COLING 2025

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)

Editors
Kengatharaiyer Sarveswaran
Surendrabikram Thapa
Sana Shams
Ashwini Vaidya
Bal Krishna Bal

CHiPSAL 2025 was co-located with the 31st International Conference on Computational Linguistics

January 19, 2025

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)

CHiPSAL 2025 was co-located with the 31st International Conference on Computational Linguistics (COLING 2025) and held virtually on 19 January 2025.

ISBN 979-8-89176-201-5

Message from the Organizing Chairs

Welcome to the proceedings of CHiPSAL 2025, the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL), held as part of the 31st International Conference on Computational Linguistics (COLING 2025) in Abu Dhabi, UAE, on January 19, 2025. This inaugural workshop, conducted in virtual mode, served as a platform to explore challenges and foster collaboration in processing South Asian languages.

The proceedings include highlights, challenges, and future directions from the workshop, presented in "A Brief Overview of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)."

CHiPSAL featured regular papers, invited keynotes, and shared task papers, with a focus on Devanagari-script language understanding. Subtasks included language identification, hate speech detection, and target classification. These contributions reflect the workshop's mission to address linguistic and cultural nuances, resource constraints, and orthographic complexities in low-resource South Asian languages while advancing multilingual NLP research.

We extend our heartfelt thanks to the program committee members worldwide for their rigorous reviews, ensuring three reviews per submission. We also express gratitude to the authors for their valuable contributions, as well as the COLING workshop chairs and the COLING 2025 organizing committees for their support in making this workshop a success.

We congratulate all authors on their accepted papers and are proud to note that CHiPSAL was a highly competitive workshop. We hope it provided a meaningful platform for discussing the challenges and future directions in South Asian language processing.

Thank you for being part of this inaugural event.

Kengatharaiyer Sarveswaran
Ashwini Vaidya
Bal Krishna Bal
Sana Shams
Surendrabikram Thapa

https://sites.google.com/view/chipsal/

Workshop Chairs

Kengatharaiyer Sarveswaran, University of Jaffna, Sri Lanka.
Ashwini Vaidya, Indian Institute of Technology, Delhi, India.
Bal Krishna Bal, Kathmandu University, Kathmandu, Nepal.
Sana Shams, University of Engineering and Technology, Lahore, Pakistan.
Surendrabikram Thapa, Virginia Tech, USA.

Program Committee Members (Alphabetical Order)

A M Abirami, Thiagarajar College of Engineering, India.
Abhai Pratap Singh, Amazon, USA.
Akaash Vishal Hazarika, Splunk, USA.
Aloka Fernando, University of Moratuwa, Sri Lanka.
Aman Shakya, Institute of Engineering, Pulchowk, Tribhuvan University, Nepal.
Anitha Dhakshina Moorthy, Thiagarajar College of Engineering, India.
Ann Sinthusha Anton Vijeevaraj, University of Vavuniya, Sri Lanka.
Annette Hautli-Janisz, University of Passau, Germany.
Ashwini Vaidya, IIT Delhi, India.
Bal Krishna Bal, Kathmandu University, Nepal.
Balaram Prasain, Tribhuvan University, Nepal.
Bareera Sadia, Al-Khawarizmi Institute of Computer Science, UET, Lahore, Pakistan.
Brinda Gurusamy, Cisco, USA.
Buddhika Karunarathne, University of Moratuwa, Sri Lanka.
Eugene Y A Charles, University of Jaffna, Sri Lanka.
Farah Adeeba, University of Engineering and Technology, KSK, Pakistan.
Farhan Jafri, Jamia Millia Islamia, India.
Gihan Dias, University of Moratuwa, Sri Lanka.
H N D Thilini, University of Colombo School of Computing, Sri Lanka.
Hariram Veeramani, UCLA, USA.
Hassan Sajjad, Dalhousie University, Canada.
Jayeeta Putatunda, Fitch Ratings, USA.
Kengatharaiyer Sarveswaran, University of Jaffna, Sri Lanka.
Krishna Chalise, Tribhuvan University, Nepal.
Kritesh Rauniyar, IIMS College, Nepal.
Lekhnath Pathak, Tribhuvan University, Nepal.
Lynnette Hui Xian Ng, CMU, USA.
Mahak Shah, Columbia University, USA.
Manjunath Chandrashekaraiah, Astera Labs, USA.
Menan Velayuthan, University of Moratuwa, Sri Lanka.
Munief Tahir, Al-Khawarizmi Institute of Computer Science, UET, Lahore, Pakistan.
Parameswari Krishnamurthy, IIIT Hyderabad, India.
Paritosh Katre, PayPal, USA.
Prakash Poudyal, Kathmandu University, Nepal.
Preetish Kakkar, Adobe, USA.
Qurat-ul-Ain Akram, University of Engineering and Technology, KSK, Pakistan.
Randil Pushpananda, University of Colombo School of Computing, Sri Lanka.
Sahar Rauf, Al-Khawarizmi Institute of Computer Science, UET, Lahore, Pakistan.
Sana Shams, Al-Khawarizmi Institute of Computer Science, UET, Lahore, Pakistan.

Shuvam Shiwakoti, Virginia Tech, USA.
Siddhant Bikram Shah, Northeastern University, USA.
Sinnathamby Mahesan, University of Jaffna, Sri Lanka.
Suganya Ramamoorthy, Vellore Institute of Technology University, India.
Surabhi Adhikari, Columbia University, USA.
Surangika Ranathunga, Massey University, New Zealand.
Surendrabikram Thapa, Virginia Tech, USA.
Tafseer Ahmed, Alexa Translations, Canada.
Toqeer Ehsan, Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates.
Usman Naseem, Macquarie University, Australia.
Uthayasanker Thayasivam, University of Moratuwa, Sri Lanka.
Vijayrajsinh Gohil, New York University, USA.

