Current Status of NLP in South East Asia with Insights from Multilingualism and Language Diversity

Alham Fikri Aji[*α], Jessica Zosa Forde[*β], Alyssa Marie Loo[*β], Lintang Sutawika[*γ], Skyler Wang[*δ,ϵ], Genta Indra Winata[*η], Zheng-Xin Yong[*β], Ruochen Zhang[*β], A. Seza Doğruöz[*κ], Yin Lin Tan[*λ,π], Jan Christian Blaise Cruz[*ϕ]

[α]MBZUAI [β]Brown University [γ]EleutherAI [δ]UC Berkeley
[ϵ] Meta AI [η]Bloomberg [κ]Ghent University [λ]Stanford University
[π]National University of Singapore [ϕ]Samsung R&D Institute Philippines

1 Motivation & Objectives

South East Asia (SEA) is a region with immense cultural and linguistic diversity—a melting pot of cultures, religions, and languages and it has a linguistic diversity hosting over 1000 languages (Table 1). In addition, multilingualism (i.e., speaking more than one language or dialect) is widely practiced on a daily basis. Despite the variety of languages, there is relatively less research on natural language processing (NLP) of the languages and their users in the area compared to languages in other regions. Many low-resource languages in the region will face potential endangerment in the long run.[1]

A significant challenge facing SEA NLP is the scarcity of available datasets and benchmarks for the region's languages, many of which are low-resource, resulting in sub-optimal performance of models. Similar to the situation in India, Europe, and Africa, most language users in SEA are multilingual; code-switching is common (Doğruöz et al., 2021; Winata et al., 2022b; Yong et al., 2023b) and it should not be seen as a challenge but as the natural way of communication in these settings. Language technology may also not be accessible to certain groups of SEA researchers due to the constraints on computing resources, hardware, training, and funding.

This tutorial will present an overview of language issues in the SEA region, link multilingualism and computational sociolinguistics with historical and societal perspectives, and provide a summary of the existing datasets for computational linguistics research, NLP systems, and evaluation benchmarks. Our goal is to inform the AACL'23 audience about challenges and opportunities for NLP research in SEA, taking the linguistic diversity in the region and multilingualism among the users and communities into account, while providing an overview of current NLP research on the languages spoken in the area. By providing an overview, we will highlight the research gaps to be tackled in the future.

Country	**Population**	**# Languages**
Brunei	0.5M	16
Cambodia	17M	28
Indonesia	267M	711
Laos	7M	87
Malaysia	32M	131
Myanmar	54M	121
Philippines	109M	184
Singapore	6M	24
Thailand	70M	73
Timor-Leste	1M	22
Vietnam	96M	110

Table 1: Language and population statistics of SEA countries, according to Ethnologue (Eberhard et al., 2021).

2 Type of Tutorial

This is a three-hour long **introductory** tutorial. The number of NLP publications focusing on SEA languages have been steadily increasing in recent years, but there is no tutorial at international venues (e.g., ACL/EMNLP/NAACL/EACL/COLING) that has systematically reviewed this research.

3 Target Audience and Prerequisites

This tutorial targets both junior and senior researchers (including NLP practitioners and lin-

[*] Equal contribution.

[1] https://www.ethnologue.com/region/SEA

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Tutorial Abstracts), pages 8–13
November 1–4, 2023. ©2023 Association for Computational Linguistics

guists) who are broadly interested in multilingual NLP and want to gain a deeper understanding of multilingualism work in the area as well as NLP research for SEA languages. Since this AACL 2023 tutorial takes place in Bali, Indonesia, where the location matches our tutorial topic, we foresee a substantial number of researchers from the SEA region attending the tutorial. We assume the audience has no previous knowledge about linguistics, multilingualism, and/or NLP in SEA. We expect that most participants will be familiar with basic issues in modeling language and in standard methods for learning from data, but no specific knowledge will be assumed.

4 Outline of Tutorial Content

The tutorial will cover four parts over the course of three hours:

1. Introduction to SEA (15 minutes)
2. Linguistic Landscape and Multilingualism in the SEA region (75 minutes)
3. Resource Availability and Collection (60 minutes)
4. (Panel Discussion) Research Ecosystem (30 minutes)

4.1 Introduction to SEA

We will start our tutorial by introducing SEA through discussing its geography, history, and culture. The topics here will be set the groundwork for what multilingualism means for SEA region, including its linguistic diversity and available resources.

4.2 Linguistic Landscape and Multilingualism in the SEA region

We will provide an overview of the language situation in SEA. We will begin by exploring the diversity of languages and language families in the region. The perspective of sociolinguistics, historical, and societal will be presented (Doğruöz and Sitaram, 2022). The multilingual nature of SEA will be highlighted through a discussion of language variation both between-country. Lastly, we will also briefly touch on language policies that have shaped the linguistic landscape of the SEA countries.

