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

1 Abstract

As part of the NWO-funded InDeep consortium[1], we aim to build upon the latest advances in explainable AI to empower end-users of neural machine translation via the application of interpretability techniques. 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 translation role while also driving quality and efficiency improvements. On the methodological side, this entails developing 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 interpretability approaches that could be generalized to text generation tasks. *Feature* and *instance attribution* methods evaluate the importance of input components and training examples, respectively, in driving model predictions, and can be applied to standard MT workflows to make them more intelligible. In particular, we find it essential to assess the relationship between importance scores produced by these methods and different translation errors. Evaluating how *faithful* explanations are in causally explaining system's outputs is another fundamental step in our investigation (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. For the behavioral part, we intend to use a combination of keylogging and possibly eye-tracking to collect granular information about the post-editing process. Our analysis will benefit from insights from recent interactive MT studies (Huang et al., 2021; Santy et al., 2019; Coppers et al., 2018) 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.

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

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

References

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.

Huang, Guoping, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. *TranSmart: A Practical Interactive Machine Translation System* Arxiv 2105.13072.

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.

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