The MAKE-NMTViz Project: Meaningful, Accurate and Knowledge-limited Explanations of NMT Systems for Translators

Gabriela Gonzalez-Saez, Fabien Lopez, Mariam Nakhle, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Caroline Rossi, Didier Schwab, Jun Yang, James Robert Turner, et al.

The MAKE-NMTViz Project: Meaningful, Accurate and Knowledge-limited Explanations of NMT Systems for Translators

Gabriela Gonzalez-Saez[1], **Fabien Lopez**[1], **Mariam Nakhle**[1 5], **Marco Dinarelli**[1], **Emmanuelle Esperança-Rodier**[1], **Sui He**[4], **Caroline Rossi**[2], **Didier Schwab**[1], **Jun Yang**[4], **James Robert Turner**[4], **Nicolas Ballier**[3]

1 Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG 38000 Grenoble, France
2 Université Grenoble Alpes
3 Université Paris Cité, LLF & CLILLAC-ARP, 75013 Paris, France
4 Swansea University
5 Lingua Custodia, France
Contact: gabriela-nicole.gonzalez-saez@univ-grenoble-alpes.fr

Abstract

This paper describes MAKE-NMTViz, a project designed to help translators visualize neural machine translation outputs using explainable artificial intelligence visualization tools initially developed for computer vision.

1 Introduction

In their meta-review Doran et al. (2017) distinguish opaque, interpretable, and comprehensible systems across various fields including computer vision and natural language processing. Neural machine translation (NMT) falls into the category of comprehensible systems, provided that adequate visualisation systems are implemented. They argue that "confidence in an interpretable learning system is a function of the user's ability to understand the machine's input/output mapping behaviour". Following the NIST report on Explainable Artificial Intelligence (XAI) (Phillips et al., 2021), we will explore their four principles of XAI, which we have rearranged to spell out our MAKE paradigm:

Meaningfulness We want our visualisation system to be understandable and comprehensible to the translator, putting meaning back at the heart of the NMT process.

Accuracy: Our system will improve our understanding of the input-process-output mapping. It may also identify aspects that are not visually apparent by utilizing visualizations that demonstrate the NMT system's internal working.

Knowledge Limits: The visualisations will be limited to the NMT system's knowledge, operating only under conditions for which it was designed and when it reaches sufficient confidence in its output, thus revealing to the user the uncertainties inherent in the results. This will be achieved by including different metrics, such as confidence and quality estimation for MT.

Explanation: We hope to provide visual evidence of the workings of the different steps (subtokenisation/encoding/decoding), thus accounting e.g. for some NMT hallucinations on the basis of the frequencies of the subtokens.

In this project, we propose to develop a platform that offers meaningful, accurate, and knowledge-limited explanations of NMT systems for translators. While current neural network visualization primarily focuses on analyzing activation patterns for classification tasks, our project expands its scope to investigate the effects of linguistic structures and neural representations. Through our platform, translators will have the capability to translate, post-edit, and evaluate while simultaneously analyzing and explaining NMT model results, thereby bridging the gap between complex Artificial Intelligence (AI) algorithms and human understanding in translation.

2 Expected Results

We intend to develop and utilize a Python-based system that leverages state-of-the-art tools to provide a comprehensive approach for analyzing input, process, and output in NMT. We draw inspiration from the *seq2seqVis* system (Strobelt et al., 2018), adapting its functionalities to analyze the decision-making process of NMT systems. We incorporate attribution methods for feature importance explanations from the *Inseq* System (Sarti et al., 2023) (e.g. saliency maps), enabling interoperability with FairSeq (Ott et al., 2019) models for analyzing relationships between NMT model components. Furthermore, we integrate

attention weight analysis strategies based on BertViz visualization (Vig, 2019), and other methods that attempt to inspect the model's internal data, thus enabling a comprehensive visualization of the data cycle for NMT. Our platform supports FairSeq models with the addition of unobtrusive probes, called *decorators*, to traceable parts of the NMT architecture, enabling flexible integration across various NMT architectures.

3 Visualiser

The visualizer is composed of four interoperable modules: translation, analysis, post-editing, and evaluation.

The *translation* module is responsible for loading the model and source text, generating the translated text, and storing the internal data of the system's working (e.g., attention weights, sequence generation probabilities). Once the translation is completed, the *analysis* module steps in to explain how the NMT model arrives at the proposed translation. The explanation comprises three parts: input, process and output, guiding the user process the system took to create the translation step-by-step (or token-by-token). The input analysis focuses on its representation and the subtokenization used. The process describes the workings of the model and its interaction with the input, connecting the model's internal data and features to the output. Finally, the output displays in the target language all the different alternatives the model generates and explains why it chose the final proposed translation.

To give the translation control to the translator we incorporate a *post-editing* module, which gives the translator the possibility to use other alternatives proposed by the NMT system, but also to force the generation of specific tokens to update the analysis module. The *evaluation* module complements the post-editing module. A post-edited text serves as the ground truth, enabling evaluation of the initially generated proposal. This reference can be replaced by standard datasets to conduct stand-alone evaluations. The evaluation contextualise the explanation of the analysis module.

4 Conclusion

In this project, we propose to change the perspective of current explainability tools and systems which do not target the translator audience. Instead, we explore the use of these systems in the translation pipeline. While existing systems are too complex to be installed and too experimental to reach the corpus linguistics and translation studies communities, our system intends to be user-centred, for a genuinely human-centred AI, and intends to add functionalities and serve as an all-in-one wrapper. The visualising tool current functionalities can be tested on Hugging Face Spaces.[1]

5 The Funding Body and Consortium

This project emanated from research supported by the MAKE-NMTVIZ project (14 months since November 2023), funded under the 2022 Grenoble-Swansea Centre for AI Call for Proposals/ GoSCAI - Grenoble-Swansea Joint Centre in Human Centred AI and Data Systems (MIAI@Grenoble Alpes (ANR-19-P3IA-0003)). This work was also supported by the CREMA project (**c**oreference **re**solution into **ma**chine translation) funded by the French National Research Agency (ANR), contract number ANR-21-CE23-0021-01. The Consortium comprises AI specialists from GETALP@UGA, linguistic experts Caroline Rossi and Nicolas Ballier (NMT specialists), and Jun Yang, Sui He and James Robert Turner, all from the Swansea Translation and Interpreting Group (STING). Its goal is to develop translation tools that integrate both linguistic and AI perspectives, fostering collaboration between translators and specialists in the field.

References

Doran, D, SC Schulz, and TR Besold. 2017. What does explainable ai really mean? a new conceptualization of perspectives. In *First International Workshop on Comprehensibility and Explanation in AI and ML 2017, CEUR Workshop Proceedings, Vol. 2071*.

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 *NAACL-HLT 2019*, pages 48–53.

Phillips, P. Jonathon, Carina Hahn, Peter Fontana, Amy Yates, Kristen K. Greene, David Broniatowski, and Mark A. Przybocki. 2021. Four principles of explainable artificial intelligence, https://doi.org/10.6028/nist.ir.8312.

Sarti, Gabriele, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proc. of 61st Meeting of the Association for Computational Linguistics*, pages 421–435.

Strobelt, Hendrik, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. 2018. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):353–363.

Vig, Jesse. 2019. A multiscale visualization of attention in the transformer model. In *Proc. of 57th Meeting of the Association for Computational Linguistics*, pages 37–42.

[1] `https://huggingface.co/gabrielanicole`