4.3 Resource Availability and Collection

We will discuss the availability of datasets, including unlabeled raw text corpora and labeled task-specific datasets, as well as the challenges of collecting, maintaining, and annotating these datasets. Additionally, we will explore where these datasets can be accessed, how various SEA countries organize their NLP resources, and to what extent they accurately represent common linguistic phenomena in SEA countries, such as code-switching and multilingualism.

Moreover, we will provide an overview of pretrained language models for SEA languages. We will delve into how these models are linguistically motivated by the multilingual society and how the‘y are designed to address different linguistic challenges in various SEA countries.

Lastly, we will introduce existing benchmark tasks for evaluating the pretrained models and touch upon the current shortcomings of the available NLP resources.

4.4 Research Ecosystem

Finally, through a panel discussion, we will discuss the meta-environment situating NLP research in SEA, including potential research collaborations and opportunities across SEA regions (Rungta et al., 2022) and the uneven distribution of research resources. We will highlight some community-based initiatives working on SEA languages. We will also involve the audience when sharing our research experience, highlighting the challenges as well as the opportunities for developing NLP technologies for SEA languages.

5 Breadth of the Tutorial

The tutorial covers a diverse set of topics related to NLP for SEA languages including their linguistics, data, models, evaluation, open research questions, etc. The tutorial also discusses the state of NLP research in different SEA regions.

6 Diversity Considerations

Instructors Our team spans across various (i) backgrounds and connections to SEA (ii) types of institutions, including academia, start-ups, and established tech companies, (iii) geographic locations, and (iv) seniority levels, gender and sexual identities.

Scope We introduce datasets, tasks, and NLP models that cover many SEA languages (including their dialects) from different language families.

Audience We target audiences interested in SEA languages from both academia and industry. We also welcome researchers who have worked with various SEA languages to share their challenges and experience during the panel discussion to foster an inclusive environment.

7 Reading List

The recommended reading list is as follows.

7.1 General Background

The following papers give a high-level overview of available resources, linguistic characteristics, and language technologies related to our tutorial.

1. One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. (Aji et al., 2022)

2. A Survey of Code-switching: Linguistic and Social Perspectives for Language Technologies (Doğruöz et al., 2021)

3. Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights. (Doğruöz and Sitaram, 2022)

4. Survey on Thai NLP Language Resources and Tools. (Arreerard et al., 2022)

5. Areal Linguistics and Mainland SEA (Enfield, 2005)

6. No Language Left Behind: Scaling Human-Centered Machine Translation (NLLB Team et al., 2022)

7.2 Resource Collection and Availability

The following papers characterize efforts in curating labeled datasets and organizing benchmark evaluation for SEA languages as well as training language models.

7.2.1 Datasets and Evaluation

1. Crowdsourcing-based Annotation of Emotions in Filipino and English Tweets. (Lapitan et al., 2016)

2. SEAME: a Mandarin-English Code-switching Speech Corpus in SEA (Lyu et al., 2010)

3. NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. (Winata et al., 2023)

4. Cross-lingual Few-Shot Learning on Unseen Languages. (Winata et al., 2022a)

5. NusaCrowd: Open Source Initiative for Indonesian NLP Resources (Cahyawijaya et al., 2023)

6. BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset (Romadhona et al., 2022)

7.2.2 Pretrained Language Models

1. IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. (Koto et al., 2020)

2. PhoBERT: Pre-trained language models for Vietnamese. (Nguyen and Tuan Nguyen, 2020)

3. WangchanBERTa: Pretraining transformer-based Thai Language Models (Lowphansirikul et al., 2021)

4. Encoder-Decoder Language Model for Khmer Handwritten Text Recognition in Historical Documents (Born et al., 2022)

5. LaoPLM: Pre-trained Language Models for Lao. (Lin et al., 2022)

6. Tagalog RoBERTa: Improving Large Scale Language Models and Resources for Filipino. (Cruz and Cheng, 2022)

8 Sharing of Tutorial Materials

All of our tutorial materials will be publicly available at `https://aacl2023-sea-nlp.github.io`.

9 Ethics Statement

Preserving multilingualism and directing attention to low-resource languages is a pertinent collective mission in NLP research today. As prominent projects like "No Language Left Behind" (NLLB Team et al., 2022) argue, prioritizing high-resource languages comes with the cost of under-privileging research on low-resource languages. As a result of linguistic inequality, large language models contain cross-lingual vulnerability (Yong et al., 2023a), and low-resource languages could face endangerment, where useful resources that could be developed for these language speakers may never materialize. Furthermore, as many SEA NLP researchers speak high-resource languages, existing academic practices and cultures may compel them to focus on these languages rather than their low-resource counterparts. The uneven distribution of research resources, such as hardware, training, and funding, may also present hurdles for interested researchers to develop NLP systems for SEA languages. Ultimately, we recognize that many low-resource SEA languages remain understudied due to a confluence of reasons, and it is our hope that this tutorial could spark productive conversations around how we can bring sustained attention to this area of research.

10 Presenter Information

Alham Fikri Aji is an assistant professor in MBZUAI. His research focuses on multi-lingual and cross-lingual NLP, especially for under-resourced languages and communities. His work area also includes data construction as well as data-efficient systems, and compute-efficient models for better accessibility. His email is alham.fikri@mbzuai.ac.ae.

