COLING 2025

South East Asian Language Processing

Proceedings of the Second Workshop

January 20, 2025

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 979-8-89176-219-0

Preface

This volume contains the proceedings of the Second Workshop in South East Asian Language Processing, held in conjunction with the 31st International Conference on Computational Linguistics (COLING 2025).

South East Asia (SEA) remains one of the most linguistically diverse regions in the world, with over 1,200 languages spoken by 680 million people. However, the diversity of South East Asian languages continues to face challenges due to the historical emphasis on national languages as lingua franca after the end of colonization, and the increasing dominance of English driven by globalization.

This year marks the second iteration of our workshop, following the success of the inaugural event in 2023. Our aim is to provide a platform for practitioners from academia, government, and industry to come together and advance the research and development of language technologies for SEA languages. The workshop also aspires to foster an inclusive and collaborative community passionate about SEA languages, increase awareness of existing works, and catalyze partnerships to bolster NLP research and development in this linguistically rich region.

The workshop received 20 submissions of technical papers (an increase of 42

The accepted papers span a diverse range of topics and languages, reflecting the vibrancy of NLP research in SEA and beyond. These include research on languages in the Philippines, Indonesia, and Thailand. The papers address a variety of NLP tasks, including morphology, script transliteration, speech transcription, question answering, dialogue summarization and generation, multilingual and multicultural language models, as well as the curation of ethical and unethical instructions in Indonesian for LLMs and a Thai commonsense reasoning dataset.

We are encouraged by the growing interest in this field and look forward to the continued evolution of this workshop as a hub for innovative and impactful research on SEA languages. We hope that future editions will attract an even broader spectrum of submissions and foster greater collaboration among researchers and practitioners dedicated to these languages.

We look forward to an enriching discussion on research in South East Asian language processing at the online event on January 20, 2025.

January 2025

Derry Wijaya, Alham Fikri Aji, Clara Vania, Genta Indra Winata, Ayu Purwarianti

Organizing Committee

Derry Wijaya, Monash University Indonesia

Alham Fikri Aji, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

Clara Vania, Amazon

Genta Indra Winata, Capital One AI Foundation

Ayu Purwarianti, Bandung Institute of Technology (ITB)

Program Committee

Peerat Limkonchotiwat, AI Singapore

Garry Kuwanto, Boston University

Samuel Cahyawijaya, Cohere

David Moeljadi, Kanda University of International Studies

Zilu Tang, Boston University

Holy Lovenia, AI Singapore

Charibeth Cheng, De La Salle University Philippines

Kemal Kurniawan, University of Melbourne

Fajri Koto, MBZUAI

Dan John Velasco, Samsung Research Philippines

Adila Krisnadhi, Universitas Indonesia

Lintang Sutawika, Carnegie Mellon University

Table of Contents

Conference Program

bAI-bAI: A Context-Aware Transliteration System for Baybayin Scripts
Jacob Simon D. Bernardo and Maria Regina Justina E. Estuar

NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural
Wilson Wongso, David Samuel Setiawan, Steven Limcorn and Ananto Joyoadikusumo

Evaluating Sampling Strategies for Similarity-Based Short Answer Scoring: a Case Study in Thailand
Pachara Boonsarngsuk, Pacharapon Arpanantikul, Supakorn Hiranwipas, Wipu Watcharakajorn and Ekapol Chuangsuwanich

Thai Winograd Schemas: A Benchmark for Thai Commonsense Reasoning
phakphum artkaew

Anak Baik: A Low-Cost Approach to Curate Indonesian Ethical and Unethical Instructions
Sulthan Abiyyu Hakim, Rizal Setya Perdana and Tirana Noor Fatyanosa

Indonesian Speech Content De-Identification in Low Resource Transcripts
Rifqi Naufal Abdjul, Dessi Puji Lestari, Ayu Purwarianti, Candy Olivia Mawalim, Sakriani Sakti and Masashi Unoki

IndoMorph: a Morphology Engine for Indonesian
Ian Kamajaya and David Moeljadi

NusaDialogue: Dialogue Summarization and Generation for Underrepresented and Extremely Low-Resource Languages
Ayu Purwarianti, Dea Adhista, Agung Baptiso, Miftahul Mahfuzh, Yusrina Sabila, Aulia Adila, Samuel Cahyawijaya and Alham Fikri Aji