LoResMT 2024

The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)

Proceedings of the Workshop

August 15, 2024

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 979-8-89176-149-0

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, we introduce seventh LoResMT workshop at ACL 2024 (https://2024.aclweb.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 require dependable evaluation benchmarks and test sets. Furthermore, MT systems rely on numerous natural language processing (NLP) tools to pre-process 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 workshop received many papers on large language model (LLM) methods for MT. The acceptance rate of LoResMT this year is 51.28%. 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 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, Valentin, Nathaniel, Atul, Chao
(On behalf of the LoResMT chairs)

Organizing Committee

Workshop Chairs

Atul Kr. Ojha, Atul Kr. Ojha, Data Science Institute, Insight SFI Research 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
Jade Abbott, Retro Rabbit
Jonathan Washington, Swarthmore College
Nathaniel Oco, De La Salle University
Valentin Malykh, Huawei Noah's Ark lab and Kazan Federal University
Varvara Logacheva Skolkovo, Institute of Science and Technology
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, Walmart Global Technology
Atul Kr. Ojha, University of Galway & Panlingua Language Processing LLP
Bogdan Babych, Heidelberg University
Chao-hong Liu, Potamu Research Ltd
Constantine Lignos, Brandeis University, USA
Daan van Esch, Google
Dana Moukheiber, Massachusetts Institute of Technology
Ekaterina Vylomova, University of Melbourne, Australia
Eleni Metheniti, CLLE-CNRS and IRIT-CNRS
Flammie Pirinen, UiT Norgga árktalaš universitehta
Jinliang Lu, Institute of automation, Chinese Academy of Sciences
John Philip McCrae, University of Galway
Jonathan Washington, Swarthmore College
Koel Dutta Chowdhury, Saarland University
Majid Latifi, UPC University
Maria Art Antonette Clariño, University of the Philippines Los Baños
Milind Agarwal, George Mason University
Nathaniel Oco, De La Salle University
Pavel Rychlý, Masaryk University and Lexical Computing
Pengwei Li, Meta
Rashid Ahmad, International Institute of Information Technology, Hyderabad
Santanu Pal, Wipro
Sangjee Dondrub, Qinghai Normal University
Sardana Ivanova, University of Helsinki
Sourabrata Mukherjee, Charles University
Thepchai Supnithi, National Electronics and Computer Technology Center
Timothee Mickus, University of Helsinki

Valentin Malykh, Huawei Noah's Ark lab and Kazan Federal University
Wen Lai, LMU Munich
Xuebo Liu, Harbin Institute of Technolgy, Shenzhen
Yalemisew Abgaz, Dublin City University
Yasmin Moslem, Bering Lab
Zhanibek Kozhirbayev, National Laboratory Astana, Nazarbayev University

Secondary Reviewers

Gaurav Negi, University of Galway

Keynote Talk: Hyperparameter Optimization for Low-Resource Machine Translation

Kevin Duh
Johns Hopkins University, USA

Abstract: Neural Machine Translation models are full of hyperparameters. To obtain a good model, one must carefully experiment with hyperparameters such as the number of layers, the number of hidden nodes, the type of non-linearity, the learning rate, and the drop-out parameter, just to name a few. I will discuss general hyperparameter optimization algorithms—including those based on evolutionary strategies, Bayesian techniques, and bandit learning–that can automate this laborious process. Further, I will argue that hyperparameter optimization is especially valuable for low-resource settings, where commonly-used hyperparameters are often suboptimal and small data sizes afford larger search spaces. Finally, I will discuss benchmarks and datasets for evaluating hyperparameter optimization algorithms in practice.

Bio: Kevin Duh is a senior research scientist at the Johns Hopkins University Human Language Technology Center of Excellence (JHU HLTCOE). He is also an assistant research professor in the Department of Computer Science and a member of the Center for Language and Speech Processing (CLSP). His research interests lie at the intersection of Natural Language Processing and Machine Learning, in particular on areas relating to machine translation, semantics, and deep learning. Previously, he was assistant professor at the Nara Institute of Science and Technology (2012-2015) and research associate at NTT CS Labs (2009-2012). He received his B.S. in 2003 from Rice University, and PhD in 2009 from the University of Washington, both in Electrical Engineering.

Keynote Talk: TBD

Loïc Barrault
Meta AI

Abstract: TBD

Bio: Loïc Barrault (M) is a Research Scientist at Meta AI. Previously, he was an Associate Professor at LIUM, University of Le Mans. He obtained his PhD at the University of Avignon in 2008 in the field of automatic speech recognition. Then he did 2 years as researcher and 9 years as Associate Professor at LIUM, Le Mans Université followed by 2 years as Senior Lecturer in the NLP group of the University of Sheffield. Loïc Barrault participated in many international projects, namely EuroMatrix+, MateCAT, DARPA BOLT, and national projects, namely ANR Cosmat, "Projet d'Investissement d'Avenir" PACTE and a large industrial project PEA TRAD. He coordinated the EU ChistERA M2CR project and is currently actively involved in the ChistERA ALLIES project and the French ANR ON-TRAC project. His research work focuses on statistical and neural machine translation, by including linguistics aspects (factored neural machine translation), by considering multiple modalities (multimodal neural machine translation) and by designing lifelong learning methods for MT. He is one of the organisers of the Multimodal Machine Translation shared task at WMT.

