COLING 2012

24th International Conference on Computational Linguistics

Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)

Workshop organizers:
Dipti Misra Sharma, Prashanth Mannem,
Joseph van Genabith, Sobha Lalitha Devi,
Radhika Mamidi and Ranjani Parthasarathi

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 Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)
Dipti Misra Sharma, Prashanth Mannem, Joseph van Genabith, Sobha Lalitha Devi, Radhika Mamidi and Ranjani Parthasarathi (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

This volume © 2012 The COLING 2012 Organizing Committee.
Licensed under the *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Nonported* license.
http://creativecommons.org/licenses/by-nc-sa/3.0/
Some rights reserved.

Contributed content copyright the contributing authors.
Used with permission.

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

Preface

Indian Languages present taxing research challenges mostly attributed to their rich variation in morphology, heavy agglutination and relatively free word order. Most of the Indian languages are digitally under-resourced, and only limited linguistic analysis resources/tools exist for some languages. The objective of the workshop is to bring together MT and parsing researchers across the globe working on Indian languages to showcase their work and exploit the synergies to interconnect state-of-the-art Indian language MT and parsing research globally.

We received good response from researchers worldwide and based on reviews from our strong program committee 4 papers were accepted for oral presentation (long papers) and 9 papers for poster presentation (short papers). A wide range of languages including Hindi, Bengali, Tamil, Telugu, Urdu were covered in the accepted papers.

The workshop also hosted a dependency parsing shared task for Hindi. As part of the shared task, a part of the Hindi Treebank (HTB) containing gold standard morphological analyses, part-of-speech tags, chunks and dependency relations labeled in the computational Paninian framework was released. Evaluation was carried out over both gold standard and automatic parts of speech (also provided by us) for all the participating systems. Seven teams from both India and abroad participated in the shared task and submitted reports on their approaches.

We thank the members of program committee for their valuable support and cooperation for the workshop. We also thank them for giving detailed reviews to the authors. Finally, we thank the organizers of COLING 2012 for giving us the opportunity to organize this workshop.

Dipti Misra Sharma
Prashanth Mannem
Josef Van Genabith
Sobha Lalitha Devi
Radhika Mamidi
Ranjani Parthasarathi

Program Committee

Adil Kak, Kashmir University, India
Anoop Sarkar, Simon Fraser University, Canada
Aravind K Joshi, University of Pennsylvania, USA
Christian Boitet, University of Grenoble, France
Fei Xia, University of Washington, USA
Geetha T.V, Anna University, India
Gurpreet Singh Lehal, Punjabi University Patiala, India
Joakim Nivre, Uppsala, Sweden
Miriram Butt, University of Konstanz, Germany
Monojit Choudhury, Microsoft Research, India
Nilandri Chatterji, IIT Delhi, India
Nitin Madnani, ETS, USA
Ondrej Bojar, Charles University, Czech Republic
Owen Rambow, Columbia University
Pushpak Bhattacharya, IIT Bombay, India
Rajeev Sangal, IIIT Hyderabad, India
Rajendran S, Amrita University, India
Rajesh Bhatt, University of Massachusetts, USA
Sarmad Hussain, National University, Pakistan
Samar Husain, University of Potsdam, Germany
Sivaji Bandyopadhyay, Jadavpur University, India
Srinivas Bangalore, AT&T Labs, USA
Sriram Venkatapathy, XRCE, France
Vijay Sundar Ram R, AU-KBC Research Center, Chennai, India

Organizing Committee

Dipti Misra Sharma, LTRC, IIIT-Hyderabad, India (*Workshop Chair*)
Josef Van Genabith, CNGL, School of Computing, Dublin City University
Radhika Mamidi, LTRC, IIIT-Hyderabad
Ranjani Parthasarathi, Anna University, Chennai
Sobha Lalitha Devi, AU-KBC Research Center, Anna University
Prashanth Mannem, LTRC, IIIT-Hyderabad

Table of Contents

Automatic Annotation of Genitives in Hindi Treebank
Nitesh Surtani and Soma Paul 1

Semantic Parsing of Tamil Sentences
Balaji Jagan, Geetha T V and Ranjani Parthasarathi 15

Tamil NER – Coping with Real Time Challenges
Malarkodi C.S, Pattabhi RK Rao and Sobha Lalitha Devi 23

Sublexical Translations for Low-Resource Language
Khan Md. Anwarus Salam, Yamada Setsuo and Nishino Tetsuro 39

Introducing Kashmiri Dependency Treebank
Shahid Mushtaq Bhat 53

A Diagnostic Evaluation Approach Targeting MT Systems for Indian Languages
Renu Balyan, Sudip Kumar Naskar, Antonio Toral and Niladri Chatterjee 61

An Approach to Discourse Parsing using Sangati and Rhetorical Structure Theory
Subalalitha C.N. and Ranjani Parthasarathi 73

Clause Boundary Identification for Malayalam Using CRF
Lakshmi S, Vijay Sundar Ram R and Sobha Lalitha Devi 83

Disambiguation of pre/post positions in English – Malayalam Text Translation
Jayan V, Sunil R and Bhadran V K 93

