AACL-IJCNLP 2020

The 3rd Workshop on Technologies for MT of Low Resource Languages (LoResMT 2020)

https://sites.google.com/view/loresmt/

Proceedings of the Workshop

December 4, 2020
Online

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 978-1-952148-96-5

Preface

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. Machine Translation (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, 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 digitalization. Today, research groups are working on MT in all continents. 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 machine translation 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 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, 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 and under-represented 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: Ashaninka, Assamese, Bambara, Bengali, Bhojpuri, Bicolano, Cebuano, English, Esperanto, French, Greek, Hiligaynon, Hindi, Ilocano, Kurdish, Manipuri, Pangasinense, Russian, Tagalog, Tamil, Vietnamese. We received both resource papers (monolingual, parallel corpora, formalisms) and methods papers, ranging from unsupervised, zero-shot to multilingual NMT. The acceptance rate of LoResMT this year is 60%.

In addition to the research papers, the workshop organised a shared task, where we solicit participants to submit novel zero-shot NMT systems for the language pairs Hindi-Bhojpuri, Hindi-Magahi and Russian-Hindi. Two shared task papers are archived in the proceedings, along with the findings of the shared task.

LoResMT hosts two invited talks. The first one is given by Grace Tang and Alp Öktem from Translators without Borders, who presented Translators without Borders' (TWB) Gamayun project where they aim to enable two-way communication using language technology for marginalized language speakers. In the second invited talk, Bonaventure Dossou and Chris Emezue from Jacobs University Bremen & Technical University of Munich describe the challenges of creating an NMT system for a new language pair, Fon-French.

We would sincerely like to thank all of our program committee members for their valuable help in reviewing the submissions and for providing their constructive feedback for improving the workshop: Alberto Poncelas, Amirhossein Tebbifakhr, Anna Currey, Arturo Oncevay, Atul Kr. Ojha, Bharathi

Raja Chakravarthi, Beatrice Savoldi, Bogdan Babych, Duygu Ataman, Eleni Metheniti, Francis Tyers, Kalika Bali, Koel Dutta Chowdhury, Jasper Kyle Catapang, John Ortega, Liangyou Li, Maria Art Antonette Clariño, Mathias Müller, Nathaniel Oco, Rico Sennrich, Sangjee Dondrub, Santanu Pal, Sardana Ivanova, Shantipriya Parida, Sunit Bhattacharya, Surafel Melaku Lakew, Tommi A Pirinen, Valentin Malykh. We are grateful to our invited speakers for their engaging presentations and insights they brought to the workshop. We would further like to thank the workshop chairs, Gao Wei and Lu Wang, for their guidance and support in organising the workshop, as well as the remote presentation chair, Zhongqing Wang, for the hard work in preparing the workshop page. Finally, we are very grateful to all the authors who submitted and presented their work to LoResMT.

On behalf of the organizing committee,
Alina Karakanta
Chao-Hong Liu

Workshop Chairs, Organizers & Editors

Alina Karakanta, Fondazione Bruno Kessler
Atul Kr. Ojha, DSI, National University of Ireland Galway & Panlingua Language Processing LLP
Chao-Hong Liu, Iconic Translation Machines, RWS Group
Jade Abbott, Retro Rabbit
John Ortega, Blackboard Insurance, Columbia University, and New York University
Jonathan Washington, Swarthmore College
Nathaniel Oco, Philippines
Surafel Melaku Lakew, Fondazione Bruno Kessler
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

Shared Task Organising Committee

Alina Karakanta, Fondazione Bruno Kessler
Atul Kr. Ojha, DSI, National University of Ireland Galway & Panlingua Language Processing LLP
Chao-Hong Liu, Iconic Translation Machines, RWS Group
Valentin Malykh, Huawei Noah's Ark lab and Kazan Federal University

Program Committee

Alberto Poncelas, ADAPT, Dublin City University
Alina Karakanta, Fondazione Bruno Kessler
Amirhossein Tebbifakhr, Fondazione Bruno Kessler
Anna Currey, Amazon Web Services
Arturo Oncevay, University of Edinburgh
Atul Kr. Ojha, DSI, National University of Ireland Galway & Panlingua Language Processing LLP
Bharathi Raja Chakravarthi, DSI, National University of Ireland Galway
Beatrice Savold, University of Trento
Bogdan Babych, Heidelberg University
Chao-Hong Liu, Iconic Translation Machines, RWS Group
Duygu Ataman, University of Zurich
Eleni Metheniti, CLLE-CNRS and IRIT-CNRS
Francis Tyers, Indiana University
Kalika Bali, MSRI Bangalore, India
Koel Dutta Chowdhury, Saarland University, Germany
Jasper Kyle Catapang, University of the Philippines
John Ortega, Blackboard Insurance, Columbia University, and New York University
Liangyou Li, Noah's Ark Lab, Huawei Technologies
Maria Art Antonette Clariño, University of the Philippines Los Baños
Mathias Müller, University of Zurich
Nathaniel Oco, Philippines
Rico Sennrich, University of Zurich
Sangjee Dondrub, Qinghai Normal University
Santanu Pal, WIPRO AI
Sardana Ivanova, University of Helsinki

