WAT 2023

MTS Machine Translation Summit 2023

September 4-8, 2023 Macau SAR, China

Proceedings of the 10th Workshop on Asian Translation (WAT2023)

September 4, 2023

Preface

Many Asian countries are rapidly growing these days and the importance of communicating and exchanging the information with these countries has intensified. To satisfy the demand for communication among these countries, machine translation technology is essential.

Machine translation technology has rapidly evolved recently and it is seeing practical use especially between European languages. However, the translation quality of Asian languages is not that high compared to that of European languages, and machine translation technology for these languages has not reached a stage of proliferation yet. This is not only due to the lack of the language resources for Asian languages but also due to the lack of techniques to correctly transfer the meaning of sentences from/to Asian languages. Consequently, a place for gathering and sharing the resources and knowledge about Asian language translation is necessary to enhance machine translation research for Asian languages.

The Conference on Machine Translation (WMT), the world's largest machine translation workshop, mainly targets on European language. The International Workshop on Spoken Language Translation (IWSLT) has spoken language translation tasks for some Asian languages using TED talk data, but there is no task for written language. The Workshop on Asian Translation (WAT) is an open machine translation evaluation campaign focusing on Asian languages. WAT gathers and shares the resources and knowledge of Asian language translation to understand the problems to be solved for the practical use of machine translation technologies among all Asian countries. WAT is unique in that it is an "open innovation platform": the test data is fixed and open, so participants can repeat evaluations on the same data and confirm changes in translation accuracy over time. WAT has no deadline for the automatic translation quality evaluation (continuous evaluation), so participants can submit translation results at any time.

Following the success of the previous WAT workshops (WAT2014 – WAT2022), WAT2023 will bring together machine translation researchers and users to try, evaluate, share and discuss brand-new ideas about machine translation. For the 10th WAT, we have a Restricted Translation task, Parallel Corpus Filtering task, Multimodal translation tasks, Document-level translation tasks, Indic translation tasks, NICT-SAP tasks, Patent translation tasks, and Non-repetitive Translation task. We had 2 teams participate in the shared tasks. About 40 translation results were submitted to the automatic evaluation server, and selected submissions were manually evaluated. In addition to the shared tasks, WAT2023 also features research papers on topics related to machine translation, especially for Asian languages. The program committee accepted 1 research papers.

We would like to thank all the authors who submitted papers. We also thank the MT-Summit 2023 organizers for their help with administrative matters.

WAT 2023 Organizers

Organizing Committee:

Toshiaki Nakazawa, The University of Tokyo, Japan

Isao Goto, Japan Broadcasting Corporation (NHK), Japan

Hideya Mino, Japan Broadcasting Corporation (NHK), Japan

Kazutaka Kinugawa, Japan Broadcasting Corporation (NHK), Japan

Chenchen Ding, National Institute of Information and Communications Technology (NICT), Japan

Raj Dabre, National Institute of Information and Communications Technology (NICT), Japan

Anoop Kunchukuttan, Microsoft AI and Research, India

Shohei Higashiyama, National Institute of Information and Communications Technology (NICT), Japan

Hiroshi Manabe, National Institute of Information and Communications Technology (NICT), Japan

Shantipriya Parida, Silo AI, Finland

Ondřej Bojar, Charles University, Czech Republic

Chenhui Chu, Kyoto University, Japan

Akiko Eriguchi, Microsoft, USA

Kaori Abe, Tohoku University, Japan

Yusuke Oda, Inspired Cognition, Japan

Makoto Morishita, NTT, Japan

Katsuhito Sudoh, Nara Institute of Science and Technology (NAIST), Japan

Sadao Kurohashi, Kyoto University, Japan

Pushpak Bhattacharyya, Indian Institute of Technology Patna (IITP), India

Technical Collaborators:

Luis Fernando D’Haro, Universidad Politécnica de Madrid, Spain
Rafael E. Banchs, Nanyang Technological University, Singapore
Haizhou Li, National University of Singapore, Singapore
Chen Zhang, National University of Singapore, Singapore

Invited talk: Machine Translation at Wikipedia

Santhosh Thottingal

Wikimedia Foundation

Abstract

Wikipedia, the multilingual encyclopedia available in over 320 languages, uses machine translation technology primarily for article translation. The translation process involves an integrated tool that utilizes various machine translation services to provide initial translations, which are then refined by editors before publication. To date, approximately 1.6 million articles have been translated. This presentation aims to introduce a human-in-the-loop product design, highlighting the provision of high-quality rich text translations through text-only machine translation, coupled with manual curation facilitated by human edits. Additionally, we will share insights and analytics pertaining to translation quality and translators. The discussion will encompass the machine translation engines employed, ranging from free and open-source systems to self-hosted services and external paid APIs. Wikipedia at present has machine translation capability to translate across 198 languages. Lastly, we will present the optimization techniques employed to scale machine translation models in order to meet the performance requirements of Wikipedia.

Biography

Santhosh Thottingal is principal engineer at Wikimedia Foundation Language team. He is based in India. At Wikimedia Foundation, he leads machine translation based projects to fill knowledge gaps in various languages. Santhosh also worked on mediawiki internationalization, technologies that help multilingual speakers to read and write content in wikipedia in their languages. Santhosh is also a typeface designer and known for his fonts for Malayalam script. He was honoured by the President of India in 2019 for his contributions to the Malayalam language.

Table of Contents

Workshop Program

September 4, 2023 [UTC+8]

14:00–14:05 **Welcome**

Overview of the 10th Workshop on Asian Translation
Toshiaki Nakazawa, Kazutaka Kinugawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Makoto Morishita, Ondřej Bojar, Akiko Eriguchi, Yusuke Oda, Chenhui Chu and Sadao Kurohashi

14:05–14:50 **Invited Talk**

Machine Translation at Wikipedia
Santhosh Thottingal

14:50–15:10 **Research Paper**

Mitigating Domain Mismatch in Machine Translation via Paraphrasing
Hyuga Koretaka, Tomoyuki Kajiwara, Atsushi Fujita and Takashi Ninomiya

15:10–16:05 **Shared Task**

Task Descriptions and Results (Hindi/Malayalam/Bengali Multimodal)
Shantipriya Parida

BITS-P at WAT 2023: Improving Indic Language Multimodal Translation by Image Augmentation using Diffusion Models
Amulya Dash, Hrithik Raj Gupta and Yashvardhan Sharma

OdiaGenAI's Participation at WAT2023
SK Shahid, Guneet Singh Kohli, Sambit Sekhar, Debasish Dhal, Adit Sharma, Shubhendra Khusawash, Shantipriya Parida, Stig-Arne Grönroos and Satya Ranjan Dash

September 4, 2023 [UTC+8] (continued)

16:05–16:10 Closing