LoResMT 2026

The Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)

Proceedings of the Workshop

March 28, 2026

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 979-8-89176-366-1

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, ACL 2024 and NAACL 2025, we introduce the LoResMT 2026 workshop at EACL 2026 (https://2026.eacl.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 57 papers covering many languages spoken worldwide. We have archived 15 scientific research papers from direct submission, 1 scientific research paper from ARR commitment and 9 system descriptions. Aside from the research and shared task 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.

Atul, Chao, Kat, Nathaniel
(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
Aishwarya Jadhav, University of California, San Diego
Alberto Poncelas, Rakuten, Singapore
Ali Hatami, University of Galway
Alina Karakanta, Leiden University
Aswarth Abhilash Dara, Apple
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
Gaurav Negi, University of Galway
John Philip McCrae, University of Galway
Koel Dutta Chowdhury, Universität des Saarlandes
Manoj Yadav, Amazon
Mathias Müller, University of Zurich
Majid Latifi, University of York
Nathaniel Oco, De La Salle University
Pengwei Li, Meta
Rico Sennrich, University of Zurich
Sardana Ivanova, University of Helsinki
Sourabrata Mukherjee, Microsoft Research India
Surangika Ranathunga, Massey University
Valentin Malykh, International IT University
Yasmin Moslem, Trinity College Dublin

Keynote Talk: How (Not) to Find Errors in LLM Outputs

Ondřej Dušek
Institute of Formal and Applied linguistics, Charles University, Prague (Czech Republic)

Abstract: While LLMs have substantially improved the quality of generated texts, they still tend to make errors in their outputs, which can be subtle and harder to find than for older approaches. This needs to be reflected in the evaluation, where standard metrics or simple scores may not capture errors easily. While human evaluation should produce better results, we find a lot of inconsistency and underspecification in practice.

Building on previous works in machine translation, we examine annotating individual spans of texts for errors in order to get more detailed evaluation feedback. We explore span annotation through both human evaluation and LLM-as-judge evaluation. We provide a unified interface for both LLM and human authored error annotations, we examine different methods of obtaining LLM-annotated spans and introduce LLM ensembles for higher robustness. We directly compare LLMs and humans on the same task, finding that LLMs are able to reach high correlation with human assessments and, depending on the domain, can match trained human crowd workers in performance. However, we also report many caveats on both the human and LLM side, and we discuss potential further improvements of the evaluation setup.

Bio: Ondřej Dušek is an Assistant Professor at Charles University in Prague, working on natural language generation and human-computer dialogue. His research focuses on generative language models including large language models, mostly applied to the data-to-text and dialogue response generation tasks. He is specifically interested in evaluating the quality of generated content, especially its semantic accuracy.

After obtaining his PhD in Prague, Ondřej spent 2 years as a postdoc at Heriot-Watt University in Edinburgh. Back in Prague, he is currently the PI of an ERC Starting Grant which aims to produce fluent, accurate and explainable natural language generation systems.

Keynote Talk: Building Tiny Aya: Optimizing for Multilinguality in Small LLMs

Julia Kreutzer
Cohere for Labs

Abstract: Do massively multilingual models need to be massive in size? In order to make AI more accessible, we need to rethink the development of multilingual models from the ground up. In this talk, we will explore how Tiny Aya was built, a massively multilingual model with 3.5B parameters. We will dive into the innovation that led to its success especially for more underrepresented languages, and we will discuss which challenges remain.

Bio: Julia Kreutzer is a Senior Research Scientist at Cohere Labs, where she focuses on research around multilingual large language models. She has a background in machine translation, with a PhD from Heidelberg University and prior work experience at Google Translate. She's passionate about advancing NLP technologies for underrepresented languages and has been part of multiple open science initiatives to work towards this goal collaboratively.

