InDeep × NMT: Empowering Human Translators via Interpretable Neural Machine Translation

Gabriele Sarti and **Arianna Bisazza**
Center for Language and Cognition (CLCG)
University of Groningen, The Netherlands
{g.sarti, a.bisazza}@rug.nl

Abstract

The NWO-funded InDeep project aims to empower users of deep-learning models of text, speech, and music by improving their ability to interact with such models and interpret their behaviors. In the translation domain, we aim at developing new tools and methodologies to improve prediction attribution, error analysis, and controllable generation for neural machine translation systems. These advances will be evaluated through field studies involving professional translators to assess gains in post-editing efficiency and enjoyability.

1 Introduction

In recent years, the widespread adoption of deep learning systems in neural machine translation (NMT) led to substantial performance gains across most language pairs. Consequently, the focus of human professionals gradually shifted towards the post-editing of machine-generated content. Despite the indisputable quality of NMT, the question of why and how these systems can effectively encode and exploit linguistic information stands unanswered. Indeed, NMT systems are intrinsically opaque due to their multi-layered nonlinear architecture. This fact significantly hinders our ability to interpret their behavior (Samek et al., 2019), an essential prerequisite to their application in real-world scenarios requiring accountability and transparency. For this reason, the interpretability of neural models has grown into a prolific field of research, developing multiple approaches aimed at analyzing models' predictions and learned representations (Belinkov et al., 2020).

While most explainable NMT studies focus on analyzing model learning and predictive behaviors to gain theoretical insights, interpretability approaches have seldom been applied from a user-centric perspective. This criticality was highlighted by exponents of the interpretability field, among which the necessity of grounding future research in practical applications found broad consensus (Doshi-Velez and Kim, 2017). In light of this, the development of methods that are *self-contained, generalizable, and scalable* would enable the identification of widespread issues characterizing NMT predictions such as hallucinations (Raunak et al., 2021), under- and over-translation, and inadequate terminology (Vamvas and Sennrich, 2021; Vamvas and Sennrich, 2022).

2 Project Description

As part of the broader consortium 'InDeep: Interpreting Deep Learning Models for Text and Sound' funded by the Dutch Research Council (NWO)[1], we aim to build upon the latest advances in interpretability studies to empower end-users of NMT via the application of interpretability techniques for neural machine translation. The InDeep project will run from 2021 to 2026, involving a number of academic and industrial partners such as the universities of Groningen and Amsterdam, KPN, Deloitte and Hugging Face. Central to this project is improving the subjective post-editing experience for human professionals, promoting a shift from a passive proofreading routine to an active role in the translation process by employing interactive and intelligible computational practices, driv-

[1]Find more details at https://interpretingdl.github.io and https://www.nwo.nl/en/projects/nwa129219399

ing further enhancements in the quality and efficiency of post-editing in real-world scenarios. On the methodological side, this entails developing and adapting tools and methodologies to improve prediction attribution, error analysis, and controllable generation for NMT systems. We will evaluate our approaches using automatic metrics, and via a field study surveying professionals in collaboration with GlobalTextware.[2]

The focus for the first part of the project will be on identifying approaches that could be generalized to conditional text generation tasks (Alvarez-Melis and Jaakkola, 2017). *Feature* and *instance attribution* methods let us establish the importance of input components and training examples, respectively, in driving model predictions. These techniques are interesting due to their practical applicability in standard translation workflows. In particular, we find it essential to assess the relationship between importance scores produced by these methods and different categories of translation errors. Evaluating the *faithfulness* for model attributions, i.e., how they are causally linked to the system's outputs, is another fundamental component of our investigation and will be pursued by employing a mix of existing and new techniques (DeYoung et al., 2020).

The second part of the project will involve a field study combining behavioral and subjective quality metrics to empirically estimate the effectiveness of our methods in real-world scenarios. For the behavioral part, we intend to use a combination of keylogging and possibly eye-tracking and mouse-tracking to collect granular information about the post-editing process. Our analysis will benefit from insights from recent interactive NMT studies (Santy et al., 2019; Coppers et al., 2018; Vandeghinste et al., 2019) to present translators with useful information while avoiding visual clutter. Our preliminary inquiry involving professionals highlighted sentence-level quality estimation and adaptive style/terminology constraints as promising directions to increase post-editing productivity and enjoyability, supporting the potential of combining interpretable and interactive modules for NMT.

[2]https://www.globaltextware.nl/

References

Alvarez-Melis, David, and Tommi Jaakkola. 2017. A Causal Framework for Explaining the Predictions of Black-Box Sequence-to-Sequence Models In *Proceedings of EMNLP 2017*, 412–421.

Belinkov, Yonatan, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and Analysis in Neural NLP In *Proceedings of ACL 2020: Tutorials*, 1–5.

Coppers, Sven, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna van der Lek-Ciudin, Tom Vanallemeersch, Vincent Vandeghinste 2018. Intellingo: An Intelligible Translation Environment In *Proceedings of CHI 2018*: 524, 1–13.

DeYoung, Jay, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, Byron C. Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models In *Proceedings of ACL 2020*, 4443–4458.

Doshi-Velez, Finale, and Been Kim. 2018. Considerations for Evaluation and Generalization in Interpretable Machine Learning *Explainable and Interpretable Models in Computer Vision and Machine Learning*, 3–17.

He, Shilin, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael Lyu, and Shuming Shi. 2019. Towards Understanding Neural Machine Translation with Word Importance In *Proceedings of EMNLP-IJCNLP 2019*, 953–962.

Samek, Wojciech, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning *Springer Nature*.

Santy, Sebastin, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. INMT: Interactive Neural Machine Translation Prediction In *Proceedings of EMNLP-IJCNLP 2019*, 103–8.

Vamvas, Jannis and Rico Sennrich. 2021. Contrastive Conditioning for Assessing Disambiguation in MT: A Case Study of Distilled Bias In *Proceedings of EMNLP 2021*, 10246–10265.

Vamvas, Jannis and Rico Sennrich. 2022. As Little as Possible, as Much as Necessary: Detecting Over- and Undertranslations with Contrastive Conditioning In *Proceedings of ACL 2022*, 10139–10155.

Vandeghinste, Vincent, Tom Vanallemeersch, Liesbeth Augustinus, Bram Bulté, Frank Van Eynde, Joris Pelemans, Lyan Verwimp. 2019. Improving the Translation Environment for Professional Translators *Informatics* 6 (2): 24, 1–36.

Vikas Raunak, Arul Menezes, Marcin Junczys-Dowmunt. 2021. The Curious Case of Hallucinations in Neural Machine Translation In *Proceedings of NAACL 2021*, 1172–1183.