From Insights to Impact: Actionable Interpretability for Neural Machine Translation

Gabriele Sarti
Center for Language and Cognition (CLCG), University of Groningen
Oude Kijk in Het Jatstraat 26, 9712 GR Groningen, Netherlands
Supervisors: Arianna Bisazza, Malvina Nissim, Grzegorz Chrupała
gabriele.sarti996@gmail.com

Modern neural machine translation (MT) systems have achieved remarkable quality but remain opaque to users, limiting their trustworthiness and usability in professional settings. While interpretability research has produced valuable insights into language model internals, these findings rarely translate into practical benefits for end users. This dissertation bridges this gap by developing and evaluating interpretability methods end-to-end to improve MT trustworthiness and controllability in real-world applications.

The thesis pursues three interconnected objectives: (1) developing methods to precisely quantify how MT systems and multilingual language models exploit contextual information during generation; (2) creating robust techniques for controlling MT output style and personalizing translations; and (3) integrating interpretability insights into professional translation workflows through evaluation with end users.

Part I: Attributing Context Usage in Multilingual Language Models. We first introduce Inseq (Sarti et al., 2023a, **ACL Demo 2023** and Sarti et al., 2024b, **XAI Demo 2024**), a Python toolkit democratizing access to feature attribution for generative language models. Inseq addresses computational and conceptual challenges through careful design and progressive disclosure of complexity, making attribution analyses accessible across diverse computational budgets and expertise levels. Its widespread adoption in works studying bias in machine translation systems is a testament to Inseq accessibility and impact. Using Inseq, we conduct an investigation on gender stereotypes in MT, finding that contrastive attribution can effectively identify biased responses in these systems. Then, we develop PECoRe (Plausibility Evaluation for Context Reliance, Sarti et al., 2024a, **ICLR 2024**), a framework that quantifies context usage in context-aware MT systems through a two-step process: detecting context-sensitive tokens using information-theoretic metrics, and then applying contrastive attribution with ecologically valid counterfactuals to identify motivating contextual cues. When tested on popular multilingual systems, PECoRe exposed critical weaknesses, including gender-agreement failures traced to incorrect anaphora resolution and occasional overreliance on target-side context cues. We adapt PECoRe for retrieval-augmented generation with MIRAGE (Qi et al., 2024, **EMNLP 2024**), demonstrating how model internals can augment answers of multilingual large language models with faithful source citations in retrieval-augmented generation settings. Unlike post-hoc rationalizations based on lexical similarity, MIRAGE grounds citations in actual context processing, achieving quality comparable to or better than self-citation approaches while ensuring faithfulness with models' inner workings.

Part II: Conditioning Generation for Personalized Machine Translation. We begin by introducing RAMP (Sarti et al., 2023b, **ACL 2023**), a prompting strategy pioneering retrieval-augmented generation for attribute-controlled translation. By leveraging multilingual LLMs with cross-lingual retrieval and explicit attribute marking, Ramp achieves zero-shot attribute control (formality, gender) without model tuning, outperforming standard prompting and specialized MT systems in both adequacy and attribute accuracy.

Building on this, we conduct a comprehensive evaluation of prompting- and interpretability-based

steering techniques for MT personalization, focusing on literary translation where subtle stylistic choices are paramount (Scalena et al., 2026, **EACL 2026**). We propose a novel approach using sparse autoencoder (SAE) latents to identify and manipulate style-relevant latent features from model activations, efficiently capturing distinctive stylistic signatures of individual translators. Our method achieves personalization comparable to few-shot prompting while avoiding inference-time overhead from long contexts. Through probing experiments, we demonstrate that both methods converge on similar internal mechanisms, suggesting that SAEs distill in-context demonstrations of translation style into interpretable latents.

Part III: Interpretability in Human Translation Workflows. We conduct DivEMT, an initial user study with 18 professional translators across six typologically diverse languages, revealing that MT's productivity contribution varies significantly by language pair (Sarti et al., 2022, **EMNLP 2022**). Typological similarity emerges as a key factor for quality: closely related languages show substantial post-editing gains while distant pairs show minimal improvement. Crucially, traditional MT quality metrics such as BLEU and COMET correlate poorly with actual productivity gains observed across languages, challenging assumptions about technical quality and the practical usefulness of these metrics for evaluating post-editing workflows.

