Proceedings of Machine Translation Summit XVIII

https://mtsummit2021.amtaweb.org

4th Workshop on Technologies for MT of Low Resource Languages

Organizers:
John Ortega, Atul Kr. Ojha, Katharina Kann and Chao-Hong Liu

Proceedings of the 4th Workshop on Technologies for Machine Translation of Low-resource Languages

Organizers
John Ortega[1] jortega@cs.nyu.edu
Atul Kr. Ojha[2,3] atulkumar.ojha@insight-centre.org
Katharina Kann[4] katharina.kann@colorado.edu
Chao-Hong Liu[5] ch.liu@acm.org
[1]New York University
[2]Data Science Institute, NUIG, Galway
[3]Panlingua Language Processing LLP, New Delhi
[4]University of Colorado Boulder
[5]Potamu Research Ltd

1 Aim of the Workshop

Based on the success of past low-resource machine translation (LoResMT) workshops at AACL-IJCNLP 2020[1], MT Summit 2019[2] and AMTA 2018[3], we introduce the 4th LoResMT workshop co-located at the MT Summit 2021[4] conference. Like its predecessors, this workshop will bring together researchers and translators of low-resource languages to compare and contrast how each use digital technology for translation. Specifically, the workshop focuses on novel advances on the coverage of even more languages than past workshops with different geographical presence, degree of diffusion and digitization.

The proceedings of LoResMT 2021 contain original work on low-resource translation which includes, but is not limited to, machine translation (MT) systems that include word tokenizers/de-tokenizers, word segmenters, and morphological analyzers. Additionally, we explicitly solicited novel work covering translations of COVID-related text and their practical use for low-resource communities.

The goal of this workshop was to begin to close the gap between low-resource translation systems and their practical use in the real world. Online systems and original research that used by native speakers of low-resource languages was of particular interest. Therefore, we encouraged the authors of research papers to include a statement about the impact of their proposed approaches on the quality of MT output and how they can be used in the real world.

The need for receiving relevant, fast and up-to-date information in one's language is today more important than ever, especially under the current crisis conditions. MT is a vital tool for facilitating communication and access to information. For most of the world's languages, the lack of training data has long posed a major obstacle to developing high quality MT systems,

[1]http://aacl2020.org
[2]https://www.mtsummit2019.com
[3]https://amtaweb.org
[4]https://amtaweb.org/mt-summit2021

excluding the speakers of these low-resource languages from the benefits of MT. In the past few years, MT performance has improved significantly, mainly due to the new possibilities opened up by neural machine translation (NMT). With the development of novel techniques, such as multilingual translation and transfer learning, the use of MT is no longer a privilege restricted to users of a dozen popular languages. Consequently, there has been an increasing interest in the MT community to expand the coverage to more languages with different geographical presence, degree of diffusion and digitization. Today, research groups on all continents are working on MT. The number of languages offered by publicly available MT engines is increasing, reaching almost 200 languages at the moment of writing. We are witnessing an interesting phenomenon of collaborative projects to promote MT for under-represented languages, involving partners from all over the globe, participating on a voluntary basis. These developments have created a colourful, promising future for low-resource languages on the MT map.

Despite all these encouraging developments in MT technologies, creating an MT system for a new language from scratch or even improving an existing system still requires a considerable amount of work in collecting the pieces necessary for building such systems. Due to the data-hungry nature of NMT approaches, the need for parallel and monolingual corpora in different domains is never saturated. The development of MT systems requires reliable test sets and evaluation benchmarks. In addition, MT systems still rely on several natural language processing (NLP) tools to pre-process human-generated texts in the forms that are required as input for MT systems and post-process the MT output in proper textual forms in the target language. These NLP tools include, but are not limited to, word tokenizers/de-tokenizers, word segmenters, and morphological analysers. The performance of these tools has a great impact on the quality of the resulting translation. There is only limited discussion on these NLP tools, their methods, their role in training different MT systems, and their coverage of support in the many languages of the world.

