for resources and tools for languages in the region such as Persian, Tibetan, Urdu, Mongolian, Assamese, Bodo, Magahi, Korean, Hindi and Bangla, of which 14 (56%) have been accepted for oral presentation through a double-blind refereeing process. We would like to thank the authors for their submissions and the Program Committee for their timely reviews. We hope that ALR workshops will continue to encourage researchers to focus on developing and sharing resources for Asian languages, an essential requirement for research in NLP in the region.

Ruvan Weerasinghe (Chair)
Sarmad Hussain (Co-Chair)
Virach Sornlertlamvanich (Co-Chair)
Rachel Edita O. Roxas (Co-Chair)

Organizing Committee

Organizers:

Ruvan Weerasinghe, University of Colombo School of Computing, Sri Lanka
Sarmad Hussain, University of Engineering and Technology, Pakistan
Virach Sornlertlamvanich, NECTEC, Thailand
Rachel Roxas, De La Salle University, Philippines

Program Committee:

Abid Khan, Univ. of Peshawar, Pakistan
Chai Wutiwiwatchai - NECTEC, Thailand
Dipti Misra Sharma, IIIT, Hyderabad, India
Francis Bond, Nanyang Technological University, Singapore
Haizhou Li - I2R, Singapore
Key-Sun Choi - KAIST, Korea
Kiyoaki Shirai - JAIST, Japan
Miriam Butt – Univ. of Konstanz, Germany
Mirna Adriani – Univ. of Indonesia, Indonesia
Rachel Edita O. Roxas – De La Salle University, Philippines
Rajeev Sangal, IIIT Hyderabad, India
Reinhard Schaler – Localization Research Centre, University of Limerick, Ireland
Ruli Marunung – Univ. of Indonesia, Indonesia
Ruvan Weerasinghe - LTRL, University of Colombo, School of Computing, Sri Lanka
Sarmad Hussain – CLE-KICS, UET Lahore, Pakistan
Steven Bird – University of Melbourne, Australia
Takenobu Tokunaga - Tokyo Institute of Technology, Japan
Virach Sornlertlamvanich - NECTEC, Thailand

Table of Contents

10th Workshop on Asian Language Resources

Program

Sunday, 9 December 2012

Session 1 – Linguistic Resources

09:00–09:30 *Korean NLP2RDF Resources*
YoungGyun Hahm, KyungTae Lim, Jungyeul Park, Yongun Yoon and Key-Sun Choi

09:30–10:00 *Building Large Scale Text Corpus for Tibetan Natural Language Processing by Extracting Text from Web Pages*
Huidan Liu, Minghua Nuo, Jian Wu and Yeping He

10:00–10:30 *A Structured Approach for Building Assamese Corpus: Insights, Applications and Challenges*
Prof. Shikhar Kr. Sarma, Himadri Bharali, Ambeswar Gogoi, Ratul Deka and Anup Kr Barman

10:30–11:00 *Corpus Building of Literary Lesser Rich Language-Bodo: Insights and Challenges*
Biswajit Brahma, Anup Kr. Barman, Prof. Shikhar Kr. Sarma and Bhatima Boro

11:00–11:30 Tea break

Session 2 – Morphology and Syntax Parsing

11:30–12:00 *Dependency Parsers for Persian*
Mojgan Seraji, Beáta Megyesi and Joakim Nivre

12:00–12:30 *A New DOP Model for Phrase-structure Parsing of Persian Sentences*
Zahra Sarabi and Morteza Analoui

12:30–13:00 *A Hybrid Dependency Parser for Bangla*
Arnab Dhar, Sanjay Chatterji, Sudeshna Sarkar and Anupam Basu

13:00–13:30 *Repairing Bengali Verb Chunks for Improved Bengali to Hindi Machine Translation*
Sanjay Chatterji, Nabanita Datta, Arnab Dhar, Biswanath Barik, Sudeshna Sarkar and Anupam Basu

13:30–14:30 Lunch

Sunday, 9 December 2012 (continued)

Session 3 – Knowledge Extraction

14:30–15:00 *Domain Specific Ontology Extractor For Indian Languages*
Brijesh Bhatt and Pushpak Bhattacharyya

15:00–15:30 *Constrained Hidden Markov Model for Bilingual Keyword Pairs Alignment*
Denny Cahyadi, Fabien Cromieres and Sadao Kurohashi

15:30–16:00 *N-gram and Gazetteer List Based Named Entity Recognition for Urdu: A Scarce Resourced Language*
Faryal Jahangir, Waqas Anwar, Usama Ijaz Bajwa and Xuan Wang

16:00–16:30 Tea break

Session 4 – Applications

16:30–17:00 *Developing a POS tagger for Magahi: A Comparative Study*
Ritesh Kumar, Bornini Lahiri and Deepak Alok

17:00–17:30 *Enhancing Lemmatization for Mongolian and its Application to Statistical Machine Translation*
Chimeddorj Odbayar and Atsushi Fujii

17:30–18:00 *Translations of Ambiguous Hindi Pronouns to Possible Bengali Pronouns*
Sanjay Chatterji, Sudeshna Sarkar and Anupam Basu