LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation

PROCEEDINGS

Editors:
Girish Nath Jha, Kalika Bali, Sobha L, S. S. Agrawal, Atul Kr. Ojha

Proceedings of the LREC 2020 WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation

Edited by: Girish Nath Jha, Kalika Bali, Sobha L, S. S. Agrawal, Atul Kr. Ojha

ISBN: 979-10-95546-67-2
EAN: 9791095546672

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Introduction

WILDRE – the 5th Workshop on Indian Language Data: Resources and Evaluation is being organized in Marseille, France on May 16th, 2020 under the LREC platform. India has a huge linguistic diversity and has seen concerted efforts from the Indian government and industry towards developing language resources. European Language Resource Association (ELRA) and its associate organizations have been very active and successful in addressing the challenges and opportunities related to language resource creation and evaluation. It is, therefore, a great opportunity for resource creators of Indian languages to showcase their work on this platform and also to interact and learn from those involved in similar initiatives all over the world.

The broader objectives of the 5th WILDRE will be

- to map the status of Indian Language Resources
- to investigate challenges related to creating and sharing various levels of language resources
- to promote a dialogue between language resource developers and users
- to provide an opportunity for researchers from India to collaborate with researchers from other parts of the world

The call for papers received a good response from the Indian language technology community. Out of nineteen full papers received for review, we selected one paper for oral, four for short oral and seven for a poster presentation.

Workshops Chairs

Girish Nath Jha, Jawaharlal Nehru University, India
Kalika Bali, Microsoft Research India Lab, Bangalore
Sobha L, AU-KBC, Anna University
S. S. Agrawal, KIIT, Gurgaon, India

Workshop Manager

Atul Kr. Ojha, Charles University, Prague, Czech Republic & Panlingua Language Processing LLP, India

Editors

Girish Nath Jha, Jawaharlal Nehru University, India
Kalika Bali, Microsoft Research India Lab, Bangalore
Sobha L, AU-KBC, Anna University
S. S. Agrawal, KIIT, Gurgaon, India
Atul Kr. Ojha, Charles University, Prague, Czech Republic & Panlingua Language Processing LLP, India

Programme Committee

Adil Amin Kak, Kashmir University
Anil Kumar Singh, IIT BHU, Benaras
Anupam Basu, Director, NIIT, Durgapur
Anoop Kunchukuttan, Microsoft AI and Research, India
Arul Mozhi, University of Hyderabad
Asif Iqbal, IIT Patna, Patna
Atul Kr. Ojha, Charles University, Prague, Czech Republic & Panlingua Language Processing LLP, India
Bogdan Babych, University of Leeds, UK
Chao-Hong Liu, ADAPT Centre, Dublin City University, Ireland
Claudia Soria, CNR-ILC, Italy
Dafydd Gibbon, Universität Bielefeld, Germany
Daan van Esch, Google, USA
Dan Zeman, Charles University, Prague, Czech Republic
Delyth Prys, Bangor University, UK
Dipti Mishra Sharma, IIIT, Hyderabad
Diwakr Mishra, Amazon-Banglore, India
Dorothee Beermann, Norwegian University of Science and Technology (NTNU)
Elizabeth Sherley, IITM-Kerala, Trivandrum
Esha Banerjee, Google, USA
Eveline Wandl-Vogt, Austrian Academy of Sciences, Austria
Georg Rehm, DFKI, Germany
Girish Nath Jha, Jawaharlal Nehru University, New Delhi
Jan Odijk, Utrecht University, The Netherlands
Jolanta Bachan, Adam Mickiewicz University, Poland
Joseph Mariani, LIMSI-CNRS, France
Jyoti D. Pawar, Goa University
Kalika Bali, MSRI, Bangalore
Khalid Choukri, ELRA, France
Lars Hellan, NTNU, Norway
M J Warsi, Aligarh Muslim University, India
Malhar Kulkarni, IIT Bombay
Manji Bhadra, Bankura University, West Bengal
Marko Tadic, Croatian Academy of Sciences and Arts, Croatia
Massimo Monaglia, University of Florence, Italy
Monojit Choudhary, MSRI Bangalore
Narayan Choudhary, CIIL, Mysore
Nicoletta Calzolari, ILC-CNR, Pisa, Italy
Niladri Shekhar Dash, ISI Kolkata
Panchanan Mohanty, GLA, Mathura
Pinky Nainwani, Cognizant Technology Solutions, Bangalore
Pushpak Bhattacharya, Director, IIT Patna
Qun Liu, Noah's Ark Lab, Huawei
Rajeev R R, ICFOSS, Trivandrum