Volunteers (alphabetical order)

Ahrane Mahaganapathy, Unversity of Jaffna, Sri Lanka.
Menan Velayuthan, Unversity of Moratuwa, Sri Lanka.
Suthakar Sivashanth, Unversity of Jaffna, Sri Lanka.

Table of Contents

CHiPSAL 2025 Program

Sunday, January 19, 2025

(GMT+4)

8:30–10:30 Morning Oral Presentations

8:30–8:50 *A Brief Overview of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*
Kengatharaiyer Sarveswaran, Surendrabikram Thapa, Sana Shams, Ashwini Vaidya and Bal Krishna Bal

8:50–9:10 *Development of Pre-Trained Transformer-based Models for the Nepali Language*
Prajwal Thapa, Jinu Nyachhyon, Mridul Sharma and Bal Krishna Bal

9:10–9:30 *Benchmarking the Performance of Pre-trained LLMs across Urdu NLP Tasks*
Munief Hassan Tahir, Sana Shams, Layba Fiaz, Farah Adeeba and Sarmad Hussain

9:30–9:50 *Bengali ChartSumm: A Benchmark Dataset and study on feasibility of Large Language Models on Bengali Chart to Text Summarization*
Nahida Akter Tanjila, Afrin Sultana Poushi, Sazid Abdullah Farhan, Abu Raihan Mostofa Kamal, Md. Azam Hossain and Md. Hamjajul Ashmafee

9:50–10:10 *DweshVaani: An LLM for Detecting Religious Hate Speech in Code-Mixed Hindi-English*
Varad Srivastava

10:10–10:30 *Improving Accuracy of Low-resource ASR using Rule-Based Character Constituency Loss (RBCCL)*
Rupak Raj Ghimire, Prakash Poudyal and Bal Krishna Bal

10:30–12:00 Break

Sunday, January 19, 2025 (continued)

12:00–14:50 Afternoon Oral Presentations

12:00–12:20 *Natural Language Understanding of Devanagari Script Languages: Language Identification, Hate Speech and its Target Detection*
Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Surabhi Adhikari, Kengatharaiyer Sarveswaran, Bal Krishna Bal, Hariram Veeramani and Usman Naseem

12:20–12:40 *SiTa - Sinhala and Tamil Speaker Diarization Dataset in the Wild*
Uthayasanker Thayasivam, Thulasithan Gnanenthiram, Shamila Jeewantha and Upeksha Jayawickrama

12:40–13:00 *Sandhi Splitting in Tamil and Telugu: A Sequence-to-Sequence Approach Leveraging Transformer Models*
Priyanka Dasari, Mupparapu Sohan Gupta, Nagaraju Vuppala, Pruthwik Mishra and Parameswari Krishnamurthy

13:00–13:20 *Bridge the GAP: Multi-lingual Models For Ambiguous Pronominal Coreference Resolution in South Asian Languages*
Rahothvarman P, Adith John Rajeev, Kaveri Anuranjana and Radhika Mamidi

13:20–13:50 Poster spotlights

14:00–15:00 Poster Presentations

Poster papers

A Dual Contrastive Learning Framework for Enhanced Hate Speech Detection in Low-Resource Languages
Krishan Chavinda and Uthayasanker Thayasivam

Abstractive Summarization of Low resourced Nepali language using Multilingual Transformers
Prakash Dhakal and Daya Sagar Baral

Structured Information Extraction from Nepali Scanned Documents using Layout Transformer and LLMs
Aayush Neupane, Aayush Lamichhane, Ankit Paudel and Aman Shakya

Domain-adaptative Continual Learning for Low-resource Tasks: Evaluation on Nepali
Sharad Duwal, Suraj Prasai and Suresh Manandhar

Sunday, January 19, 2025 (continued)

POS-Aware Neural Approaches for Word Alignment in Dravidian Languages
Antony Alexander James and Parameswari Krishnamurthy

neDIOM: Dataset and Analysis of Nepali Idioms
Rhitabrat Pokharel and Ameeta Agrawal

Bridging the Bandwidth Gap: A Mixed Band Telephonic Urdu ASR Approach with Domain Adaptation for Banking Applications
Ayesha Khalid, Farah Adeeba, Najm Ul Sehar and Sarmad Hussain

Impacts of Vocoder Selection on Tacotron-based Nepali Text-To-Speech Synthesis
Ganesh Dhakal Chhetri and Prakash Poudyal

EmoTa: A Tamil Emotional Speech Dataset
Jubeerathan Thevakumar, Luxshan Thavarasa, Thanikan Sivatheepan, Sajeev Kugarajah and Uthayasanker Thayasivam

Benchmarking Whisper for Low-Resource Speech Recognition: An N-Shot Evaluation on Pashto, Punjabi, and Urdu
Najm Ul Sehar, Ayesha Khalid, Farah Adeeba and Sarmad Hussain

Leveraging Machine-Generated Data for Joint Intent Detection and Slot Filling in Bangla: A Resource-Efficient Approach
A H M Rezaul Karim and Ozlem Uzuner

Challenges in Adapting Multilingual LLMs to Low-Resource Languages using LoRA PEFT Tuning
Omkar Khade, Shruti Jagdale, Abhishek Phaltankar, Gauri Takalikar and Raviraj Joshi