LoResMT 2025

The Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)

Proceedings of the Workshop

May 3, 2025

The LoResMT organizers gratefully acknowledge the support from the following organizations.

In cooperation with

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 979-8-89176-230-5

Preface

Based on the success of past low-resource machine translation (MT) workshops at AMTA 2018, MT Summit 2019, AACL-IJCNLP 2020, AMTA 2021, COLING 2022, EACL 2023 and ACL 2024, we introduce the LoResMT 2025 workshop at NAACL 2025 (https://2025.naacl.org/). In the past few years, machine translation (MT) performance has improved significantly. With the development of new techniques such as multilingual translation and transfer learning, the use of MT is no longer a privilege for users of popular languages. However, the goal of expanding MT coverage to more diverse languages is hindered by the fact that MT methods require large amounts of data to train quality systems. This has made developing MT systems for low-resource languages challenging. Therefore, the need for developing comparable MT systems with relatively small datasets remains highly desirable.

Despite the advancements in MT technologies, creating an MT system for a new language or enhancing an existing one still requires a significant amount of effort to gather the necessary resources. The data-intensive nature of neural machine translation (NMT) approaches necessitates parallel and monolingual corpora in various domains, which are always in high demand. Developing MT systems also requires dependable evaluation benchmarks and test sets. Furthermore, MT systems rely on numerous natural language processing (NLP) tools to preprocess human-generated texts into the required input format and post-process MT output into the appropriate textual forms in the target language. These tools include word tokenizers/de-tokenizers, word segmenters, and morphological analyzers, among others. The quality of these tools significantly impacts the translation output, yet there is a limited discourse on their methods, their role in training different MT systems, and their support coverage in different languages.

LoResMT is a platform that aims to facilitate discussions among researchers who are working on machine translation (MT) systems and methods for low-resource, under-represented, ethnic, and endangered languages. The goal of the platform is to address the challenges associated with the development of MT systems for languages that have limited resources or are at risk of being lost.

This year, LoResMT received research papers covering many languages spoken worldwide. The acceptance rate of LoResMT this year is 70.83%. Aside from the research papers, LoResMT also featured two invited talks. These talks allowed participants to hear from experts in the field of MT and learn about the latest developments and challenges in MT for low-resource languages.

The program committee members play a crucial role in ensuring the success of the peer-review workshop. They review the submissions and provide constructive feedback to help the authors refine their papers and ensure they meet the set standards. Without their dedication, expertise, and hard work, the workshop would not be possible. The authors who submitted their work to LoResMT are also an integral part of the workshop's success. Their research and contributions offer new insights into the field of machine translation for low-resource languages, and their participation enriches the discussions and fosters collaboration. We are sincerely grateful to both the program committee members and the authors for their invaluable contributions and for making LoResMT a success.

Kat, Nathaniel, Atul, Chao
(On behalf of the LoResMT chairs)

Organizing Committee

Workshop Chairs

Atul Kr. Ojha, Data Science Institute, Insight Research Ireland Centre for Data Analytics, University of Galway
Chao-hong Liu, Potamu Research Ltd
Ekaterina Vylomova, University of Melbourne, Australia
Flammie Pirinen, UiT Norgga árktalaš universitehta
Jonathan Washington, Swarthmore College
Nathaniel Oco, De La Salle University
Xiaobing Zhao, Minzu University of China

Program Committee

Abigail Walsh, ADAPT Centre, Dublin City University, Ireland
Alberto Poncelas, Rakuten, Singapore
Ali Hatami, University of Galway
Alina Karakanta, Fondazione Bruno Kessler (FBK), University of Trento
Anna Currey, AWS AI Labs
Aswarth Abhilash Dara, School of Computer Science, Carnegie Mellon University
Atul Kr. Ojha, University of Galway & Panlingua Language Processing LLP
Chao-hong Liu, Potamu Research Ltd
Constantine Lignos, Brandeis University, USA
Daan van Esch, Google
Ekaterina Vylomova, University of Melbourne, Australia
Flammie Pirinen, UiT Norgga árktalaš universitehta
John Philip McCrae, University of Galway
Luis Chiruzzo, Facultad de Ingeniería - Universidad de la República - Uruguay
Maitrey Mehta, University of Utah
Milind Agarwal, George Mason University
Nathaniel Oco, De La Salle University
Pavel Rychlý, Masaryk University and Lexical Computing
Pengwei Li, Meta
Santhosh Kakarla, George Mason University
Satya Subrahmanya Gautama Shastry Bulusu Venkata, George Mason University
Sourabrata Mukherjee, Charles University
Surangika Ranathunga, Massey University
Timothee Mickus, University of Helsinki
Valentin Malykh, Huawei Noah's Ark lab and Kazan Federal University
Wen Lai, Technical University of Munich
Yasmin Moslem, Bering Lab
Zeerak Talat, University of Edinburgh, University of Edinburgh

Secondary Reviewer

Gaurav Negi, University of Galway

Keynote Talk: Low-Resource NLP: hot takes and anecdotes from Google Translate

Isaac Caswell
Google Translate

Abstract: Come with me as I opine that Low-Resource NLP is still the best place to be, and then give lots of anecdotes about the things I've run into while adding hundreds of languages to Google Translate.

