Introduction to the Special Issue on Machine Translation for Low-Resource Languages

Chao-Hong Liu¹ · Alina Karakanta² · Audrey N. Tong³ · Oleg Aulov³ · Ian M. Soboroff³ · Jonathan Washington⁴ · Xiaobing Zhao⁵

Published online: 6 January 2021

© The Author(s), under exclusive licence to Springer Nature B.V. part of Springer Nature 2021

Machine translation (MT) technologies have improved significantly in the last two decades, with developments in phrase-based statistical MT (SMT) and recently neural MT (NMT). However, most of these methods rely on the availability of large parallel data for training the MT systems, resources which are not available for the majority of language pairs, and hence current technologies often fall short in their ability to be applied to low-resource languages. Developing MT technologies using relatively small corpora still presents a major challenge for the MT community but also presents opportunities to improve the technology. In addition, many methods for developing MT systems still rely on several natural language processing (NLP) tools to pre-process texts

✉ Chao-Hong Liu
ch.liu@acm.org

Alina Karakanta
akarakanta@fbk.eu

Audrey N. Tong
audrey.tong@nist.gov

Oleg Aulov
oleg.aulov@nist.gov

Ian M. Soboroff
ian.soboroff@nist.gov

Jonathan Washington
jonathan.washington@swarthmore.edu

Xiaobing Zhao
13522842727@163.com

¹ Potamu Research, Dublin 24, Ireland

² Centro Materiali e Microsistemi, Fondazione Bruno Kessler, Trento, Italy

³ National Institute of Standards and Technology, Gaithersburg, USA

⁴ Swarthmore College, Swarthmore, USA

⁵ Minzu University of China, Beijing, China

in source languages and post-process MT outputs in target languages. The performance of these tools often has a great impact on the quality of the resulting translation. The availability of MT technologies and NLP tools can facilitate equal access to information that may be otherwise unavailable for monolingual speakers.

In recent years, several workshops and evaluations have been organized to promote research on low-resource languages. NIST conducted Low Resource Human Language Technology evaluations (LoReHLT) annually from 2016 to 2019. In LoReHLT evaluations, there is no training data in the evaluation language. Participants receive training data in related languages and need to bootstrap their systems to handle the surprise evaluation language at the start of the evaluation. Methods for this include cross-language transfer and taking advantage of linguistic universals. The evaluations were supported by DARPA's Low Resource Languages for Emergent Incidents (LORELEI) program, which sought to advance technologies that were less dependent on large data resources and that could be quickly pivoted to new languages within a very short amount of time so that information from any language could be extracted in a timely manner to provide situation awareness to emergent incidents. There are also the Workshop on Technologies for MT of Low-Resource Languages (LoResMT) and the Workshop on Deep Learning Approaches for Low-Resource Natural Language Processing (DeepLo), which provide a venue for sharing research and working on the research and development in this field.

This special issue solicited original research papers on MT systems/methods and related NLP tools for low-resource languages in general. We were pleased to receive 42 submissions with diverse topics in the area of MT for low-resource languages. They can be broadly categorized in four main areas.

- Specific (low-resource) languages (18 submissions)
- Corpora/Evaluations (6)
- Methods/Platforms not for specific languages (12)
- Methods incorporating linguistic information (6)

Due to the tremendous work that has been done by diverse research teams from all over the world, we have decided to split the special issue into three parts. The acceptance rate of the special issue was around 35%. This issue of the MT Journal is the first part, which archives five papers working on methodologies of MT for low-resource languages.

We would like to express our sincere gratitude to the many researchers and colleagues in Springer who participated as authors, editors and reviewers and made the special issue successful. For the first part of the special issue, they are Alberto Ponce-las, Atul Kr. Ojha, Chao-Hong Liu, Guodong Xie, Longyue Wang, Jeyasri Varadarajan, Jolanda Voogd, Kokila Kothandaraman, Oleg Aulov, Özgür Özdemir, Peyman Pass-ban, Rejwanul Haque, as well as colleagues who helped during the course. We also like to thank Prof. Andy Way and especially our Springer colleague Jeyasri Varadarajan for all the help in the preparing of the special issue. We are deeply grateful for your help to

deliver this special issue to the MT community despite an on-going pandemic. Thank you all very much!

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.