MTEE : Open Machine Translation Platform for Estonian Government

Toms Bergmanis, Mārcis Pinnis, Roberts Rozis, Jānis Šlapiņš, Valters Šics,
Berta Bernāne, Guntars Pužulis, Endijs Titomers
Tilde, Latvia {name.surname}@tilde.lv
Andre Tättar, Taido Purason, Hele-Andra Kuulmets, Agnes Luhtaru, Liisa Rätsep,
Maali Tars, Annika Laumets-Tättar, Mark Fishel
University of Tartu, Estonia {name.surname}@ut.ee

Abstract

We present the MTEE project—a research initiative funded via an Estonian public procurement to develop machine translation technology that is open-source and free of charge. The MTEE project delivered an open-source platform serving state-of-the-art machine translation systems supporting four domains for six language pairs translating from Estonian into English, German, and Russian and vice-versa. The platform also features grammatical error correction and speech translation for Estonian and allows for formatted document translation and automatic domain detection. The software, data and training workflows for machine translation engines are all made publicly available for further use and research.

1 Project Background

MTEE is an Estonian governmental project to develop high-quality machine translation (MT) platform that is open-source and free of charge. The project was motivated by the COVID-19 pandemic. It was aimed to address the country's need for fast and cheap translation of information to and from Estonian and the languages most relevant to Estonia's society: English, German, and Russian. MTEE was funded by the Ministry of Education and Research via a public procurement through the Language Technology Competence Center at the Institute of the Estonian Language. The duration of MTEE project was nine months, and it concluded in January 2022. It was fulfilled as a collaboration between Tilde and the Institute of Computer Science of the University of Tartu. A demonstration of the platform[1] is made publicly available by hosting using the infrastructure of the High Performance Computing Center of the University of Tartu.

2 Data

To train MT systems, we used parallel data from OPUS (Tiedemann, 2009), ELRC-SHARE (Piperidis et al., 2018) and EU Open Data Portal,[2] as well as data donors and industry partners. In contrast, monolingual data were mainly obtained from the public web. To classify data as belonging to legal, military, crisis, or general domains, we used its source information. Furthermore, we used terminology provided by the Institute of the Estonian Language to automatically obtain additional data for individual domains. The resulting data sets ranged from 5 to 20 million parallel sentences for the general domain. However, data sets were much smaller for niche domains and language pairs, such as the German–Estonian crisis domain, where only a few dozen sentence pairs were identified. We observed a similar pattern for the monolingual data, for which data sizes ranged from 50 million sentences for the general domain to only 8 thousand sentences for the Russian military domain.

We used random held-out subsets of training data for testing and development, which, depending on the language pair and domain, were 500 to 2000 sentences large. Held-out subsets, however, are part of pre-existing parallel corpora, which

[1] https://mt.cs.ut.ee/
[2] https://data.europa.eu/

may be present in training data of other (also third party) MT systems, which would make a fair comparison of the MT system quality impossible. For this reason, we also created entirely novel translation benchmarks[3] by ordering professional translations of recent news.

3 Models

Following the implementation by Lyu et al. (2020), we trained modular multilingual transformer-based models (Vaswani et al., 2017) using fairseq (Ott et al., 2019) with separate encoders and decoders for each input and output language. We selected this architecture because it showed better results for lower-resourced language pairs and domains. The final set of models was trained on a combination of parallel and back-translated data and fine-tuned for each domain.

To evaluate MTEE MT systems, we compared them against the public systems by Tilde, Google, DeepL and Neurotõlge.[4] The evaluation using the newly created translation benchmarks yielded results[5] on average favouring MTEE systems for all domains. These results suggest that, at least as these tests can tell, MTEE systems are competitive and of high quality.

4 Platform

The MTEE platform serves the MT systems and provides functionality for text, document (.docx, .xlsx, .odt, .tmx, .pptx, .txt), and web page translation for all domains and language pairs. Before the translation request is routed to the corresponding MT model, adherence to one of the four domains is automatically detected using a fine-tuned XLM-RoBERTa (Conneau et al., 2020) language model. For translation directions where Estonian is the source language, the platform also provides hints for grammatical error correction[6] and speech translation via a cascade of automatic speech recognition[7] followed by an MT system. These components can be accessed through the translation website or their REST APIs. All components developed for the platform are dockerized and released under the MIT license.[8]

5 Current Status of MTEE

The MTEE project concluded in January 2022, and its results were handed over to the Language Technology Competence Center at the Institute of the Estonian Language.

The High Performance Computing Center of the University of Tartu is hosting the MTEE platform's demonstration for at least another year. Tilde and the Institute of Computer Science of the University of Tartu also continue to provide their technical and scientific support during this period.

Ultimately, when the Institute of the Estonian Language has approbated the technical and scientific results of the project, they should possess the knowledge and the know-how to extend and maintain the platform independently.

References

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL 2020*, pages 8440–8451.

Lyu, Sungwon, Bokyung Son, Kichang Yang, and Jaekyoung Bae. 2020. Revisiting Modularized Multilingual NMT to Meet Industrial Demands. In *Proceedings of EMNLP 2020*, pages 5905–5918, November.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL 2019 (Demonstrations)*, pages 48–53.

Piperidis, Stelios, Penny Labropoulou, Miltos Deligiannis, and Maria Giagkou. 2018. Managing Public Sector Data for Multilingual Applications Development. In *Proceedings of LREC 2018*, pages 1289–1293.

Tiedemann, Jörg. 2009. News from OPUS-A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30.

[3] https://github.com/Project-MTee/MTee_translation_benchmarks

[4] https://www.neurotolge.ee

[5] https://raw.githubusercontent.com/wiki/Project-MTee/mtee-platform/WP3.pdf

[6] https://github.com/tartunlp/grammar-api/pkgs/container/grammar-api

[7] https://github.com/tartunlp/speech-to-text-api/pkgs/container/speech-to-text-api

[8] https://github.com/orgs/Project-MTee/packages