MATEO: MAchine Translation Evaluation Online

Bram Vanroy, Arda Tezcan, Lieve Macken
LT[3], Language and Translation Technology Team
Ghent University
Belgium
{firstname.lastname}@ugent.be

Abstract

We present MAchine Translation Evaluation Online (MATEO), a project that aims to facilitate machine translation (MT) evaluation by means of an easy-to-use web interface that can evaluate given machine translations with a battery of automatic metrics. It caters to both experienced and novice users who are working with MT, such as MT system builders, teachers and students of (machine) translation, and researchers.

1 Introduction

Due to the swift development of evaluation metrics for machine translation (MT) and the absence of up-to-date and user-friendly interfaces, this project aims to bridge the gap by joining together a diverse set of automatic, reference-based MT evaluation metrics, including both established and cutting-edge methods, into a single, easily accessible web interface. It is intended for researchers and practitioners in the Social Sciences and Humanities (SSH) and beyond, also including MT developers and researchers, translation scholars, and experts in the fields of digital humanities and (computational) social sciences. Furthermore, the tool can serve as an instructional resource for educators and students because it emphasises the importance of evaluating language resources. It improves the digital literacy of users: being able to easily evaluate machine-generated translations should make users aware not to blindly use MT systems but critically evaluate them for a task, topic, or domain at hand.

MATEO's web interface is open-source,[1] GPLv3 licensed, and will be hosted at CLARIN.eu infrastructure.

This project was kick-started with a Sponsorship 2021 grant from the European Association of Machine Translation. A substantial follow-up grant was acquired from the CLARIN.eu Bridging Gaps initiative. The secured funding accounts for half-time employment for one year at Ghent University for the first author of this paper, who is the developer of this project. The project will end at the end of June 2023.

2 Related Platforms

Similar platforms exist but they are either not maintained or do not provide all the functionality that we are interested in providing. In the past we made use of Asiya Online[2] for teaching MT classes, which provided similar functionality as we are aiming for but unfortunately the service does not work anymore. It also does not support more recent metrics which we would like to include. Tilde MT also provides an interface to evaluate MT but it is limited to BLEU.[3] MT-ComparEval (Klejch et al., 2015) is an open-source tool that is similar to our plans but it is rather dated when compared to the current, rapidly evolving landscape of MT evaluation metrics by only providing BLEU, precision, recall and F-scores.[4] Finally, MutNMT provides an interface to train MT systems in an educational setting but its evaluation methods are limited to BLEU, TER and ChrF.[5]

[1] https://github.com/BramVanroy/mateo-demo
[2] https://asiya.cs.upc.edu/demo/asiya_online.php
[3] https://www.letsmt.eu/Bleu.aspx
[4] https://ufal.mff.cuni.cz/tools/mt-compareval
[5] https://github.com/Prompsit/mutnmt

3 Progress

MATEO is currently in active development. Below we describe the work that has been done and which next steps are planned. At the time of writing the beta version of the tool is available[6], which will change considerably in the coming months after submitting this paper. The final version will be delivered in time for the EAMT conference.

Underlying the interface, the tool currently makes use of a general purpose evaluation framework "evaluate" by Hugging Face for evaluating given machine translations.[7] As part of the MATEO project, more MT evaluation metrics were added to that framework: NIST (Doddington, 2002), TER (Snover et al., 2006), ChrF (Popović, 2017), CharacTER, CharCut. Other metrics such as BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), BERTScore (Zhang et al., 2020) were already present in the library and are included in MATEO.

In terms of the web interface, we have created a Streamlit[8] website that contains information about the project, the metrics and supported languages, and that allows users to translate and evaluate single-sentence, multi-system machine translations. The translation engine is Facebook's M2M model which we included so that users have access to an open-source multilingual baseline system without having to open other translation services. In terms of evaluation, this first version supports SacreBLEU metrics (BLEU, ChrF, TER) BERTScore, BLEURT, COMET. Other metrics, as mentioned above, may be added for the final version. Users get a bar-chart visualization of the evaluation scores of multiple systems and can download the results as an Excel file.

The first version of the tool was used in classes on MT at Ghent University. Students used MATEO for assignments to improve their MT (evaluation) literacy in late December/early January. We discuss findings from their work in (Macken et al., 2023). They were also asked to give feedback about the usability of the tool which we will analyze in detail and incorporate in new versions of the tool.

To complete the project, we have improvements planned, some of which are inspired by the students' feedback. Importantly, file uploads for system-wide evaluations will be enabled, and the translation and evaluation components will be separated. The translation engine will be replaced by a more up-to-date model. Also, the results of the WMT22 Metrics Shared task (Freitag et al., 2022) will be evaluated and promising metrics will be added to "evaluate" (if they are open-source), and ultimately also to the MATEO interface. Visualizations for edit operations as well as and different export options will also be included in the interface. Finally, the tool will be hosted on CLARIN.eu infrastructure.

[6] https://lt3.ugent.be/mateo/

[7] https://github.com/huggingface/evaluate

[8] https://streamlit.io/

References

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of the 2nd HLT*, HLT '02, pages 138–145, San Francisco, CA, USA, March. Morgan Kaufmann Publishers Inc.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chikiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task. In *Proc. of the 7th WMT*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December. ACL.

Klejch, Ondřej, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. MT-ComparEval. *The Prague Bulletin of Math. Ling.*, 104(1):63–74, October.

Macken, Lieve, Bram Vanroy, and Arda Tezcan. 2023. Adapting machine translation education to the neural era: A case study of MT quality assessment. In *Proc. of the 24th EAMT*, Tampere, Finland, June. EAMT.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU. In *Proc. of the 40th ACL*, ACL '02, pages 311–318, USA, July. ACL.

Popović, Maja. 2017. chrF++. In *Proc. of the 2nd Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. ACL.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET. In *Proc. of EMNLP 2020*, pages 2685–2702, Online, November. ACL.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT. In *Proc. of the 58th ACL*, pages 7881–7892, Online, July. ACL.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA 2006*.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore. In *Proc. of ICLR 2020*, pages 1–43.