LoResMT provides a discussion panel for researchers working on MT systems/methods for low-resource languages in general. This year we received research papers covering a wide range of languages spoken in Asia, Latin America, Africa and Europe. These languages are: Arabic, Albanian, Ashaninka, Bengali, Dutch, Eastern Pokomchi, English, French, German, Gujarati, Hindi, Inuktitut, Indonesian, Irish, Japanese, Kannada, Khasi, Konkani, Korean, Mayan, Malayalam, Marathi, Odia, Punjabi, Quechua, Sanskrit, Spanish, Tamil, Telugu, Turkish and Urdu. We received both resource papers (monolingual, parallel corpora, social media, sentiment and formalisms) and methods papers, ranging from unsupervised, zero-shot to multilingual NMT, MT evaluation. The acceptance rate of LoResMT this year is 43.4%.

In addition to soliciting research papers, we organized a shared task to be presented at the workshop, where we asked participants to build novel MT systems for COVID-related texts in low-resource languages, including one sign language. The shared task aimed to encourage research on MT systems for three language pairs: English↔Irish, English↔Marathi and Taiwanese Sign Language↔Traditional Chinese. The corpora along with additional information on downloading videos for sign language for machine translation tasks are freely available on Github[5]. Six shared task papers are published as part of the proceedings, along with the findings of the shared task.

2 Invited Speakers (listed alphabetically)

We are happy our dear colleagues Barry Haddow, Catherine Muthoni Gitau, Mathias Müller and Mona Diab have prepared talks on four important topics for LoResMT 2021.

[5] https://github.com/loresmt/loresmt-2021

2.1 Barry Haddow, Aveni and University of Edinburgh

MT for Low-resource Languages: Progress and Open Problems

Current machine translation (MT) systems have reached the stage where researchers are now debating whether or not they can rival human translators in performance. However these MT systems are typically trained on data sets consisting of tens or even hundreds of millions of parallel sentences. There is an increasing body of research which considers the problem of training MT systems on much smaller data sets. The aim of this talk is to provide a broad survey of the techniques that have been applied to low-resource MT, presenting a data-centric taxonomy, and indicating gaps. The talk is based on a survey paper which is currently being finalised, and that we aim to release during August 2021.

About the Speaker

Barry Haddow is a senior researcher in the School of Informatics at the University of Edinburgh. He has worked in machine translation for more than 10 years, and his current interests include low-resource MT, spoken language translation and evaluation of MT. Barry coordinates the annual WMT conference on machine translation and associated shared tasks.

2.2 Catherine Muthoni Gitau, African Institute for Mathematical Sciences

Challenges and Advances in MT Systems for African Languages

Africa is known to be the highest linguistically diverse continent with over 2,000 languages across the continent representing about 30% of the languages spoken around the world and despite this, African languages account only a small fraction of available language resources making them low-resourced. There's minimal attention that's being given to machine translation for African languages and therefore, there is not much work or research regarding the problems that arise when using machine translation techniques. However, there's been an increase in work around machine translation for African languages in the last couple of years with the aim of addressing some of these challenges. In this talk, I will present on the challenges currently being faced in the development of machine translation systems for African languages as well as work that's being done to alleviate some of these challenges. I will go into detail about the work of the Masakhane community whose mission is to strengthen and spur NLP research in African languages, for Africans, by Africans with a focus on work that's being done on machine translation.

About the Speaker

Catherine Gitau is a natural language processing researcher and engineer at Proto, a company that builds multilingual AI chatbots for contact centers. She recently completed her Masters' in Machine Intelligence at the African Institute of Mathematical Sciences (AIMS) under the African Masters' in Machine Intelligence (AMMI) program and is an active member of the Masakhane Community whose mission is to strengthen NLP research in African Languages. Her research interests include natural language processing and low-resource machine translation.

2.3 Mathias Müller, Institut für Computerlinguistik, Universität Zürich

On Meaningful Evaluation of Machine Translation Systems

In this talk Mathias will discuss best practices for evaluating machine translation systems. The goal of defining such best practices is to ensure that conclusions drawn from experiments are valid, and that perceived scientific progress is in fact real. Areas we will touch on during the talk include selecting data for experiments, significance testing and the special role of low-resource

experiments. Mathias is looking forward to a lively discussion, leading to a set of practices that we can all advocate in the future and implement in our own research.

About the Speaker

Mathias is a post-doc and lecturer at the University of Zurich. His current main interests are 1) the meta-sciences of scientific integrity, methodology and reproducibility applied to machine translation and 2) sign language translation. In his personal life, as a father of two, he advocates the best practice of not working on weekends.