Table of Contents

Program

Saturday, March 28, 2026

09:00 - 09:10 *Opening Remarks*

09:10 - 10:10 *Invited Talk 1: Ondřej Dušek (Institute of Formal and Applied Linguistics, Charles University, Prague (Czech Republic))*

10:10 - 10:30 *Session 1: Findings of Turkic Low Resource Machine Translation Challenge*

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

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

Are Small Language Models the Silver Bullet to Low-Resource Languages Machine Translation?
Yewei Song, Lujun LI, Cedric Lothritz, Saad Ezzini, Lama Sleem, Niccolo' Gentile, Radu State, Tegawendé F. Bissyandé and Jacques Klein

Tao–Filipino Neural Machine Translation: Strategies for Ultra–Low-Resource Settings
Adrian Denzel Macayan, Luis Andrew Sunga Madridijo, Ellexandrei Esponilla and Zachary Mitchell Francisco

Building and Evaluating a High Quality Parallel Corpus for English Urdu Low Resource Machine Translation
Munief Hassan Tahir, Hunain Azam, Sana Shams and Sarmad Hussain

Text Filter Based on Automatically Acquired Vocabularies for Multilingual Machine Translation
Kenji Imamura and Masao Utiyama

Improving Indigenous Language Machine Translation with Synthetic Data and Language-Specific Preprocessing
Aashish Dhawan, Christopher Driggers-Ellis, Christan Grant and Daisy Zhe Wang

Adapting Multilingual NMT to Language Isolates: The Role of Proxy Language Selection and Dialect Handling for Nivkh
Eleonora Izmailova, Alexey Sorokin and Pavel Grashchenkov

12:30 - 14:00 *Lunch*

14:00 - 15:00 *Invited Talk 2: Julia Kreutzer (Cohere for Labs)*

Saturday, March 28, 2026 (continued)

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

Can LLMs Translate Italy's Language Varieties?
Edoardo Signoroni and Pavel Rychlý

Balancing Fluency and Adherence: Hybrid Fallback Term Injection in Low-Resource Terminology Translation
Kurt Abela, Marc Tanti and Claudia Borg

Assessing and Improving Punctuation Robustness in English-Marathi Machine Translation
Kaustubh Shivshankar Shejole, Sourabh Deoghare and Pushpak Bhattacharyya

Can Linguistically Related Languages Guide LLM Translation in Low-Resource Settings?
Aishwarya Ramasethu, Rohin Garg, Niyathi Allu, Harshwardhan Fartale and Dun Li Chan

Navigating Data Scarcity in Low-Resource English-Tatar Translation using LLM Fine-Tuning
Ahmed Khaled Khamis

No One-Size-Fits-All: Building Systems For Translation to Bashkir, Kazakh, Kyrgyz, Tatar and Chuvash Using Synthetic And Original Data
Dmitry Karpov

DevLake at LoResMT 2026: The Impact of Pre-training and Model Scale on Russian-Bashkir Low-Resource Translation
Vyacheslav Tyurin

A Comparative Evaluation of Open-Source Models for Russian-Kazakh Translation
Gleb Shanshin

Script Correction and Synthetic Pivoting: Adapting Tencent HY-MT for Low-Resource Turkic Translation
Bolgov Maxim

Machine Translation for Low Resource Turkic Languages: English-Tatar
Alexander Dikov

Saturday, March 28, 2026 (continued)

Data-Centric Approach at the LoResMT 2026 Turkic Translation Challenge: Russian-Kyrgyz
Dmitry Novokshanov

LoResMT 2026 Shared Task System Description
Vladimir Panov

Ensemble Methods for Low-Resource Russian-Kyrgyz Machine Translation: When Diverse Models Beat Better Models
Adilet Metinov

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

16:00 - 17:15 *Session 4: Scientific Research Papers*

Comparing LLM-Based Translation Approaches for Extremely Low-Resource Languages
Jared Coleman, Ruben Rosales, Kira Toal, Diego Cuadros, Nicholas Leeds, Bhaskar Krishnamachari and Khalil Iskarous

Context Volume Drives Performance: Tackling Domain Shift in Extremely Low-Resource Translation via RAG
David Samuel Setiawan, Raphael Merx and Jey Han Lau

Semi-Automatic construction of a Quechua-Spanish dictionary
Maximiliano Duran and Max Silberztein

A Fine-Grained Linguistic Evaluation of Low-Resource Luxembourgish–English MT
Nils Rehlinger

CTC Regularization for Low-Resource Speech-to-Text Translation
Zachary William Hopton and Rico Sennrich

17:15 - 17:25 *Closing remarks*