IndoNLP 2025

Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages (IndoNLP2025)

Proceedings of the Workshop

January 20, 2025

©2025 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-214-5

The workshop is supported in part by the Informatics Institute of Technology, Colombo, Sri Lanka

Preface

The rapid advancement of Natural Language Processing (NLP) and Large Language Models (LLMs) has transformed the landscape of computational linguistics. However, Indo-Aryan and Dravidian Languages (IADL), which represent a significant portion of South Asia's linguistic heritage, remain under-resourced and under-researched in these technological developments. This workshop aims to bridge this gap by bringing together researchers, linguists, and technologists to focus on the unique challenges and opportunities. Participants will explore innovative methods for creating and annotating digital corpora, develop speech and language technologies suited to IADL, and promote interdisciplinary collaborations. By leveraging LLMs, we seek to address the complexities of syntax, morphology, and semantics in these languages to enhance the performance of NLP applications. Furthermore, the workshop will provide a platform for sharing best practices, tools, and resources, enhancing the digital infrastructure necessary for language preservation. Through collaborative efforts, we aim to build a research community to advance NLP for IADL, contributing to linguistic diversity and cultural preservation in the digital age.

In parallel with the workshop, we have also organised a shared task to address key challenges in transliteration for Indian languages. The primary objectives of the shared task are to develop a real-time transliterator, effectively manage linguistic variations, and improve typing accuracy. A significant focus of the task is on enabling the transliterator to handle ad-hoc transliterations, which involve short typing scripts and diverse typing patterns, with or without vowel combinations. This initiative aims to create a robust transliteration system that accommodates the dynamic and complex nature of typing practices in Indian languages.

We received 27 submissions for the workshop and shared task. Following the review process, we accepted 15 papers and 4 shared task submissions to appear in the workshop proceedings.

The success of IndoNLP 2025 would not have been possible without the contributions of several exceptional individuals who supported this initiative. First and foremost, we extend our heartfelt gratitude to the authors who submitted their work to the workshop, driving forward research in low-resource languages across diverse areas of study. We are equally thankful to the program committee members, whose dedicated efforts were instrumental to the success of this workshop. Their timely engagement in the review process and constructive feedback not only enhanced the quality of the submissions but also ensured that the papers met the highest academic standards. Moreover we would like to thank to Prof. Pushpak Bhattacharyya for accepting our invitation to be as the keynote speaker in the workshop. Finally, we would like to express our sincere gratitude to the Informatics Institute of Technology, Colombo, for their generous sponsorship of the workshop. We are truly thankful to everyone who contributed to the success of IndoNLP 2025 through their invaluable support and encouragement.

Organizing Committee

Ruvan Weerasinghe,Informatics Institute of Technology, Sri Lanka
Isuri Anuradha, Lancaster University, UK
Deshan Sumanathilaka, Swansea University, UK
Mo El-Haj, Lancaster University, UK
Chamila Liyanage, University of Colombo School of Computing, Sri Lanka
Fahad Khan, Istituto di Linguistica Computazionale in CNR, Italy
Andrew Hardie, Lancaster University, UK
Asim Abbas, Birmingham University, UK
Ruslan Mitkov Lancaster University, UK
Paul Rayson, Lancaster University, UK
Julian Hough, Swansea University, UK
Nicholas Micallef, Swansea University, UK
Naomi Krishnarajah, Informatics Institute of Technology, Sri Lanka

Program Committee

Abdullah Alzahrani, Swansea University, Wales, UK
Abdul Nazeer, National Institute of Technology, Calicut, India
Arka Majhi, Indian Institute of Technology, Bombay, India
Anand Kumar, National Institute of Technology, Karnataka, India
Asanka Wasala, Dell Technologies, Ireland
Arjumand Younus, University College Dublin, Ireland
Ayush Agarwal, Walmart, USA
Dulip Herath, Queensland University, Australia
Daisy Lal, Lancaster University, UK
Damith Premasiri, Lancaster University, UK
Gayanath Chandrasena, University of Helsinki, Finland
Girish Nath Jha, School for Sanskrit and Indic Studies, JNU, India
Jiby Mariya Jose, Indian Institute of Information Technology, India
Kishorjit Nongmeikapam, Indian Institute of Information Technology (IIIT) Manipur, India
Kengatharaiyer Sarveswaran, University of Jaffna, Sri Lanka
Kaza Sri Sai Swaroop, IBM, India
Lochandaka Ranathunga, University of Moratuwa, Sri Lanka
Nishantha Medagoda, Auckland University of Technology, New Zealand
Pabitra Mitra, Indian Institute of Technology, Kharagpur, India
Prasan Yapa, Kyoto University of Advance Science, Japan
Pumudu Fernando,Informatics Institute of Technology, Sri Lanka
Randil Pushpanandha, University of Colombo, Sri Lanka
Saman Galgodage, Swansea University, UK
Sinnathamby Mahesan, University of Jaffna, Sri Lanka
Torin Wirasinghe,Informatics Institute of Technology, Sri Lanka
Tanmoy Chakraborty, Indian Institute of Technology, Delhi, India
Tirthankar Dasgupta, Indian Institute of Technology, Kharagpur, India
Venkatesh Raju, Stealth Mode AI Startup, India

