Proceedings of Machine Translation Summit XVIII

https://mtsummit2021.amtaweb.org

Volume 1: MT Research Track

Editors:
Kevin Duh and Francisco Guzmán (Research Track Co-chairs)
Stephen Richardson (General Conference Chair)

Welcome to the 18th biennial conference of the International Association of Machine Translation (IAMT) – MT Summit 2021 Virtual!

Dear MT Colleagues and Friends,

This year's MT Summit is hosted by the Association for Machine Translation in the Americas (AMTA). Every two years, the Summit is hosted on a rotating basis by one of the three sister organizations comprising IAMT: the European Association for Machine Translation (EAMT), the Asian-Pacific Association for Machine Translation (AAMT), and of course, AMTA. While each of these organizations holds its own conferences annually or biennially, the Summit is always held in odd-numbered years, and this year, AMTA is grateful to have that honor.

After a tremendously successful MT Summit XVII held in Dublin in 2019, we anticipated an equally successfully Summit in 2021 given the rapidly accelerating interest in and research and development of neural machine translation (NMT) in both academia and industry. But as you all know, the year 2020 brought a major surprise that no one anticipated. Our biennial AMTA conference, scheduled for the fall of 2020 in Orlando, Florida was transformed into a completely virtual conference after much consternation followed by a great deal of effort. We successfully rescheduled the MT Summit 2021 conference at the same venue for the following year, thinking that it would at least be a "hybrid" conference, but alas, here we are once again with a completely virtual conference. This decision was made late in the game last April when, based on the results of a survey of likely participants, it become obvious that the vast majority would not be attending in person. Recent spikes in the cases of COVID throughout the world have further justified our decision to go completely virtual.

There have been some silver linings to this COVID cloud, however, the main one being that our AMTA 2020 virtual attendance was double that of previous years, and we anticipate that attendance for the virtual Summit will be at least double what it was in Dublin. We are also grateful that once again, we were able to reschedule our intended venue in Orlando, Florida for AMTA 2022. We hope that many of you will join us there in person! And yes, we will still add a virtual component to the conference for those who are yet unable to travel.

But enough of this COVID-related confusion! We are very pleased with the response we have had to our calls for papers, presentations, workshops, tutorials, and exhibitions for MT Summit 2021 and we are sure you'll agree that the program is brimming with relevant, exciting, and useful information, not to mention the many opportunities to view the latest technology demonstrations and opportunities to network with colleagues both old and new from across the MT spectrum. The most unique aspect of these conferences is that they are truly global gatherings of MT researchers, developers, providers, and users. Academics, students, and commercial researchers and developers are able to share their latest results and offerings with colleagues, in addition to receiving and understanding real-world user requirements. Individual MT users, as well as those from language services providers, enterprises, and governments, benefit from updates on leading-edge R&D in machine translation and have a chance to present and discuss their use cases.

At this point, I need to give some serious thanks to many organizations and individuals who have made this conference possible. First, we have received amazing support from our sponsors, for which we are tremendously grateful! Our visionary sponsor, Microsoft, made it possible for the first 150 students to register for the conference at a very significant discount, and those students quickly took advantage of this generous offer. Our Leader-level sponsors, who will be sponsoring our conference tracks, include: Apple, Intento, Lilt, Pangeanic, (RWS) Language Weaver, Systran, Vistatec, and Yandex Cloud. Our Patron-level sponsors are: Amazon (AWS), Facebook AI, Google, Kudo, Lengoo, Logrus Global, Star, and Welocalize. To all these companies we express our most sincere gratitude for their support of MT Summit 2021. Many of them will also give demonstrations of their systems and software during our Technology Exhibition Fair, and we hope that all our attendees will take advantage of this great opportunity to see the very latest commercial offerings and advancements in the world of MT. We are grateful to have three additional exhibitors in the Fair as well: CustomMT, KantanMT, and XTM.

Finally, I need to give special thanks and recognition to the members of our organizing committee, all of whom have worked very hard and given many hours and days of their time, for the most part voluntarily, to make MT Summit 2021 a success. Listing their names and official positions doesn't really seem to be an adequate reflection of their work and sacrifice, but it's the best I can do here, and I trust they know how much their efforts are truly appreciated.