Shantipriya Parida, Idiap Research Institute
Surafel Melaku Lakew, Fondazione Bruno Kessler
Tommi A Pirinen, University of Hamburg
Valentin Malykh, Huawei Noah's Ark lab and Kazan Federal University

Additional Reviewers:

Sunit Bhattacharya, Charles University, Prague

Invited Speakers

Grace Tang, Translators Without Borders
Alp Öktem, Translators Without Borders
Bonaventure Dossou, Jacobs University Bremen
Chris Emezue, Technical University of Munich

Table of Contents

Conference Program

Friday, December 04, 2020 (Times GMT+8)

08:00–08:15 Opening remarks

09:00–11:00 Q&A Session 1 (Asia, Americas)

Overcoming Resistance: The Normalization of an Amazonian Tribal Language
John Ortega, Richard Alexander Castro-Mamani and Jaime Rafael Montoya Samame

Bridging Philippine Languages With Multilingual Neural Machine Translation
Renz Iver Baliber, Charibeth Cheng, Kristine Mae Adlaon and Virgion Mamonong

Neural Machine Translation for Extremely Low-Resource African Languages: A Case Study on Bambara
Allahsera Auguste Tapo, Bakary Coulibaly, Sébastien Diarra, Christopher Homan, Julia Kreutzer, Sarah Luger, Arthur Nagashima, Marcos Zampieri and Michael Leventhal

11:00–13:45 *Lunch*

13:45–14:30 *Invited talk: Fon-French Neural Machine Translation*
Bonaventure Dossou and Chris Emezue

14:30–15:30 Findings of the shared task

Findings of the LoResMT 2020 Shared Task on Zero-Shot for Low-Resource languages
Atul Kr. Ojha, Valentin Malykh, Alina Karakanta and Chao-Hong Liu

Zero-Shot Neural Machine Translation: Russian-Hindi @LoResMT 2020
Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray and Sivaji Bandyopadhyay

Unsupervised Approach for Zero-Shot Experiments: Bhojpuri–Hindi and Magahi–Hindi@LoResMT 2020
Amit Kumar, Rajesh Kumar Mundotiya and Anil Kumar Singh

15:30–16:30 *Invited talk: Gamayun: using language technology to improve humanitarian communication*
Grace Tang and Alp Öktem

Friday, December 04, 2020 (Times GMT+8) (continued)

16:30–17:20 ***Coffee break***

17:20–19:00 **Q&A Session 2 (Asia, Europe)**

Zero-shot translation among Indian languages
Rudali Huidrom and Yves Lepage

Improving Multilingual Neural Machine Translation For Low-Resource Languages: French, English - Vietnamese
Thi-Vinh Ngo, Phuong-Thai Nguyen, Thanh-Le Ha, Khac-Quy Dinh and Le-Minh Nguyen

EnAsCorp1.0: English-Assamese Corpus
Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray and Sivaji Bandyopadhyay

Unsupervised Neural Machine Translation for English and Manipuri
Salam Michael Singh and Thoudam Doren Singh

Effective Architectures for Low Resource Multilingual Named Entity Transliteration
Molly Moran and Constantine Lignos

19:00–21:00 ***Dinner***

21:00–22:00 ***Panel discussion***

Friday, December 04, 2020 (Times GMT+8) (continued)

22:00–22:15 Closing remarks

22:15–00:00 Q&A Session 3

Towards Machine Translation for the Kurdish Language
Sina Ahmadi and Maraim Masoud

An Ensemble Method for Producing Word Representations focusing on the Greek Language
Michalis Lioudakis, Stamatis Outsios and Michalis Vazirgiannis

Using Multiple Subwords to Improve English-Esperanto Automated Literary Translation Quality
Alberto Poncelas, Jan Buts, James Hadley and Andy Way

Investigating Low-resource Machine Translation for English-to-Tamil
Akshai Ramesh, Venkatesh Balavadhani parthasa, Rejwanul Haque and Andy Way