2.4 Mona Diab, Facebook, George Washington University

Trustworthy Human Evaluation Frameworks for MT

How do we establish trust in our machine translation systems performance? Typical evaluations rely on reference translations that are curated from humans, serving as gold data annotations. In this talk I will examine this assumption and propose ways to ensure we have trustworthy reference data with closer to real translation perception (higher meaningfulness gauging). I will propose a holistic view of translation evaluation as an ecosystem and a framework especially for low resource scenarios.

About the Speaker

Mona is a Research Scientist with Facebook AI and she is also a full Professor of CS at the George Washington University where she heads the CARE4Lang NLP Lab. Before joining FB, she led the Lex Conversational AI project within Amazon AWS AI. Her interests span building robust technologies for low resource scenarios with a special interest in Arabic technologies, (mis)information propagation, computational socio-pragmatics, NLG evaluation metrics, and resource creation. She has served the community in several capacities: Elected President of SIGLEX and SIGSemitic. She currently serves as the elected VP-Elect for ACL SIGDAT, the board supporting EMNLP conferences. She has delivered tutorials and organized numerous workshops and panels around Arabic processing. She is a cofounder of CADIM (Consortium on Arabic Dialect Modeling, previously known as Columbia University Arabic Dialects Modeling Group), in 2005, which served as a world renowned reference point on Arabic Language Technologies. Moreover she helped establish two research trends in NLP, namely computational approaches to Code Switching and Semantic Textual Similarity. She is also a founding member of the *SEM conference, one of the top tier conferences in NLP. She currently serves as the senior area chair for multiple top tier conferences. She has published more than 230 peer reviewed articles.

3 Co-organizing Committee

- Jade Abbott, Retro Rabbit
- Jonathan Washington, Swarthmore College
- Nathaniel Oco, National University (Philippines)
- Surafel Melaku Lakew, Amazon AI
- Tommi A Pirinen, University of Hamburg
- 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

4 Program Committee

- Alberto Poncelas, Rakuten
- Alina Karakanta, Fondazione Bruno Kessler
- Amirhossein Tebbifakhr, Fondazione Bruno Kessler
- Anna Currey, Amazon AI
- Atul Kr. Ojha, National Univerity of Ireland Galway
- Beatrice Savoldi, beatrice savoldi
- Bharathi Raja Chakravarthi, National University of Ireland Galway
- Bogdan Babych, Heidelberg University
- Chao-Hong Liu, Potamu Research Ltd
- Duygu Ataman, UZH
- Eleni Metheniti, CLLE - CNRS, IRIT
- Félix Arturo Oncevay Marcos, University of Edinburgh
- Flammie A Pirinen, UiT–Norgga árktalaš universitehta
- Jasper Kyle Catapang, University of Birmingham
- John McCrae, National University of Ireland Galway
- John E Ortega, New York University
- Jonathan N Washington, Swarthmore College
- Katharina Kann, University of Colorado Boulder
- Koel Dutta Chowdhury, Saarland University
- Liangyou Li, Huawei Noah's Ark Lab
- Majid Latifi, National College of Ireland (NCI)
- Maria Art Antonette, Clariño University of the Philippines Los Baños
- Mathias Müller, University of Zurich
- Mehdi Rezagholizadeh, Huawei Noah's Ark Lab
- Nathaniel Oco, Philippines
- Priya Rani, National University of Ireland Galway
- Rico Sennrich, University of Zurich
- Sangjie Duanzhu, Qinghai Normal University
- Santanu Pal, Wipro

- Sardana Ivanova, University of Helsinki
- Shabnam Tafreshi, University of Maryland
- Shantipriya Parida, Idiap Research Institute, Martigny, Switzerland
- Sina Ahmadi, Insight Centre for Data Analytics
- Sunit Bhattacharya, Charles University
- Surafel M Lakew, Amazon AI
- Thepchai Supnithi, NECTEC, National Science and Technology Development Agency
- Tsz Kin Lam, Heidelberg University
- Valentin Malykh, Huawei Noah's Ark lab and Kazan Federal University
- Vlad Tyshkevich, Brooklyn College, City University of New York

Contents