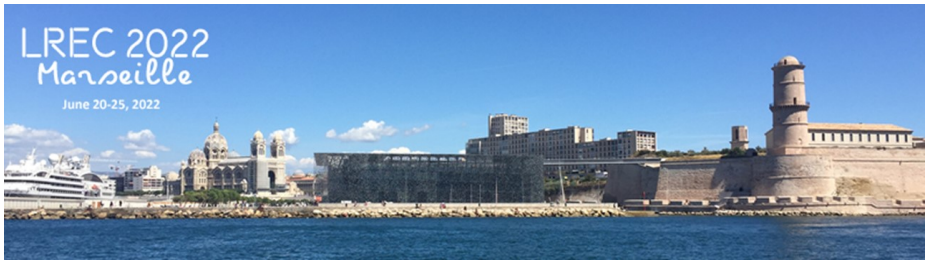

LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

6th Workshop on Indian Language Data: Resources and Evaluation (WILDRE-6)

PROCEEDINGS

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

Proceedings of the 6thWorkshop on Indian Language Data: Resources and Evaluation (WILDRE-6 2022)

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

ISBN: 979-10-95546-87-0
EAN: 9791095546870

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

Preface

WILDRE – the 6th Workshop on Indian Language Data: Resources and Evaluation is being organized in Marseille, France on June 20th, 2022 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 6th 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. This year, we selected only three papers for oral, and thirteen for a poster presentations.

Workshop Chairs

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

Workshop Organizers

Atul Kr. Ojha, National University of Ireland Galway, Ireland & Panlingua Language Processing LLP, India
Girish Nath Jha, Jawaharlal Nehru University, India
Kalika Bali, Microsoft Research India Lab, Bangalore
Sobha L, AU-KBC, Anna University

Program Committee

Adil Amin Kak, Kashmir University
Alessandro Panunzi, University of Florence, Italy
Arul Mozhi, University of Hyderabad
Atul Kr. Ojha, National University of Ireland Galway, Ireland & Panlingua Language Processing LLP, India
Bharathi Raja Asoka Chakravarthi, National University of Ireland Galway, Irelnad
Bogdan Babych, Heidelberg University, Germany
Chao-Hong Liu, Potamu Research Ltd, 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
Dorothee Beermann, Norwegian University of Science and Technology (NTNU)
Elizabeth Sherley, IITM-Kerala, Trivandrum
Esha Banerjee, Google
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, Microsoft Research India Lab, Bangalore
Khalid Choukri, ELRA, France
Lars Hellan, NTNU, Norway
Malhar Kulkarni, IIT Bombay
Massimo Monaglia, University of Florence, Italy
Monojit Choudhary, MSRI Bangalore
Nicoletta Calzolari, ILC-CNR, Pisa, Italy
Niladri Shekhar Dash, ISI Kolkata
Partha Talukdar, Google Research, India

Pinky Nainwani, Cognizant Technology Solutions, Bangalore
Pushpak Bhattacharya, IIT Bombay
Rajeev R R, ICFOSS, Trivandrum
Ritesh Kumar, Agra University
Shantipriya Parida, Silo AI, Helsinki, Finland
S.S. Agrawal, KIIT, Gurgaon, India
Sobha L, AU-KBC Research Centre, Anna University
Stelios Piperidis, ILSP, Greece
Subhash Chandra, Delhi University
Vishal Goyal, Punjabi University, Patiala
Zygmunt Vetulani, Adam Mickiewicz University, Poland

Table of Contents

Introducing EM-FT for Manipuri-English Neural Machine Translation
Rudali Huidrom and Yves Lepage 1

L3Cube-HingCorpus and HingBERT: A Code Mixed Hindi-English Dataset and BERT Language Models
Ravindra Nayak and Raviraj Joshi 7

Leveraging Sub Label Dependencies in Code Mixed Indian Languages for Part-Of-Speech Tagging using Conditional Random Fields.
Akash Kumar Gautam 13

HindiWSD: A package for word sense disambiguation in Hinglish & Hindi
Mirza Yusuf, Praatibh Surana and Chethan sharma 18

Pāṇinian Phonological Changes: Computation and Development of Online Access System
Sanju . and Subhash Chandra 24

L3Cube-MahaNER: A Marathi Named Entity Recognition Dataset and BERT models
Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade and Raviraj Joshi 29

Identifying Emotions in Code Mixed Hindi-English Tweets
Sanket Sonu, Rejwanul Haque, Mohammed Hasanuzzaman, Paul Stynes and Pramod Pathak . . . 35

Digital Accessibility and Information Mining of Dharmaśāstric Knowledge Traditions
Arooshi Nigam and Subhash Chandra 42

Language Resource Building and English-to-Mizo Neural Machine Translation Encountering Tonal Words
Vanlalmuansangi Khenglawt, Sahinur Rahman Laskar, Santanu Pal, Partha Pakray and Ajoy Kumar Khan 48