Bio: Isaac Caswell has been a Researcher at Google Translate since 2017, and has a background in Linguistics and Computer Science. He focuses on Low Resource Languages, and is responsible for the last 146 languages that were added to the product. His research focuses on staring at data and trying to push the boundaries of multilinguality, yielding curious paper titles like MadLad, Smol, Gatitos, and FUNLangID, along with normal-sounding papers with silly tables, like that one with all the LangID mistakes and the other one with the animals that keep turning into crocodiles. Outside of work he focuses on language learning, singing, nature, community living, and Cat.

Keynote Talk: Low-resource MT: A perspective from the Americas

Arturo Oncevay
JP Moragn

Abstract: Exploring the challenges and opportunities of MT for Indigenous languages in the Americas through lessons from organizing shared tasks at AmericasNLP.

Bio: TBD

Table of Contents

Comparative Evaluation of Machine Translation Models Using Human-Translated Social Media Posts as References: Human-Translated Datasets
Shareefa Ahmed Al Amer, Mark G. Lee and Phillip Smith 1

Enhanced Zero-Shot Machine Translation via Fixed Prefix Pair Bootstrapping
Van-Hien Tran and Masao Utiyama 10

UTER: Capturing the Human Touch in Evaluating Morphologically Rich and Low-Resource Languages
Samy Ouzerrout 16

From Text to Multi-Modal: Advancing Low-Resource-Language Translation through Synthetic Data Generation and Cross-Modal Alignments
Bushi Xiao, Qian Shen and Daisy Zhe Wang 24

Wenzhou Dialect Speech to Mandarin Text Conversion
Zhipeng Gao, Akihiro Tamura and Tsuneo Kato 36

Fostering Digital Inclusion for Low-Resource Nigerian Languages: A Case Study of Igbo and Nigerian Pidgin
Ebelechukwu Nwafor and Minh Phuc Nguyen 44

Low-resource Machine Translation: what for? who for? An observational study on a dedicated Tetun language translation service
Raphael Merx, Adérito José Guterres Correia, Hanna Suominen and Ekaterina Vylomova 54

Jamo-Level Subword Tokenization in Low-Resource Korean Machine Translation
Junyoung Lee, Marco Cognetta, Sangwhan Moon and Naoaki Okazaki 66

Beyond English: The Impact of Prompt Translation Strategies across Languages and Tasks in Multilingual LLMs
Itai Mondshine, Tzuf Paz-Argaman and Reut Tsarfaty 81

ModeLing: A Novel Dataset for Testing Linguistic Reasoning in Language Models
Nathan Andrew Chi, Teodor Malchev, Riley Kong, Ryan Andrew Chi, Lucas Huang, Ethan A Chi, R. Thomas McCoy and Dragomir Radev 105

Multilingual State Space Models for Structured Question Answering in Indic Languages
Arpita Vats, Rahul Raja, Mrinal Mathur, Aman Chadha and Vinija Jain 115

Parallel Corpora for Machine Translation in Low-Resource Indic Languages: A Comprehensive Review
Rahul Raja and Arpita Vats 129

Low-Resource Transliteration for Roman-Urdu and Urdu Using Transformer-Based Models
Umer Butt, Stalin Varanasi and Günter Neumann 144

Building Data Infrastructure for Low-Resource Languages
Sarah K. K. Luger, Rafael Mosquera and Pedro Ortiz Suarez 154

Encoder-Aware Sequence-Level Knowledge Distillation for Low-Resource Neural Machine Translation
Menan Velayuthan, Nisansa De Silva and Surangika Ranathunga 161

PahGen: Generating Ancient Pahlavi Text via Grammar-guided Zero-shot Translation
Farhan Farsi, Parnian Fazel, Farzaneh Goshtasb, Nadia Hajipour, Sadra Sabouri, Ehsaneddin Asgari and Hossein Sameti 171

Limitations of Religious Data and the Importance of the Target Domain: Towards Machine Translation for Guinea-Bissau Creole
Jacqueline Rowe, Edward Gow-Smith and Mark Hepple 183

Program

Saturday, May 3, 2025

09:00 - 09:10 *Opening Remarks*

09:10 - 10:10 *Invited Talk 1: Isaac Caswell (Google Translate)*

10:10 - 10:30 *Session 1: Booster Presentations*

PahGen: Generating Ancient Pahlavi Text via Grammar-guided Zero-shot Translation
Farhan Farsi, Parnian Fazel, Farzaneh Goshtasb, Nadia Hajipour, Sadra Sabouri, Ehsaneddin Asgari and Hossein Sameti