Ritesh Kumar, Agra University
Shantipriya Parida, Idiap Research Institute, Switzerland
S.K. Shrivastava, Head, TDIL, MEITY, Govt of India
S.S. Agrawal, KIIT, Gurgaon, India
Sachin Kumar, EZDI, Ahmedabad
Santanu Chaudhury, Director, IIT Jodhpur
Shivaji Bandhopadhyay, Director, NIT, Silchar
Sobha L, AU-KBC Research Centre, Anna University
Stelios Piperidis, ILSP, Greece
Subhash Chandra, Delhi University
Swaran Lata, Retired Head, TDIL, MCIT, Govt of India
Virach Sornlertlamvanich, Thammasat University, Bangkok, Thailand
Vishal Goyal, Punjabi University, Patiala
Zygmunt Vetulani, Adam Mickiewicz University, Poland

Table of Contents

Workshop Program

Saturday, May 16, 2020

14:00–14:45 **Inaugural session**

14:00–14:05 ***Welcome by Workshop Chairs***

14:05–14:25 ***Inaugural Address***

14:25–14:45 ***Keynote Lecture***

14:45–16:15 **Paper Session**

Part-of-Speech Annotation Challenges in Marathi
Gajanan Rane, Nilesh Joshi, Geetanjali Rane, Hanumant Redkar, Malhar Kulkarni and Pushpak Bhattacharyya

A Dataset for Troll Classification of TamilMemes
Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae and Paul Buitelaar

OdiEnCorp 2.0: Odia-English Parallel Corpus for Machine Translation
Shantipriya Parida, Satya Ranjan Dash, Ondřej Bojar, Petr Motlicek, Priyanka Pattnaik and Debasish Kumar Mallick

Handling Noun-Noun Coreference in Tamil
Vijay Sundar Ram and Sobha Lalitha Devi

Malayalam Speech Corpus: Design and Development for Dravidian Language
Lekshmi K R, Jithesh V S and Elizabeth Sherly

16:15–16:25 ***Break***

Saturday, May 16, 2020 (continued)

16:25–17:45 **Poster Session**

Multilingual Neural Machine Translation involving Indian Languages
Pulkit Madaan and Fatiha Sadat

Universal Dependency Treebanks for Low-Resource Indian Languages: The Case of Bhojpuri
Atul Kr. Ojha and Daniel Zeman

A Fully Expanded Dependency Treebank for Telugu
Sneha Nallani, Manish Shrivastava and Dipti Sharma

Determination of Idiomatic Sentences in Paragraphs Using Statement Classification and Generalization of Grammar Rules
Naziya Shaikh

Polish Lexicon-Grammar Development Methodology as an Example for Application to other Languages
Zygmunt Vetulani and Grażyna Vetulani

Abstractive Text Summarization for Sanskrit Prose: A Study of Methods and Approaches
Shagun Sinha and Girish Jha

A Deeper Study on Features for Named Entity Recognition
Malarkodi C S and Sobha Lalitha Devi

17:45–18:30 ***Panel discussion***

Saturday, May 16, 2020 (continued)

18:30–18:40 **Valedictory Address**

18:40–18:45 ***Vote of Thanks***