Table of Contents

Crossing Language Boundaries: Evaluation of Large Language Models on Urdu-English Question Answering
Samreen kazi, Maria Rahim and Shakeel Ahmed Khoja 1

Hindi Reading Comprehension: Do Large Language Models Exhibit Semantic Understanding?
Daisy Monika Lal, Paul Rayson and Mo El-Haj 12

Machine Translation and Transliteration for Indo-Aryan Languages: A Systematic Review
Sandun Sameera Perera and Deshan Koshala Sumanathilaka 23

Investigating the Effect of Backtranslation for Indic Languages
Sudhansu Bala Das, Samujjal Choudhury, Dr Tapas Kumar Mishra and Dr Bidyut Kr Patra 34

BERTopic for Topic Modeling of Hindi Short Texts: A Comparative Study
Atharva Mutsaddi, Anvi Jamkhande, Aryan Shirish Thakre and Yashodhara Haribhakta 49

Evaluating Structural and Linguistic Quality in Urdu DRS Parsing and Generation through Bidirectional Evaluation
Muhammad Saad Amin, Luca Anselma and Alessandro Mazzei 60

Studying the Effect of Hindi Tokenizer Performance on Downstream Tasks
Rashi Goel and Fatiha Sadat 71

Adapting Multilingual LLMs to Low-Resource Languages using Continued Pre-training and Synthetic Corpus: A Case Study for Hindi LLMs
Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar and Eileen Long 77

OVQA: A Dataset for Visual Question Answering and Multimodal Research in Odia Language
Shantipriya Parida, Shashikanta Sahoo, Sambit Sekhar, Kalyanamalini Sahoo, Ketan Kotwal, Sonal Khosla, Satya Ranjan Dash, Aneesh Bose, Guneet Singh Kohli, Smruti Smita Lenka and Ondřej Bojar 85

Advancing Multilingual Speaker Identification and Verification for Indo-Aryan and Dravidian Languages
Braveenan Sritharan and Uthayasanker Thayasivam 93

Sentiment Analysis of Sinhala News Comments Using Transformers
Isuru Bandaranayake and Hakim Usoof 100

ExMute: A Context-Enriched Multimodal Dataset for Hateful Memes
Riddhiman Swanan Debnath, Nahian Beente Firuj, Abdul Wadud Shakib, Sadia Sultana and Md Saiful Islam 109

Studying the capabilities of Large Language Models in solving Combinatorics Problems posed in Hindi
Yash Kumar and Subhajit Roy 116

From Scarcity to Capability: Empowering Fake News Detection in Low-Resource Languages with LLMs
Hrithik Majumdar Shibu, Shrestha Datta, Md. Sumon Miah, Nasrullah Sami, Mahruba Sharmin Chowdhury and Md Saiful Islam 126

Enhancing Participatory Development Research in South Asia through LLM Agents System: An Empirically-Grounded Methodological Initiative from Field Evidence in Sri Lankan
Xinjie Zhao, Hao Wang, Shyaman Maduranga Sriwarnasinghe, Jiacheng Tang, Shiyun Wang, Sayaka Sugiyama and So Morikawa 134

Identifying Aggression and Offensive Language in Code-Mixed Tweets: A Multi-Task Transfer Learning Approach
Bharath Kancharla, Prabhjot Singh, Lohith Bhagavan Kancharla, Yashita Chama and Raksha Sharma 148

Team IndiDataMiner at IndoNLP 2025: Hindi Back Transliteration - Roman to Devanagari using LLaMa
Saurabh Kumar, Dhruvkumar Babubhai Kakadiya and Sanasam Ranbir Singh 155

IndoNLP 2025 Shared Task: Romanized Sinhala to Sinhala Reverse Transliteration Using BERT
Sandun Sameera Perera, Lahiru Prabhath Jayakodi, Deshan Koshala Sumanathilaka and Isuri Anuradha 161

Sinhala Transliteration: A Comparative Analysis Between Rule-based and Seq2Seq Approaches
Widanalage Mario Yomal De Mel, Kasun Imesha Wickramasinghe, Nisansa de Silva and Surangika Dayani Ranathunga 168

Romanized to Native Malayalam Script Transliteration Using an Encoder-Decoder Framework
Bajiyo Baiju, Kavya Manohar, Leena G. Pillai and Elizabeth Sherly 177

