COLING

International Conference on Computational Linguistics

Proceedings of the Conference and Workshops

COLING

Volume 29 (2022), No. 2

Proceedings of The Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)

The 29th International Conference on Computational Linguistics

October 12 - 17, 2022
Gyeongju, Republic of Korea

ISSN 2951-2093

Preface

Based on the success of past low-resource machine translation (MT) workshops at AMTA 2018 (https://amtaweb.org/), MT Summit 2019 (https://www.mtsummit2019.com), AACL-IJCNLP 2020 (http://aacl2020.org/), and AMTA 2021, we introduce the Fifth LoResMT workshop at COLING 2022. 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. Consequently, there has been an increasing interest in the community to expand the coverage to more languages with different geographical presences, degrees of diffusion and digitalization. However, the goal to increase MT coverage for more users speaking diverse languages is limited by the fact the MT methods demand vast amounts of data to train quality systems, which has posed a major obstacle to developing MT systems for low-resource languages. Therefore, developing comparable MT systems with relatively small datasets is still highly desirable.

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 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, under-represented, ethnic and endangered 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 Cebuano, English, Filipino, Gujarati, Haitian, Indonesian, Jamaican, Kannada, Lambani, Luhya, Malaysian, Marathi, Persian, Romanian, Spanish Sign and Swahili. We received both resource papers (monolingual, parallel corpora, formalisms) and methods papers, ranging from unsupervised, transfer-learning, and zero-shot to multilingual NMT. The workshop also received papers on Sign language and evaluation methods for MT. The acceptance rate of LoResMT this year is 53%. In addition to the research papers, the workshop hosts two invited talks. Vishrav Chaudhary gives the first invited talk from Microsoft Turing, who described the Mining Methods for Low Resource MT. In the second invited talk, Pushpak Bhattacharyya from the Indian Institute of Technology Bombay explains multilingual computation, focusing on Machine Translation, in a low-resource setting.

We would sincerely like to thank all of our program committee members for their valuable help in reviewing the submissions and providing their constructive feedback for improving the workshop: Alberto Poncelas, Alina Karakanta, Amirhossein Tebbifakhr, Anna Currey, Arturo Oncevay, Aswath Abhilash Dara, Barry Haddow, Beatrice Savoldi, Bogdan Babych, Constantine Lignos, Daan van Esch, Diptesh Kanojia, Ekaterina Vylomova, Eleni Metheniti, Eva Vanmassenhove, Jasper Kyle Catapang, Liangyou Li, Majid Latifi, Maria Art Antonette Clariño, Mathias Müller, Monojit Choudhury, Nathaniel Oco, Rico Sennrich, Saliha Muradoglu, Sangjee Dondrub, Santanu Pal, Sardana Ivanova, Shantipriya

Parida, Sunit Bhattacharya, Surafel M. Lakew, Thepchai Supnithi, Valentin Malykh, Vukosi Marivate, Wen Lai, Xiaobing Zhao. We are grateful to our invited speakers for their engaging presentations and the insights they brought to the workshop. We would further like to thank the workshop chairs, Sadao Kurohashi, Seung-Hoon Na, and Damira Mrsic, for their guidance and support in organising the workshop, as well as the remote presentation chair, for the hard work in preparing the workshop page. Finally, we are grateful to all the authors who submitted and presented their work to LoResMT.

Atul Kr. Ojha and Chao-Hong Liu
(On behalf of the workshop chairs)

Organizing Committee

Workshop Chairs

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
Jade Abbott, Retro Rabbit
Jonathan Washington, Swarthmore College
Nathaniel Oco, National University (Philippines)
Tommi A Pirinen, UiT The Arctic University of Norway, Tromsø
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

Alberto Poncelas, Rakuten, Singapore
Alina Karakanta, Fondazione Bruno Kessler
Amirhossein Tebbifakhr, Fondazione Bruno Kessler
Anna Currey, Amazon Web Services
Aswarth Abhilash Dara, Amazon
Arturo Oncevay, University of Edinburgh
Atul Kr. Ojha, Data Science Institute, Insight Centre for Data Analytics, 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
Francis Tyers, Indiana University
Kalika Bali, MSRI Bangalore, India
Koel Dutta Chowdhury, Saarland University (Germany)
Jade Abbott, Retro Rabbit
Jasper Kyle Catapang, University of the Philippines
John P. McCrae, DSI, University of Galway
Liangyou Li, Noah's Ark Lab, Huawei Technologies
Majid Latifi, University of York, York, UK
Maria Art Antonette Clariño, University of the Philippines Los Baños
Mathias Müller, University of Zurich
Monojit Choudhury, Microsoft Turing
Nathaniel Oco, National University (Philippines)
Rico Sennrich, University of Zurich
Saliha Muradoglu, The Australian National University
Sangjee Dondrub, Qinghai Normal University
Santanu Pal, WIPRO AI
Sardana Ivanova, University of Helsinki

Shantipriya Parida, Silo AI
Sunit Bhattacharya, Charles University
Surafel Melaku Lakew, Amazon AI
Tommi A Pirinen, UiT The Arctic University of Norway, Tromsø
Wen Lai, Center for Information and Language Processing, LMU Munich
Valentin Malykh, Huawei Noah's Ark lab and Kazan Federal University

Invited Speakers

1. Pushpak Bhattacharya, IIT-Bomaby

   Title: Low Resource Machine Translation- A Perspective

   Abstract: This talk is on multilingual computation, focussing on Machine Translation, in low-resource settings. Tasks in this area have to grapple with the characteristic problem of disambiguation in the face of resource scarcity, which is the reality for most languages and also arguably for ANY language when it comes to very high-end NLP tasks like, say, 100% disambiguation as demanded by pure interlingua based MT. Starting with our early work on rule-based MT, we move to our research in Low Resource Machine Translation, covering SMT, NMT, segmenting, pivoting, and semi and unsupervised MT. The findings covered in this talk are based on contributions by many students and researchers over many years, reported in top conferences and journals.

