IJCNLP-08 Workshop

On

NLP for Less Privileged Languages

Proceedings of the Workshop

11 January 2008
IIIT, Hyderabad, India

Introduction

Welcome to the IJCNLP Workshop on NLP for Less Privileged Languages, a meeting held in conjunction with the Third International Joint Conference on Natural Language Processing at Hyderabad, India. The goal of this workshop is to ascertain the progress made in providing computational support for less computerized or ‘less privileged’ languages and in building language resources and Natural Language Processing tools etc. for such languages. An introductory article explains the background of and motivation for this workshop. It also presents an overview of the papers selected for the workshop.

The workshop attracted a lot of interest from around the world. There were a relatively large number of submissions and each paper was reviewed by three reviewers, which ensured that the quality of papers selected was comparable to other successful workshops held previously on similar themes. The selected papers include a variety of topics and covers a wide range of languages. Another major feature of the workshop is that it includes three invited talks by speakers from different regions of the world and on very different topics.

We would like to thank the program committee members for all the hard work that they did during the reviewing process. We would also like to thank all the people involved in organizing the IJCNLP conference. We hope that this workshop will be able to achieve its goal and will stimulate and encourage even more interest in the theme of the workshop so that the gap between languages like English and the less computerized ‘less privileged’ or languages can be reduced at a rapid pace.

Anil Kumar Singh (Chair)

Organizer:

Anil Kumar Singh, IIIT, Hyderabad, India

Program Committee:

Steven Bird, University of Melbourne, Australia
Rajeev Sangal, IIIT, Hyderabad, India
Michael Maxwell, University of Maryland, USA
Bente Maegaard, CST, University of Copenhagen, Denmark
Lakshmi Bai, IIIT, Hyderabad, India
Emily M. Bender, University of Washington, USA
Nicoletta Calzolari, Istituto di Linguistica Computazionale del CNR - Pisa, Italy
Alexander Gelbukh, Center for Computing Research, National Polytechnic Institute, Mexico
Sarmad Hussain, CRULP, Pakistan
Greville Corbett, University of Surrey, UK
Anil Kumar Singh, IIIT, Hyderabad, India
Sobha L., AU-KBC, Chennai, India
Rachel Edita Roxas, Dela Salle University, Manila, Philippines
Sivaji Bandyopadhyay, Jadavpur University, Kolkata, India
Nicholas Thieberger, University of Melbourne, Australia
Monojit Choudhury, Indian Institute of Technology, Kharagpur, India
Xabier Arregi, University of the Basque Country, Spain
Khalid Choukri, ELRA - Paris, France
Samar Husain, IIIT, Hyderabad, India
Indra Budi, University of Indonesia, Indonesia
Rajat Mohanty, Indian Institute of Technology, Mumbai, India
Jeff Good, University at Buffalo, USA
Prasad Pingali, IIIT, Hyderabad, India
Harshit Surana, IIIT, Hyderabad, India

Special Acknowledgment:

Samar Husain, IIIT, Hyderabad, India
Harshit Surana, IIIT, Hyderabad, India

Invited Speakers:

Anne David and Michael Maxwell, CASL, University of Maryland, USA
Virach Sornlertlamvanich, TCL, NICT, Thailand
Monojit Choudhury, Microsoft Research, India

Table of Contents

Conference Program

Friday, January 11, 2008

09:00-09:10 **Opening Remarks**: *Natural Language Processing for Less Privileged Languages: Where do we come from? Where are we going?*
Anil Kumar Singh

Session 1:

09:10-09:40 **Invited Talk**: *Building Language Resources: Ways to move forward*
Anne David and Micheal Maxwell

09:40-10:00 *KUI: an ubiquitous tool for collective intelligence development*
Thatsanee Charoenporn, Virach Sornlertlamvanich, Hitoshi Isahara and Kergrit Robkop

10:00-10:20 *Prototype Machine Translation System From Text-To-Indian Sign Language*
Tirthankar Dasgupta, Sandipan Dandpat and Anupam Basu

10:20-11:00 **Break**

Session 2:

11:00-11:30 **Invited Talk**: *Cross Language Resource Sharing*
Virach Sornlertlamvanich

11:30-11:50 *Joint Grammar Development by Linguists and Computer Scientists*
Michael Maxwell and Anne David

11:50-12:10 *Cross-Language Parser Adaptation between Related Languages*
Daniel Zeman and Philip Resnik

12:10-12:30 *SriShell Primo: A Predictive Sinhala Text Input System*
Sandeva Goonetilleke, Yoshihiko Hayashi, Yuichi Itoh and Fumio Kishino

12:30-14:00 **Lunch**

Session 3:

14:00-14:30 **Invited Talk**: *Breaking the Zipfian Barrier of NLP*
Monojit Choudhury

Friday, January 11, 2008 (continued)

14:30-15:30 **Poster Display and Discussion**

An Optimal Order of Factors for the Computational Treatment of Personal Anaphoric Devices in Urdu Discourse
Mohammad Naveed Ali, M. A. Khan and Muhammad Aamir Khan

Morphology Driven Manipuri POS Tagger
Thoudam Doren Singh and Sivaji Bandyopadhyay

Acharya - A Text Editor and Framework for working with Indic Scripts
Krishnakumar V and Indrani Roy

Implementing a Speech Recognition System Interface for Indian Languages
R.K. Aggarwal and M. Dave

Indigenous Languages of Indonesia: Creating Language Resources for Language Preservation
Hammam Riza

Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields
Chirag Patel and Karthik Gali

Speech to speech machine translation: Biblical chatter from Finnish to English
David Ellis, Mathias Creutz, Timo Honkela and Mikko Kurimo

15:30-16:00 **Break**

Session 4:

16:00-16:20 *A Rule-based Syllable Segmentation of Myanmar Text*
Zin Maung Maung and Yoshiki Mikami

16:20-16:40 *Strategies for sustainable MT for Basque: incremental design, reusability, standardization and open-source*
I. Alegria, X. Arregi, X. Artola, A. Diaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor and K. Sarasola

16:40-17:00 *Design of a Rule-based Stemmer for Natural Language Text in Bengali*
Sandipan Sarkar and Sivaji Bandyopadhyay

17:00-17:20 *Finite State Solutions For Reduplication In Kinyarwanda Language*
Jackson Muhirwe and Trond Trosterud

17:20-18:00 **Closing Discussion**