Resolution for Pronouns in Tamil Using CRF
Akilandeswari A and Sobha Lalitha Devi 103

Morphological Processing for English-Tamil Statistical Machine Translation
Loganathan Ramasamy, Ondřej Bojar and Zdeněk Žabokrtský 113

Dative Case in Telugu: A Parsing Perspective
Umamaheshwar Rao Garapati, Rajyarama Koppaka and Srinivas Addanki 123

Evaluation of Two Bengali Dependency Parsers
Arjun Das, Arabinda Shee and Utpal Garain 133

CUNI: Feature Selection and Error Analysis of a Transition-Based Parser
Daniel Zeman 143

Parsing Hindi with MDParser
Alexander Volokh and Günter Neumann 149

A Three Stage Hybrid Parser for Hindi
Sanjay Chatterji, Arnad Dhar, Sudeshna Sarkar and Anupam Basu 155

Two-stage Approach for Hindi Dependency Parsing Using MaltParser
Naman Jain, Karan Singla, Aniruddha Tammewar and Sambhav Jain 163

Hindi Dependency Parsing using a combined model of Malt and MST
B. Venkata Seshu Kumari and Rajeswara Rao Ramisetty 171

Ensembling Various Dependency Parsers: Adopting Turbo Parser for Indian Languages
Puneeth Kukkadapu, Deepak Malladi and Aswarth Dara 179

ISI-Kolkata at MTPIL-2012
Arjun Das, Arabinda Shee and Utpal Garain 185

Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)

Program

Saturday, 15 December 2012

Session 1

09:30–10:30 **Invited Talk**
NLP in India: Past, Present and Future
Rajeev Sangal, IIIT-Hyderabad

10:30–11:00 *Automatic Annotation of Genitives in Hindi Treebank*
Nitesh Surtani and Soma Paul

11:00–11:30 *Semantic Parsing of Tamil Sentences*
Balaji Jagan, Geetha T V and Ranjani Parthasarathi

11:30–12:00 Tea break

Session 2

12:00–12:30 *Tamil NER – Coping with Real Time Challenges*
Malarkodi C.S, Pattabhi RK Rao and Sobha Lalitha Devi

12:30–13:00 *Sublexical Translations for Low-Resource Language*
Khan Md. Anwarus Salam, Yamada Setsuo and Nishino Tetsuro

13:00–13:10 *Introducing Kashmiri Dependency Treebank*
Shahid Mushtaq Bhat

13:10–13:20 *A Diagnostic Evaluation Approach Targeting MT Systems for Indian Languages*
Renu Balyan, Sudip Kumar Naskar, Antonio Toral and Niladri Chatterjee

13:20–13:30 *An Approach to Discourse Parsing using Sangati and Rhetorical Structure Theory*
Subalalitha C.N. and Ranjani Parthasarathi

13:30–14:30 Lunch

Saturday, 15 December 2012 (continued)

Session 3

14:30–14:40 *Clause Boundary Identification for Malayalam Using CRF*
Lakshmi S, Vijay Sundar Ram R and Sobha Lalitha Devi

14:40–14:50 *Disambiguation of pre/post positions in English – Malayalam Text Translation*
Jayan V, Sunil R and Bhadran V K

14:50–15:00 *Resolution for Pronouns in Tamil Using CRF*
Akilandeswari A and Sobha Lalitha Devi

15:00–15:10 *Morphological Processing for English-Tamil Statistical Machine Translation*
Loganathan Ramasamy, Ondřej Bojar and Zdeněk Žabokrtský

15:10–15:20 *Dative Case in Telugu: A Parsing Perspective*
Umamaheshwar Rao Garapati, Rajyarama Koppaka and Srinivas Addanki

15:20–15:30 *Evaluation of Two Bengali Dependency Parsers*
Arjun Das, Arabinda Shee and Utpal Garain

15:30–16:30 Poster session

16:30–17:00 Tea break

Session 5: Hindi Parsing Shared Task

17:00–17:15 *Overview of the Hindi Parsing Shared Task - 2012*
Akshar Bharati, Prashanth Mannem and Dipti Misra Sharma

17:15–17:25 *CUNI: Feature Selection and Error Analysis of a Transition-Based Parser*
Daniel Zeman

17:25–17:35 *Parsing Hindi with MDParser*
Alexander Volokh and Günter Neumann

17:35–17:45 *A Three Stage Hybrid Parser for Hindi*
Sanjay Chatterji, Arnad Dhar, Sudeshna Sarkar and Anupam Basu

17:45–17:55 *Two-stage Approach for Hindi Dependency Parsing Using MaltParser*
Naman Jain, Karan Singla, Aniruddha Tammewar and Sambhav Jain

17:55–18:05 *Hindi Dependency Parsing using a combined model of Malt and MST*
B. Venkata Seshu Kumari and Rajeswara Rao Ramisetty

18:05–18:15 *Ensembling Various Dependency Parsers: Adopting Turbo Parser for Indian Languages*
Puneeth Kukkadapu, Deepak Malladi and Aswarth Dara

18:15–18:25 *ISI-Kolkata at MTPIL-2012*
Arjun Das, Arabinda Shee and Utpal Garain