Conference Program

8.45–9.00 **Opening Remark**

9.00–10.00 **Keynote Speech**

Theme: Language Processing and Evaluation

10.00–10.15 *Crossing Language Boundaries: Evaluation of Large Language Models on Urdu-English Question Answering*
Samreen kazi, Maria Rahim and Shakeel Ahmed Khoja

10.15–10.30 *Hindi Reading Comprehension: Do Large Language Models Exhibit Semantic Understanding?*
Daisy Monika Lal, Paul Rayson and Mo El-Haj

Coffee Break

11.00–11.15 *Machine Translation and Transliteration for Indo-Aryan Languages: A Systematic Review*
Sandun Sameera Perera and Deshan Koshala Sumanathilaka

11.15–11.30 *Investigating the Effect of Backtranslation for Indic Languages*
Sudhansu Bala Das, Samujjal Choudhury, Dr Tapas Kumar Mishra and Dr Bidyut Kr Patra

11.30–11.45 *BERTopic for Topic Modeling of Hindi Short Texts: A Comparative Study*
Atharva Mutsaddi, Anvi Jamkhande, Aryan Shirish Thakre and Yashodhara Haribhakta

11.45–12.00 *Evaluating Structural and Linguistic Quality in Urdu DRS Parsing and Generation through Bidirectional Evaluation*
Muhammad Saad Amin, Luca Anselma and Alessandro Mazzei

12.00–12.15 *Studying the Effect of Hindi Tokenizer Performance on Downstream Tasks*
Rashi Goel and Fatiha Sadat

12.15–12.30 *Adapting Multilingual LLMs to Low-Resource Languages using Continued Pretraining and Synthetic Corpus: A Case Study for Hindi LLMs*
Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar and Eileen Long

12.30–12.45 *OVQA: A Dataset for Visual Question Answering and Multimodal Research in Odia Language*
Shantipriya Parida, Shashikanta Sahoo, Sambit Sekhar, Kalyanamalini Sahoo, Ketan Kotwal, Sonal Khosla, Satya Ranjan Dash, Aneesh Bose, Guneet Singh Kohli, Smruti Smita Lenka and Ondřej Bojar

12.45–13.00 *Advancing Multilingual Speaker Identification and Verification for Indo-Aryan and Dravidian Languages*
Braveenan Sritharan and Uthayasanker Thayasivam

13.00–14.00 Lunch Break

Theme: Applications and Societal Impact: Applying NLP to Real-World Problems and Societal Challenges

14.00–14.15 *Sentiment Analysis of Sinhala News Comments Using Transformers*
Isuru Bandaranayake and Hakim Usoof

14.15–14.30 *ExMute: A Context-Enriched Multimodal Dataset for Hateful Memes*
Riddhiman Swanan Debnath, Nahian Beente Firuj, Abdul Wadud Shakib, Sadia Sultana and Md Saiful Islam

14.30–14.45 *Studying the capabilities of Large Language Models in solving Combinatorics Problems posed in Hindi*
Yash Kumar and Subhajit Roy

14.45–15.00 *From Scarcity to Capability: Empowering Fake News Detection in Low-Resource Languages with LLMs*
Hrithik Majumdar Shibu, Shrestha Datta, Md. Sumon Miah, Nasrullah Sami, Mahruba Sharmin Chowdhury and Md Saiful Islam

15.00–15.15 *Enhancing Participatory Development Research in South Asia through LLM Agents System: An Empirically-Grounded Methodological Initiative from Field Evidence in Sri Lankan*
Xinjie Zhao, Hao Wang, Shyaman Maduranga Sriwarnasinghe, Jiacheng Tang, Shiyun Wang, Sayaka Sugiyama and So Morikawa

15.15–15.30 *Identifying Aggression and Offensive Language in Code-Mixed Tweets: A Multi-Task Transfer Learning Approach*
Bharath Kancharla, Prabhjot Singh, Lohith Bhagavan Kancharla, Yashita Chama and Raksha Sharma

15.30–16.00 Coffee Break

Shared Task Discussion

Team IndiDataMiner at IndoNLP 2025: Hindi Back Transliteration - Roman to Devanagari using LLaMa
Saurabh Kumar, Dhruvkumar Babubhai Kakadiya and Sanasam Ranbir Singh

IndoNLP 2025 Shared Task: Romanized Sinhala to Sinhala Reverse Transliteration Using BERT
Sandun Sameera Perera, Lahiru Prabhath Jayakodi, Deshan Koshala Sumanathilaka and Isuri Anuradha

Sinhala Transliteration: A Comparative Analysis Between Rule-based and Seq2Seq Approaches
Widanalage Mario Yomal De Mel, Kasun Imesha Wickramasinghe, Nisansa de Silva and Surangika Dayani Ranathunga

Romanized to Native Malayalam Script Transliteration Using an Encoder-Decoder Framework
Bajiyo Baiju, Kavya Manohar, Leena G. Pillai and Elizabeth Sherly

Final Remark