Multilingual State Space Models for Structured Question Answering in Indic Languages
Arpita Vats, Rahul Raja, Mrinal Mathur, Aman Chadha and Vinija Jain

Parallel Corpora for Machine Translation in Low-Resource Indic Languages: A Comprehensive Review
Rahul Raja and Arpita Vats

Comparative Evaluation of Machine Translation Models Using Human-Translated Social Media Posts as References: Human-Translated Datasets
Shareefa Ahmed Al Amer, Mark G. Lee and Phillip Smith

Fostering Digital Inclusion for Low-Resource Nigerian Languages: A Case Study of Igbo and Nigerian Pidgin
Ebelechukwu Nwafor and Minh Phuc Nguyen

Beyond English: The Impact of Prompt Translation Strategies across Languages and Tasks in Multilingual LLMs
Itai Mondshine, Tzuf Paz-Argaman and Reut Tsarfaty

Building Data Infrastructure for Low-Resource Languages
Sarah K. K. Luger, Rafael Mosquera and Pedro Ortiz Suarez

From Text to Multi-Modal: Advancing Low-Resource-Language Translation through Synthetic Data Generation and Cross-Modal Alignments
Bushi Xiao, Qian Shen and Daisy Zhe Wang

10:30 - 11:00 *Coffee/Tea Break*

Saturday, May 3, 2025 (continued)

11:00 - 12:30 *Session 2: Scientific Research Papers*

Low-resource Machine Translation: what for? who for? An observational study on a dedicated Tetun language translation service
Raphael Merx, Adérito José Guterres Correia, Hanna Suominen and Ekaterina Vylomova

Enhanced Zero-Shot Machine Translation via Fixed Prefix Pair Bootstrapping
Van-Hien Tran and Masao Utiyama

Limitations of Religious Data and the Importance of the Target Domain: Towards Machine Translation for Guinea-Bissau Creole
Jacqueline Rowe, Edward Gow-Smith and Mark Hepple

Jamo-Level Subword Tokenization in Low-Resource Korean Machine Translation
Junyoung Lee, Marco Cognetta, Sangwhan Moon and Naoaki Okazaki

Wenzhou Dialect Speech to Mandarin Text Conversion
Zhipeng Gao, Akihiro Tamura and Tsuneo Kato

12:30 - 14:00 *Lunch*

14:00 - 15:00 *Invited Talk 2: Arturo Oncevay (JP Moragn)*

15:00 - 16:00 *Session 3: Poster Session*

PahGen: Generating Ancient Pahlavi Text via Grammar-guided Zero-shot Translation
Farhan Farsi, Parnian Fazel, Farzaneh Goshtasb, Nadia Hajipour, Sadra Sabouri, Ehsaneddin Asgari and Hossein Sameti

Multilingual State Space Models for Structured Question Answering in Indic Languages
Arpita Vats, Rahul Raja, Mrinal Mathur, Aman Chadha and Vinija Jain

Parallel Corpora for Machine Translation in Low-Resource Indic Languages: A Comprehensive Review
Rahul Raja and Arpita Vats

Saturday, May 3, 2025 (continued)

Comparative Evaluation of Machine Translation Models Using Human-Translated Social Media Posts as References: Human-Translated Datasets
Shareefa Ahmed Al Amer, Mark G. Lee and Phillip Smith

Fostering Digital Inclusion for Low-Resource Nigerian Languages: A Case Study of Igbo and Nigerian Pidgin
Ebelechukwu Nwafor and Minh Phuc Nguyen

Beyond English: The Impact of Prompt Translation Strategies across Languages and Tasks in Multilingual LLMs
Itai Mondshine, Tzuf Paz-Argaman and Reut Tsarfaty

Building Data Infrastructure for Low-Resource Languages
Sarah K. K. Luger, Rafael Mosquera and Pedro Ortiz Suarez

From Text to Multi-Modal: Advancing Low-Resource-Language Translation through Synthetic Data Generation and Cross-Modal Alignments
Bushi Xiao, Qian Shen and Daisy Zhe Wang

15:30 - 16:00 *Coffee/Tea Break*

16:00 - 17:12 *Session 4: Scientific Research Papers*

ModeLing: A Novel Dataset for Testing Linguistic Reasoning in Language Models
Nathan Andrew Chi, Teodor Malchev, Riley Kong, Ryan Andrew Chi, Lucas Huang, Ethan A Chi, R. Thomas McCoy and Dragomir Radev

Low-Resource Transliteration for Roman-Urdu and Urdu Using Transformer-Based Models
Umer Butt, Stalin Varanasi and Günter Neumann

Encoder-Aware Sequence-Level Knowledge Distillation for Low-Resource Neural Machine Translation
Menan Velayuthan, Nisansa De Silva and Surangika Ranathunga

UTER: Capturing the Human Touch in Evaluating Morphologically Rich and Low-Resource Languages
Samy Ouzerrout

17:13 - 17:23 *Closing remarks*