Patti O'Neill-Brown, AMTA VP, Networking chair
Natalia Levitina, AMTA Secretary
Jen Doyon, AMTA Treasurer
Kevin Duh, Research Track Co-chair
Paco Guzman, Research Track Co-chair
Janice Campbell, Users and Providers Track Co-chair
Jay Marciano, Users and Providers Track Co-chair, Workshops and Tutorials Chair
Konstantin Savenkov, Users and Providers Track Co-chair
Alex Yanishevsky, Users and Providers Track Co-chair, Conference Online Platform Chair
Ben Huyck, Government Track Co-chair
Steve La Rocca, Government Track Co-chair
Ray Flournoy, Sponsorships Chair
Kenton Murray, Student Mentoring Chair
Elaine O'Curran, AMTA Counselor, Publications Chair
Alon Lavie, AMTA Consultant
Konstantin Dranch, Communications Chair
Kate Ozerova, Marketing Lead
Darius Hughes, Webmaster

Again, welcome one and all to MT Summit XVIII 2021! I look forward to "seeing" you online and hopefully, too, in person in the future.

Steve Richardson
IAMT President and MT Summit 2021 General Conference Chair

Introduction

The research track at MTSummit 2021 continues the tradition of bringing MT practitioners together from academia, industry and government from around the world.

This year we have a very rich program with 24 papers from a variety of topics. The most popular subject this year is low-resource machine translation, with papers spanning unsupervised MT, bilingual lexicon induction and curriculum learning. In addition, we have many works discussing modeling (e.g. transfer learning, domain adaptation and reinforcement learning); others discussing morphology (e.g. target-side inflection, subword tokenization); domain-specific translation (e.g. user-generated content translation, product-reviews); and papers performing error analyses of modern NMT systems and understanding their limitations. We are also excited about our invited keynote speakers for the research track: Lucia Specia (Imperial College London) will talk about Multimodal Simultaneous MT, while Graham Neubig (Carnegie Mellon University) will discuss Context-aware MT.

We hope that this conference brings many productive exchanges of ideas and sparks future collaborations.

We would like to thank the hard work of individuals that made this happen: the authors, the reviewers, the MT Summit organizing committee. We would also like to thank Michael Denkowski for numerous pieces of advice on organizing the research track.

Sincerely,

Kevin Duh and Francisco Guzmán (Research Track Co-Chairs)

Program Committee

Yuki Arase (Osaka University)
Nguyen Bach (Alibaba US)
Pushpak Bhattacharya (IIT Bombay)
Alexandra Birch (University of Edinburgh)
Marine Carpuat (University of Maryland)
Francisco Casacuberta (UPV)
Daniel Cer (Google Research; UC Berkeley)
Vishrav Chaudhary (Facebook)
Boxing Chen (Alibaba Group)
Colin Cherry (Google)
Raj Dabre (NICT)
Asif Ekbal (IIT Patna)
Akiko Eriguchi (Microsoft)
Angela Fan (Facebook)
Atsushi Fujita (NICT)
Hongyu Gong (Facebook)
Matthias Huck (SAP SE)
Katharina Kann (Univ. of Colorado Boulder)
Rebecca Knowles (NRC)
Philipp Koehn (Johns Hopkins University)
Shankar Kumar (Google)
Anoop Kunchukuttan (Microsoft)
Alon Lavie (Unbabel)
Gurpreet Lehal (PU)
Yang Liu (Tsinghua University)
Kelly Marchisio (Johns Hopkins University)
Daniel Marcu (Amazon)
Josep Maria Crego (Systran)
Benjamin Marie (NICT)
Marianna Martindale (University of Maryland)
Haitao Mi (Ant Group)
Tetsuji Nakagawa (Google)
Toshiaki Nakazawa (The University of Tokyo)
Jan Niehues (Maastricht University)
Vassilina Nikoulina (Naver Labs Europe)
Atul Ojha (National Univ. of Ireland Galway)
Shantipriya Parida (Idiap Research Institute)
Stephan Peitz (Apple Inc.)
Juan Pino (Facebook)
Maja Popovic (ADAPT Centre @ DCU)
Jean Senellart (Systran)
Rico Sennrich (University of Zurich)
Christophe Servan (Qwant)
Patrick Simianer (Lilt)
Matthias Sperber (Apple)
Katsuhito Sudoh (NAIST)
Christoph Tillmann (IBM Research)
Marco Turchi (FBK)
Josef van Genabith (Saarland University)
Yogarshi Vyas (Amazon AI)
Xinyi Wang (Carnegie Mellon University)
Taro Watanabe (NAIST)
Derek F. Wong (University of Macau)
François YVON (CNRS)
Jiajun Zhang (Institute of Automation, CAS)
Bing Zhao (SRI International)

