IJCNLP-AACL 2023

The First Workshop on South East Asian Language Processing

Proceedings of the Workshop

November 1, 2023

The IJCNLP-AACL organizers gratefully acknowledge the support from the following sponsors.

Platinum

vivo

Gold

面壁智能
ModelBest

Silver

translated.

ZHIPU · AI

小牛翻译
NiuTrans.com

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 979-8-89176-023-3

Preface

This volume contains the proceedings of the First Workshop in South East Asian Language Processing, held in conjunction with the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023).

South East Asia (SEA) is 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 has long been at risk due to the emphasis on national languages as the lingua franca in South East Asian countries at the end of colonization, and the increasing prominence of English due to the necessities of globalization.

This year is the first year we are conducting this workshop with the aim of bringing together practitioners from academia, government, and industry interested in the research and development of language technologies for SEA languages. The workshop also aims to build an inclusive community of everyone passionate about SEA languages, increase community awareness of works that have been developed to date on these languages, and foster collaborations that will strengthen and spur NLP research and development in SEA languages.

The workshop received 14 submissions of technical papers (12 long, 2 short), of which 10 were accepted (8 long, 2 short), for an acceptance rate of 71%. We thank the Programme Committee members who provided extremely valuable reviews in terms of technical content.

The accepted papers cover a wide range of natural language processing research, including research on languages in the Philippines and Indonesia, as well as languages in South Asia such as Bangla, Hindi, and Indic languages. The papers tackle a variety of tasks in NLP, including named entity recognition, word sense and synset induction, grammatical error correction, machine translation, dialogue systems, sentiment analysis, text generation, large language models alignment, as well as datasets, benchmarks, and language resources construction. In the future, we hope that this workshop will continue to attract submissions of diverse research works on SEA languages.

We look forward to an enriching discussion on research in South East Asian language processing at the hybrid event on November 1, 2023!

November 2023

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, Bloomberg

Ayu Purwarianti, Bandung Institute of Technology (ITB)

Program Committee

Peerat Limkonchotiwat, Vidyasirimedhi Institute of Science and Technology (VISTEC)

Jan Christian Blaise Cruz, Samsung R&D Institute Philippines

Garry Kuwanto, Boston University

Phu Mon Htut, Amazon Web Services AI Labs (AWS AI Labs)

Samuel Cahyawijaya, Hong Kong University of Science and Technology (HKUST)

David Moeljadi, Kanda University of International Studies

Haryo Akbarianto Wibowo, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

Radityo Eko Prasojo, Pitik.id

Andros Tjandra, Facebook AI Research (FAIR, Meta)

Rahmad Mahendra, RMIT University

Zilu Tang, Boston University

Holy Lovenia, AI Singapore, National University of Singapore

Samuel Louvan, Amazon

Table of Contents

Conference Program

14:00–14:05 ***Welcome Remarks***

14:05–14:20 *Towards Automatic Construction of Filipino WordNet: Word Sense Induction and Synset Induction Using Sentence Embeddings*
Dan John Velasco, Axel Alba, Trisha Gail Pelagio, Bryce Anthony Ramirez, Jan Christian Blaise Cruz, Unisse Chua, Briane Paul Samson and Charibeth Cheng

14:20–14:35 *Low-Resource Clickbait Spoiling for Indonesian via Question Answering*
Ni Putu Intan Maharani, Ayu Purwarianti and Alham Fikri Aji

14:35–14:50 *Developing a Named Entity Recognition Dataset for Tagalog*
Lester James Miranda

14:50–15:05 *Balarila: Deep Learning for Semantic Grammar Error Correction in Low-Resource Settings*
Andre Dominic H. Ponce, Joshue Salvador A. Jadie, Paolo Edni Andryn Espiritu and Charibeth Cheng

15:05–15:20 *Unsupervised Approach to Evaluate Sentence-Level Fluency: Do We Really Need Reference?*
Gopichand Kanumolu, Lokesh Madasu, Pavan Baswani, Ananya Mukherjee and Manish Shrivastava

15:20–15:50 ***Coffee Break***

15:50–16:05 *Utilizing Weak Supervision to Generate Indonesian Conservation Datasets*
Mega Fransiska, Diah Pitaloka, Saripudin Saripudin, Satrio Putra and Lintang Sutawika*

16:05–16:20 *InstructAlign: High-and-Low Resource Language Alignment via Continual Crosslingual Instruction Tuning*
Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung and Pascale Fung

16:20–16:35 *SentMix-3L: A Novel Code-Mixed Test Dataset in Bangla-English-Hindi for Sentiment Analysis*
Md Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos and Marcos Zampieri

16:35–16:50 *IndoToD: A Multi-Domain Indonesian Benchmark For End-to-End Task-Oriented Dialogue Systems*
Muhammad Kautsar, Rahmah Nurdini, Samuel Cahyawijaya, Genta Winata and Ayu Purwarianti

16:50–17:05 *Replicable Benchmarking of Neural Machine Translation (NMT) on Low-Resource Local Languages in Indonesia*
Lucky Susanto, Ryandito Diandaru, Adila Krisnadhi, Ayu Purwarianti and Derry Tanti Wijaya

No Day Set (continued)

17:05–17:10 *Best Paper Award Announcement*

17:10–17:20 *Closing Remarks*