Classification of Multiword Expressions in Malayalam
Treesa Cyriac and Sobha Lalitha Devi 55

Bengali and Magahi PUD Treebank and Parser
Pritha Majumdar, Deepak Alok, Akanksha Bansal, Atul Kr. Ojha and John P. McCrae 60

Makadi: A Large-Scale Human-Labeled Dataset for Hindi Semantic Parsing
Shashwat Vaibhav and Nisheeth Srivastava 68

Automatic Identification of Explicit Connectives in Malayalam
Kumari Sheeja S and Sobha Lalitha Devi 74

Web based System for Derivational Process of Kṛdanta based on Pāṇinian Grammatical Tradition
Sumit Sharma and Subhash Chandra 80

Universal Dependency Treebank for Odia Language
Shantipriya Parida, Kalyanamalini Shabadi, Atul Kr. Ojha, Saraswati Sahoo, Satya Ranjan Dash and Bijayalaxmi Dash 84

Computational Referencing System for Sanskrit Grammar
Baldev Khandoliyan and Ram Kishor 90

L3Cube-MahaCorpus and MahaBERT: Marathi Monolingual Corpus, Marathi BERT Language Models, and Resources
Raviraj Joshi 97

Conference Program

Monday, June 20, 2022

14:00–14:45 Inaugural session

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

14:25–15:00 ***Keynote Lecture***

15:00–16:00 Oral Session-I

15:00–15:30 *Introducing EM-FT for Manipuri-English Neural Machine Translation*
Rudali Huidrom and Yves Lepage

15:30–16:00 *L3Cube-HingCorpus and HingBERT: A Code Mixed Hindi-English Dataset and BERT Language Models*
Ravindra Nayak and Raviraj Joshi

16:00–16:30 Coffee break/Poster Session

16:00–16:30 *Leveraging Sub Label Dependencies in Code Mixed Indian Languages for Part-Of-Speech Tagging using Conditional Random Fields.*
Akash Kumar Gautam

16:00–16:30 *HindiWSD: A package for word sense disambiguation in Hinglish & Hindi*
Mirza Yusuf, Praatibh Surana and Chethan sharma

16:00–16:30 *Pāṇinian Phonological Changes: Computation and Development of Online Access System*
Sanju . and Subhash Chandra

16:00–16:30 *L3Cube-MahaNER: A Marathi Named Entity Recognition Dataset and BERT models*
Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade and Raviraj Joshi

16:00–16:30 *Identifying Emotions in Code Mixed Hindi-English Tweets*
Sanket Sonu, Rejwanul Haque, Mohammed Hasanuzzaman, Paul Stynes and Pramod Pathak

Monday, June 20, 2022 (continued)

16:00–16:30 *Digital Accessibility and Information Mining of Dharmaśāstric Knowledge Traditions*
Arooshi Nigam and Subhash Chandra

16:00–16:30 *Language Resource Building and English-to-Mizo Neural Machine Translation Encountering Tonal Words*
Vanlalmuansangi Khenglawt, Sahinur Rahman Laskar, Santanu Pal, Partha Pakray and Ajoy Kumar Khan

16:00–16:30 *Classification of Multiword Expressions in Malayalam*
Treesa Cyriac and Sobha Lalitha Devi

16:00–16:30 *Bengali and Magahi PUD Treebank and Parser*
Pritha Majumdar, Deepak Alok, Akanksha Bansal, Atul Kr. Ojha and John P. McCrae

16:00–16:30 *Makadi: A Large-Scale Human-Labeled Dataset for Hindi Semantic Parsing*
Shashwat Vaibhav and Nisheeth Srivastava

16:00–16:30 *Automatic Identification of Explicit Connectives in Malayalam*
Kumari Sheeja S and Sobha Lalitha Devi

16:00–16:30 *Web based System for Derivational Process of Kṛdanta based on Pāṇinian Grammatical Tradition*
Sumit Sharma and Subhash Chandra

16:00–16:30 *Universal Dependency Treebank for Odia Language*
Shantipriya Parida, Kalyanamalini Shabadi, Atul Kr. Ojha, Saraswati Sahoo, Satya Ranjan Dash and Bijayalaxmi Dash

16:00–16:30 *Computational Referencing System for Sanskrit Grammar*
Baldev Khandoliyan and Ram Kishor

Monday, June 20, 2022 (continued)

16:30–17:00 Oral Session-II

16:30–17:00 *L3Cube-MahaCorpus and MahaBERT: Marathi Monolingual Corpus, Marathi BERT Language Models, and Resources*
Raviraj Joshi

17:00–17:45 Panel discussion

17:45–17:55 ***Valedictory Session***

17:55–18:00 ***Vote of Thanks***