   About the Speaker: Prof Pushpak Bhattacharyya (http://www.cse.iitb.ac.in/ pb) is a Professor of Computer Science and Engineering at IIT Bombay. He has done extensive research in Natural Language Processing and Machine Learning. Some of his noteworthy contributions are IndoWordnet, Eye Tracking assisted NLP, Low Resource MT and Knowledge Graph-Deep Learning Synergy in Information Extraction and Question Answering. He has published close to 400 research papers, has authored/co-authored 6 books including a textbook on machine translation, and has guided more than 350 students for their Ph.D., master's, and Undergraduate thesis. Prof. Bhattacharyya is a Fellow of the National Academy of Engineering, Abdul Kalam National Fellow, Distinguished Alumnus of IIT Kharagpur, and past President of ACL.

2. Vishrav Chaudhary, Microsoft Turing

   Title: Mining Methods for Low Resource MT

   Abstract: TBD

   About the Speaker: Vishrav Chaudhary, Senior Principal Researcher at Microsoft Turing, leading efforts around large-scale multilingual models. In the past, Vishrav's research has been focused on several aspects of Machine Translation including Low-Resource translation and Quality Estimation, Cross-lingual understanding, Transfer Learning, Efficient model architectures and Domain Adaptation.

Table of Contents

Conference Program

Sunday, October 16, 2022 (GMT+9)

09:00–10:05 **Inagural Session**
Chair: Atul Kr. Ojha

09:00–09:15 *Opening remarks*
Workshop Chairs

09:15–10:05 *Keynote talk: Mining Methods for Low Resource MT*
Vishrav Chaudhary, Microsoft Turing

10:00–10:30 **Q&A Session 1**
Chair: Ekaterina Vylomova

10:05–10:15 *Very Low Resource Sentence Alignment: Luhya and Swahili*
Everlyn Chimoto and Bruce Bassett

10:15–10:30 *Multiple Pivot Languages and Strategic Decoder Initialization Helps Neural Machine Translation*
Shivam Mhaskar and Pushpak Bhattacharyya

10:30–11:00 ***COFFEE/TEA BREAK***

11:00–12:30 **Q&A Session 2**
Chair: Jonathan Washington

11:00–11:30 *Known Words Will Do: Unknown Concept Translation via Lexical Relations*
Winston Wu and David Yarowsky

11:30–12:00 *The Only Chance to Understand: Machine Translation of the Severely Endangered Low-resource Languages of Eurasia*
Anna Mosolova and Kamel Smaili

12:00–12:30 *Data-adaptive Transfer Learning for Translation: A Case Study in Haitian and Jamaican*
Nathaniel Robinson, Cameron Hogan, Nancy Fulda and David R. Mortensen

12:30–14:00 ***LUNCH BREAK***

Sunday, October 16, 2022 (GMT+9) (continued)

14:00–14:55 **Keynote talk**
Chair: Chao-Hong Liu

14:00–14:55 *Low Resource Machine Translation- A Perspective*
Pushpak Bhattacharyya, Indian Institute of Technology Bombay

14:55–15:30 **Q&A Session 3**
Chair: Nathaniel Oco

14:55–15:05 *Augmented Bio-SBERT: Improving Performance for Pairwise Sentence Tasks in Biomedical Domain*
Sonam Pankaj and Amit Gautam

15:05–15:15 *Machine Translation for a Very Low-Resource Language - Layer Freezing Approach on Transfer Learning*
Amartya Chowdhury, Deepak K. T., Samudra Vijaya K and S. R. Mahadeva Prasanna

15:15–15:30 *HFT: High Frequency Tokens for Low-Resource NMT*
Edoardo Signoroni and Pavel Rychlý

15:30–16:00 ***COFFEE/TEA BREAK***

16:00–17:00 **Q&A Session 4**
Chair: Valentin Malykh

16:00–16:30 *Romanian Language Translation in the RELATE Platform*
Vasile Pais, Maria Mitrofan and Andrei-Marius Avram

16:30–17:00 *Translating Spanish into Spanish Sign Language: Combining Rules and Data-driven Approaches*
Luis Chiruzzo, Euan McGill, Santiago Egea-Gómez and Horacio Saggion

Sunday, October 16, 2022 (GMT+9) (continued)

17:00–18:00 **Q&A Session 5**
Chair: Xiaobing Zhao

17:00–17:12 *Benefiting from Language Similarity in the Multilingual MT Training: Case Study of Indonesian and Malaysian*
Alberto Poncelas and Johanes Effendi

17:12–17:24 *A Preordered RNN Layer Boosts Neural Machine Translation in Low Resource Settings*
Mohaddeseh Bastan and Shahram Khadivi

17:24–17:36 *Exploring Word Alignment towards an Efficient Sentence Aligner for Filipino and Cebuano Languages*
Jenn Leana Fernandez and Kristine Mae M. Adlaon

17:36–17:50 *Aligning Word Vectors on Low-Resource Languages with Wiktionary*
Mike Izbicki

17:50–18:00 *Valedictory Session*
Workshop Chairs