LoResMT 2023

The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)

Proceedings of the Workshop

May 6, 2023

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

In cooperation with

©2023 Association for Computational Linguistics

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 978-1-959429-55-5

Preface

Based on the success of past low-resource machine translation (MT) workshops at AMTA 2018, MT Summit 2019, AACL-IJCNLP 2020, AMTA 2021, and COLING 2022, we introduce LoResMT 2023 workshop at EACL 2023 (https://2023.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 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 a wide range of languages spoken around the world. In addition to research papers, the workshop also accepts relevant findings papers at EACL 2023 to be presented at LoResMT. 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)

Program Committee

Workshop Chairs

Atul Kr. Ojha, Atul Kr. Ojha, Data Science Institute, Insight Centre for Data Analytics, University of Galway & Panlingua Language Processing LLP
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
Alina Karakanta, Fondazione Bruno Kessler (FBK), University of Trento
Amirhossein Tebbifakhr, Fondazione Bruno Kessler
Anna Currey, AWS AI Labs
Aswarth Abhilash Dara, Amazon
Atul Kr. Ojha, University of Galway & Panlingua Language Processing LLP
Bharathi Raja Chakravarthi, University of Galway
Bogdan Babych, Heidelberg University
Chao-hong Liu, Potamu Research Ltd
Constantine Lignos, Brandeis University, USA
Daan van Esch, Google
Diptesh Kanojia, University of Surrey, UK
Duygu Ataman, University of Zurich
Ekaterina Vylomova, University of Melbourne, Australia
Eleni Metheniti, CLLE-CNRS and IRIT-CNRS
Flammie Pirinen, UiT Norgga árktalaš universitehta
Jade Abbott, Retro Rabbit
Jasper Kyle Catapang, University of the Philippines
Jindřich Libovický, Charles Univeristy
Jonathan Washington, Swarthmore College
Majid Latifi, UPC University
Maria Art Antonette Clariño, University of the Philippines Los Baños
Mathias Müller, University of Zurich
Nathaniel Oco, De La Salle University
Rajdeep Sarkar, University of Galway
Rico Sennrich, University of Zurich
Saliha Muradoglu, The Australian National University
Sangjee Dondrub, Qinghai Normal University
Sardana Ivanova, University of Helsinki
Shantipriya Parida, Silo AI
Sunit Bhattacharya, Charles University

Surafel M. Lakew, Amazon.com, Inc
Wen Lai, LMU Munich
Valentin Malykh, Huawei Noah's Ark lab and Kazan Federal University
Varvara Logacheva Skolkovo, Institute of Science and Technology

Secondary Reviewers

Gaurav Negi, University of Galway

Table of Contents

Train Global, Tailor Local: Minimalist Multilingual Translation into Endangered Languages
Zhong Zhou, Jan Niehues and Alexander Waibel 1

Multilingual Bidirectional Unsupervised Translation through Multilingual Finetuning and Back-Translation
Bryan Li, Mohammad Sadegh Rasooli, Ajay Patel and Chris Callison-burch 16

PEACH: Pre-Training Sequence-to-Sequence Multilingual Models for Translation with Semi-Supervised Pseudo-Parallel Document Generation
Alireza Salemi, Amirhossein Abaskohi, Sara Tavakoli, Azadeh Shakery and Yadollah Yaghoobzadeh 32

A Simplified Training Pipeline for Low-Resource and Unsupervised Machine Translation
Àlex R. Atrio, Alexis Allemann, Ljiljana Dolamic and Andrei Popescu-belis 47

Language-Family Adapters for Low-Resource Multilingual Neural Machine Translation
Alexandra Chronopoulou, Dario Stojanovski and Alexander Fraser 59

Improving Neural Machine Translation of Indigenous Languages with Multilingual Transfer Learning
Wei-rui Chen and Muhammad Abdul-mageed 73

Investigating Lexical Replacements for Arabic-English Code-Switched Data Augmentation
Injy Hamed, Nizar Habash, Slim Abdennadher and Ngoc Thang Vu 86

Measuring the Impact of Data Augmentation Methods for Extremely Low-Resource NMT
Annie Lamar and Zeyneb Kaya 101

Findings from the Bambara - French Machine Translation Competition (BFMT 2023)
Ninoh Agostinho Da Silva, Tunde Ajayi, Alex Antonov, Panga Azazia Kamate, Moussa Coulibaly, Mason Del Rio, Yacouba Diarra, Sebastian Diarra, Chris Emezue and Joel Hamilcaro 110

Evaluating Sentence Alignment Methods in a Low-Resource Setting: An English-YorùBá Study Case
Edoardo Signoroni and Pavel Rychlý 123

Program

Saturday, May 6, 2023

09:00 - 09:15 *Opening Remarks*

09:15 - 10:05 *Invited Talk 1*

10:05 - 10:30 *Session 1: Finding Papers*

10:30 - 11:15 *COFFEE/TEA BREAK*

11:15 - 12:45 *Session 2: Scientific Research Papers*

Train Global, Tailor Local: Minimalist Multilingual Translation into Endangered Languages
Zhong Zhou, Jan Niehues and Alexander Waibel

Measuring the Impact of Data Augmentation Methods for Extremely Low-Resource NMT
Annie Lamar and Zeyneb Kaya

Language-Family Adapters for Low-Resource Multilingual Neural Machine Translation
Alexandra Chronopoulou, Dario Stojanovski and Alexander Fraser

Multilingual Bidirectional Unsupervised Translation through Multilingual Finetuning and Back-Translation
Bryan Li, Mohammad Sadegh Rasooli, Ajay Patel and Chris Callison-burch

12:45 - 14:15 *Lunch*

14:15 - 15:00 *Invited Talk 2*

15:00 - 15:30 *Session 3: Finding Papers*

A Simplified Training Pipeline for Low-Resource and Unsupervised Machine Translation
Àlex R. Atrio, Alexis Allemann, Ljiljana Dolamic and Andrei Popescu-belis

15:45 - 16:30 *COFFEE/TEA BREAK*

Saturday, May 6, 2023 (continued)

16:30 - 18:05 *Session 4: Scientific Research Papers*

Improving Neural Machine Translation of Indigenous Languages with Multilingual Transfer Learning
Wei-rui Chen and Muhammad Abdul-mageed

PEACH: Pre-Training Sequence-to-Sequence Multilingual Models for Translation with Semi-Supervised Pseudo-Parallel Document Generation
Alireza Salemi, Amirhossein Abaskohi, Sara Tavakoli, Azadeh Shakery and Yadollah Yaghoobzadeh

Investigating Lexical Replacements for Arabic-English Code-Switched Data Augmentation
Injy Hamed, Nizar Habash, Slim Abdennadher and Ngoc Thang Vu

Evaluating Sentence Alignment Methods in a Low-Resource Setting: An English-YorùBá Study Case
Edoardo Signoroni and Pavel Rychlý

Findings from the Bambara - French Machine Translation Competition (BFMT 2023)
Ninoh Agostinho Da Silva, Tunde Ajayi, Alex Antonov, Panga Azazia Kamate, Moussa Coulibaly, Mason Del Rio, Yacouba Diarra, Sebastian Diarra, Chris Emezue and Joel Hamilcaro

18:05 - 18:15 *Closing remarks*