Table of Contents

Program

Thursday, August 15, 2024

09:00 - 09:10 *Opening Remarks*

09:10 - 10:05 *Invited Talk 1: Kevin Duh (Johns Hopkins University)*

10:05 - 10:30 *Session 1: Booster Presentations*

KpopMT: Translation Dataset with Terminology for Kpop Fandom
JiWoo Kim, Yunsu Kim and JinYeong Bak

HeSum: a Novel Dataset for Abstractive Text Summarization in Hebrew
Itai Mondshine, Tzuf Paz-Argaman, Asaf Achi Mordechai and Reut Tsarfaty

Challenges in Urdu Machine Translation
Abdul Basit, Abdul Hameed Azeemi and Agha Ali Raza

Low-Resource Cross-Lingual Summarization through Few-Shot Learning with Large Language Models
Gyutae Park, Seojin Hwang and Hwanhee Lee

Enhancing Low-Resource NMT with a Multilingual Encoder and Knowledge Distillation: A Case Study
Aniruddha Roy, Pretam Ray, Ayush Maheshwari, Sudeshna Sarkar and Pawan Goyal

Rule-Based, Neural and LLM Back-Translation: Comparative Insights from a Variant of Ladin
Samuel Frontull and Georg Moser

AGE: Amharic, Ge'ez and English Parallel Dataset
Henok Biadglign Ademtew and Mikiyas Girma Birbo

Adopting Ensemble Learning for Cross-lingual Classification of Crisis-related Text On Social Media
Shareefa Ahmed Al Amer, Mark G. Lee and Phillip Smith

Rosetta Balcanica: Deriving a "Gold Standard" Neural Machine Translation (NMT) Parallel Dataset from High-Fidelity Resources for Western Balkan Languages
Edmon Begoli, Maria Mahbub and Sudarshan Srinivasan

Thursday, August 15, 2024 (continued)

Irish-based Large Language Model with Extreme Low-Resource Settings in Machine Translation
Khanh-Tung Tran, Barry O'Sullivan and Hoang D. Nguyen

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

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

Tuning LLMs with Contrastive Alignment Instructions for Machine Translation in Unseen, Low-resource Languages
Zhuoyuan Mao and Yen Yu

Linguistically Informed Transformers for Text to American Sign Language Translation
Abhishek Bharadwaj Varanasi, Manjira Sinha and Tirthankar Dasgupta

Leveraging Mandarin as a Pivot Language for Low-Resource Machine Translation between Cantonese and English
King Yiu Suen, Rudolf Chow and Albert Y.s. Lam

Enhancing Turkish Word Segmentation: A Focus on Borrowed Words and Invalid Morpheme
Soheila Behrooznia, Ebrahim Ansari and Zdenek Zabokrtsky

Super donors and super recipients: Studying cross-lingual transfer between high-resource and low-resource languages
Vitaly Protasov, Elisei Stakovskii, Ekaterina Voloshina, Tatiana Shavrina and Alexander Panchenko

12:30 - 14:00 *Lunch*

14:00 - 15:00 *Invited Talk 2: Loïc Barrault (Meta AI)*

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

KpopMT: Translation Dataset with Terminology for Kpop Fandom
JiWoo Kim, Yunsu Kim and JinYeong Bak

HeSum: a Novel Dataset for Abstractive Text Summarization in Hebrew
Itai Mondshine, Tzuf Paz-Argaman, Asaf Achi Mordechai and Reut Tsarfaty

Challenges in Urdu Machine Translation
Abdul Basit, Abdul Hameed Azeemi and Agha Ali Raza

Low-Resource Cross-Lingual Summarization through Few-Shot Learning with Large Language Models
Gyutae Park, Seojin Hwang and Hwanhee Lee

Enhancing Low-Resource NMT with a Multilingual Encoder and Knowledge Distillation: A Case Study
Aniruddha Roy, Pretam Ray, Ayush Maheshwari, Sudeshna Sarkar and Pawan Goyal

Rule-Based, Neural and LLM Back-Translation: Comparative Insights from a Variant of Ladin
Samuel Frontull and Georg Moser

AGE: Amharic, Ge'ez and English Parallel Dataset
Henok Biadglign Ademtew and Mikiyas Girma Birbo

Adopting Ensemble Learning for Cross-lingual Classification of Crisis-related Text On Social Media
Shareefa Ahmed Al Amer, Mark G. Lee and Phillip Smith

Rosetta Balcanica: Deriving a "Gold Standard" Neural Machine Translation (NMT) Parallel Dataset from High-Fidelity Resources for Western Balkan Languages
Edmon Begoli, Maria Mahbub and Sudarshan Srinivasan

Irish-based Large Language Model with Extreme Low-Resource Settings in Machine Translation
Khanh-Tung Tran, Barry O'Sullivan and Hoang D. Nguyen

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

16:00 - 17:30 *Session 4: Scientific Research Papers*

Tokenisation in Machine Translation Does Matter: The impact of different tokenisation approaches for Maltese
Kurt Abela, Kurt Micallef, Marc Tanti and Claudia Borg

Machine Translation Through Cultural Texts: Can Verses and Prose Help Low-Resource Indigenous Models?
Antoine Cadotte, Nathalie André and Fatiha Sadat

Learning-From-Mistakes Prompting for Indigenous Language Translation
You Cheng Liao, Chen-Jui Yu, Chi-Yi Lin, He-Feng Yun, Yen-Hsiang Wang, Hsiao-Min Li and Yao-Chung Fan

Finetuning End-to-End Models for Estonian Conversational Spoken Language Translation
Tiia Sildam, Andra Velve and Tanel Alumäe

Benchmarking Low-Resource Machine Translation Systems
Ana Alexandra Morim Da Silva, Nikit Srivastava, Tatiana Moteu Ngoli, Michael Röder, Diego Moussallem and Axel-Cyrille Ngonga Ngomo

17:30 - 17:40 *Closing remarks*