Contents

1 Learning Curricula for Multilingual Neural Machine Translation Training

Gaurav Kumar, Philipp Koehn and Sanjeev Khudanpur

10 Investigating Active Learning in Interactive Neural Machine Translation

Kamal Gupta, Dhanvanth Boppana, Rejwanul Haque, Asif Ekbal and Pushpak Bhattacharyya

23 Crosslingual Embeddings are Essential in UNMT for distant languages: An English to IndoAryan Case Study

Tamali Banerjee, Rudra V Murthy and Pushpak Bhattacharya

35 Neural Machine Translation in Low-Resource Setting: a Case Study in English-Marathi Pair

Aakash Banerjee, Aditya Jain, Shivam Mhaskar, Sourabh Dattatray Deoghare, Aman Sehgal and Pushpak Bhattacharya

48 Transformers for Low-Resource Languages: Is Féidir Linn!

Seamus Lankford, Haithem Alfi and Andy Way

61 The Effect of Domain and Diacritics in Yoruba--English Neural Machine Translation

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya and Cristina España-Bonet

76 Integrating Unsupervised Data Generation into Self-Supervised Neural Machine Translation for Low-Resource Languages

Dana Ruiter, Dietrich Klakow, Josef van Genabith and Cristina España-Bonet

92 Surprise Language Challenge: Developing a Neural Machine Translation System between Pashto and English in Two Months

Alexandra Birch, Barry Haddow, Antonio Valerio Miceli Barone, Jindrich Helcl, Jonas Waldendorf, Felipe Sánchez Martínez, Mikel Forcada, Víctor Sánchez Cartagena, Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Wilker Aziz, Lina Murady, Sevi Sariisik, Peggy van der Kreeft and Kay Macquarrie

103 Like Chalk and Cheese? On the Effects of Translationese in MT Training

Samuel Larkin, Michel Simard and Rebecca Knowles

114 Investigating Softmax Tempering for Training Neural Machine Translation Models

Raj Dabre and Atsushi Fujita

127 Scrambled Translation Problem: A Problem of Denoising UNMT

Tamali Banerjee, Rudra V Murthy and Pushpak Bhattacharya

139 Make the Blind Translator See The World: A Novel Transfer Learning Solution for Multimodal Machine Translation

Minghan Wang, Jiaxin Guo, Yimeng Chen, Chang Su, Min Zhang, Shimin Tao and Hao Yang

150 Sentiment Preservation in Review Translation using Curriculum-based Re-inforcement Framework

Divya Kumari, Soumya Chennabasavaraj, Nikesh Garera and Asif Ekbal

163 On nature and causes of observed MT errors

Maja Popovic

176 A Comparison of Sentence-Weighting Techniques for NMT

Simon Rieß, Matthias Huck and Alex Fraser

188 Sentiment-based Candidate Selection for NMT

Alexander G Jones and Derry Wijaya

202 Studying The Impact Of Document-level Context On Simultaneous Neural Machine Translation

Raj Dabre, Aizhan Imankulova and Masahiro Kaneko

215 Attainable Text-to-Text Machine Translation vs. Translation: Issues Beyond Linguistic Processing

Atsushi Fujita

231 Modeling Target-side Inflection in Placeholder Translation

Ryokan Ri, Toshiaki Nakazawa and Yoshimasa Tsuruoka

243 Product Review Translation using Phrase Replacement and Attention Guided Noise Augmentation

Kamal Gupta, Soumya Chennabasavaraj, Nikesh Garera and Asif Ekbal

256 Optimizing Word Alignments with Better Subword Tokenization

Anh Khoa Ngo Ho and François Yvon

270 Introducing Mouse Actions into Interactive-Predictive Neural Machine Translation

Ángel Navarro and Francisco Casacuberta

282 Neural Machine Translation with Inflected Lexicon

Artur Nowakowski and Krzysztof Jassem

293 An Alignment-Based Approach to Semi-Supervised Bilingual Lexicon Induction with Small Parallel Corpora

Kelly V Marchisio, Philipp Koehn and Conghao Xiong