Our second user study, QE4PE (Quality Estimation for Post-Editing), investigates how word-level error highlights affect productivity, quality and usability for 42 professional post-editors working across two translation directions (Sarti et al., 2025a, **TACL 2025**). We compare supervised state-of-the-art supervised methods, unsupervised uncertainty-based techniques, and oracle human annotations to assess the impact of error-detection quality on these metrics, finding no meaningful differences between highlight sources in coarse-grained metrics but a 15-20% reduction in critical errors across all highlighted conditions compared to regular post-editing. Our findings suggest that highlights help translators catch mistakes they would otherwise miss, though at the cost of higher cognitive effort.

We conclude with a comprehensive evaluation of unsupervised quality estimation methods for detecting translation errors across 12 translation directions (Sarti et al., 2025b, **EMNLP 2025**). Token log-probability variance estimated with Monte Carlo Dropout proves particularly robust, matching state-of-the-art supervised approaches. Proper calibration dramatically improves performance to near human inter-annotator agreement levels, while analysis reveals that annotators' subjective judgments substantially influence metric rankings.

Broader Impact. This thesis pioneers the systematic assessment of interpretability methods by examining their effects on professional users' productivity, decision-making, and satisfaction, positioning the field of machine translation as a prime context for such an assessment. By demonstrating concrete improvements in professional translation workflows, it provides both a proof of concept and a blueprint for the next generation of transparent and controllable translation systems. Its technical contributions address critical challenges in accessibility and computational cost that typically hinder the production and deployment of these methods, charting a path toward improved human-AI collaboration in translation workflows.

The full dissertation is available at: https://hdl.handle.net/11370/97754e4b-f978-477e-a394-83f67cca257c

Acknowledgements. The author acknowledges the continued support of his supervisors Arianna Bisazza, Malvina Nissim and Grzegorz Chrupała, which was instrumental to the success of this research. Moreover, we would like to acknowledge the helpful discussion and comments of members of thesis and defense committees — Barbara Plank, Ivan Titov, Michael Biehl, Willem Zuidema, Eva Vanmassenhove, Rik van Noord and Tommaso Caselli — and of colleagues at GroNLP and at the InDeep consortium. This work was primarily supported by the Dutch Research Council (NWO) as part of the InDeep project (NWA.1292.19.399), with additional support provided by the Imminent Research Center and the Netherlands eScience Center. We also thank the Center for Information Technology of the University of Groningen and SURF for providing access to the Hábrók and Snellius high-performance computing clusters.

References

Qi, Jirui, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Conference on Empirical Methods in NLP (EMNLP)*.

Sarti, Gabriele, Arianna Bisazza, Ana Guerberof-Arenas, and Antonio Toral. 2022. DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In *Conference on Empirical Methods in NLP (EMNLP)*.

Sarti, Gabriele, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023a. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the Association for Computational Linguistics (ACL): Demos*.

Sarti, Gabriele, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, and Maria Nadejde. 2023b. RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Sarti, Gabriele, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. 2024a. Quantifying the plausibility of context reliance in neural machine translation. In *12th International Conference on Learning Representations (ICLR)*, May.

Sarti, Gabriele, Nils Feldhus, Jirui Qi, Malvina Nissim, and Arianna Bisazza. 2024b. Democratizing advanced attribution analyses of generative language models with the inseq toolkit. In *2nd World Conference on XAI: Demos*. CEUR.org.

Sarti, Gabriele, Vilém Zouhar, Grzegorz Chrupała, Ana Guerberof-Arenas, Malvina Nissim, and Arianna Bisazza. 2025a. QE4PE: Word-level quality estimation for human post-editing. *Transactions of the Association of Computational Linguistics (TACL)*.

Sarti, Gabriele, Vilém Zouhar, Malvina Nissim, and Arianna Bisazza. 2025b. Unsupervised word-level quality estimation for machine translation through the lens of annotators (dis)agreement. *Conference on Empirical Methods in NLP (EMNLP)*.

Scalena, Daniel, Gabriele Sarti, Arianna Bisazza, Elisabetta Fersini, and Malvina Nissim. 2026. Steering large language models for machine translation personalization. *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.