Jessica Zosa Forde is a PhD Candidate at Brown University. Jessica's research focuses on the evaluation of deep learning models, to improve their reliability in high stakes domains. Jessica presented a tutorial on reproduciblity in NLP at ACL in 2022. Her email is jessica_forde@brown.edu.

Alyssa Marie Loo is an undergraduate in Linguistics and Computer Science at Brown University. Her research focuses on interpretability of large language models and their alignment with human linguistic behavior. Her email is alyssa_loo@brown.edu.

Lintang Sutawika is Researcher at EleutherAI. He is a proponent of open source software and machine learning artifacts. His work has comprised of extending language models to more languages, interpreting language models and maintaining software for language model evaluation. His email is lintang@eleuther.ai

Samson Tan is an Applied Scientist at AWS AI Research & Education. His research focuses on linguistic variation and their effect on the robustness and evaluation of NLP models. His email is samson@amazon.com.

Skyler Wang is a PhD Candidate at UC Berkeley and a Visiting Sociologist at Meta AI. Broadly, Skyler's research focuses on creating socially and ethically-grounded machine translation technologies for low-resource language communities. He is a Sociologist on Meta AI's "No Language Left Behind" team. His email is skyler.wang@berkeley.edu.

Genta Indra Winata is a Senior Research Scientist at Bloomberg. His research focuses on multilingual, cross-lingual, language models, dialogue system, and low-resource NLP. His work area includes few-shot learning and evaluation of large language models. His email is gwinata@bloomberg.net.

Zheng-Xin Yong is a PhD Candidate at Brown University. His research focuses on cross-lingual NLP, large language models, and AI safety. His email is contact.yong@brown.edu

Ruochen Zhang is a PhD Candidate at Brown University. Her research interests lie in multi-/cross-lingual learning, evaluation and application of large language models. Her email is ruochen_zhang@brown.edu.

A. Seza Doğruöz is a tenured Associate Professor at Ghent University. She conducts interdisciplinary research on multilingualism, sociolinguistics and computational linguistics. Her email is as.dogruoz@ugent.be.

Yin Lin Tan is a PhD student at Stanford University. Her research focuses on sociolinguistic variation, phonetics, and multilingualism. Her email is yltan@stanford.edu.

Jan Christian Blaise Cruz is an AI Research Engineer at Samsung R&D Institute Philippines. His research revolves around low-resource techniques for translation and language generation. His email is jcb.cruz@samsung.com.

References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Ratchakrit Arreerard, Stephen Mander, and Scott Piao. 2022. Survey on Thai NLP language resources and tools. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6495–6505, Marseille, France. European Language Resources Association.

Seanghort Born, Dona Valy, and Phutphalla Kong. 2022. Encoder-decoder language model for khmer handwritten text recognition in historical documents. In *2022 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, pages 234–238.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Jan Christian Blaise Cruz and Charibeth Cheng. 2022. Improving large-scale language models and resources for Filipino. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6548–6555, Marseille, France. European Language Resources Association.

A. Seza Doğruöz and Sunayana Sitaram. 2022. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97, Marseille, France. European Language Resources Association.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World. Twenty-fourth edition.* Dallas, Texas: SIL International.

Nick J Enfield. 2005. Areal linguistics and mainland southeast asia. *Annual review of anthropology*, 34:181–206.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fermin Roberto Lapitan, Riza Theresa Batista-Navarro, and Eliezer Albacea. 2016. Crowdsourcing-based annotation of emotions in Filipino and English tweets. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 74–82, Osaka, Japan. The COLING 2016 Organizing Committee.

Nankai Lin, Yingwen Fu, Chuwei Chen, Ziyu Yang, and Shengyi Jiang. 2022. LaoPLM: Pre-trained language models for Lao. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6506–6512, Marseille, France. European Language Resources Association.

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. 2021. Wangchanberta: Pretraining transformer-based thai language models. *arXiv preprint arXiv:2101.09635*.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li. 2010. Seame: a mandarin-english code-switching speech corpus in south-east asia. In *Eleventh Annual Conference of the International Speech Communication Association*.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arxiv preprint arxiv:2207.04672*.

Nanda Putri Romadhona, Sin-En Lu, Bo-Han Lu, and Richard Tzong-Han Tsai. 2022. BRCC and SentiBahasaRojak: The first Bahasa rojak corpus for pretraining and sentiment analysis dataset. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4418–4428, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mukund Rungta, Janvijay Singh, Saif M Mohammad, and Diyi Yang. 2022. Geographic citation gaps in nlp research. *arXiv preprint arXiv:2210.14424*.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preoţiuc-Pietro. 2022a. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 777–791.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2022b. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *arXiv preprint arXiv:2212.09660*.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023a. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Samuel Cahyawijaya, Holy Lovenia, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Long Phan, Yin Lin Tan, and Alham Fikri Aji. 2023b. Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages.