Analyzing and Improving Cross-lingual Knowledge Transfer for Machine Translation

David Stap
dd.stap@gmail.com
Current affiliation: NXAI
Supervisor: prof. dr. Christof Monz (c.monz@uva.nl)
Co-supervisor: dr. Vlad Niculae (v.niculae@uva.nl)

1 Goals and Objectives

This thesis (Stap, 2025) addresses a core challenge in machine translation: building systems that translate between thousands of languages when parallel training data is abundant for only a small fraction of them. Our objective is to advance cross-lingual knowledge transfer—the ability of multilingual models to use representations learned from one language to improve translation for another—and extend machine translation to low-resource languages. We study cross-lingual knowledge transfer from four angles: (1) measuring and enhancing transfer through representational similarities in multilingual neural machine translation (NMT); (2) leveraging transfer in retrieval-augmented translation with k-nearest-neighbor methods; (3) mitigating negative transfer when fine-tuning large language models (LLMs) for translation; and (4) characterizing how language diversity during fine-tuning affects cross-lingual generalization.

2 Methodology

Representational analysis (Chapter 3): We define Representational Transfer Potential (RTP), a metric that quantifies knowledge transfer between languages from similarities in their cross-attention context vectors (Stap et al., 2023). Across a wide range of language pairs, we establish a predictive relationship between representational similarity and translation quality. A regression analysis identifies dataset characteristics (multi-parallel overlap, source subword overlap) and linguistic features (genetic distance) that predict transfer. Building on these findings, we design an auxiliary training objective that uses multi-parallel data to increase representational alignment across languages.

Multilingual k-nearest-neighbor machine translation (Chapter 4): We extend k-NN-MT to multilingual settings by introducing cross-lingual and multilingual datastore architectures (Stap and Monz, 2023). To improve retrieval across languages, we align representations with linear mappings. We then compare bilingual, cross-lingual, and multilingual datastores across language pairs and resource conditions, measuring both translation quality and inference cost.

LLM fine-tuning and capability preservation (Chapter 5): We analyze how fine-tuning on parallel translation data affects the qualitative capabilities of LLMs across model scales (7B to 65B parameters) and data regimes (Stap et al., 2024). The capabilities we examine include formality steering, few-shot domain adaptation, document-level translation, and non-literal translation of idiomatic expressions. Based on these findings, we develop a mixed-data fine-tuning strategy that combines parallel translation data with monolingual text, balancing translation quality against capability preservation.

Language diversity in LLM fine-tuning (Chapter 6): Across 132 controlled translation directions, we vary language diversity during fine-tuning and measure its effect on both seen and unseen language pairs (Stap and Monz, 2025). We then analyze how representations change across model layers, examining how middle layers adapt during fine-tuning and how cross-lingual overlap grows as more languages are seen.

We evaluate with BLEU and COMET, and validate findings across multiple language families and

resource conditions.

3 Results

Quantifying and enhancing representational transfer: RTP scores correlate strongly with translation quality improvements (Spearman's $\rho = .77$, $p < 0.001$), indicating that transfer operates at the representational level. Multi-parallel overlap emerges as a strong yet under-explored predictor of transfer. The auxiliary similarity loss improves low-resource translation while maintaining high-resource performance.

Multilingual *k*-NN-MT: Cross-lingual datastores built from linguistically similar high-resource languages consistently outperform bilingual datastores for low-resource translation. Multilingual datastores benefit both low- and high-resource languages at once. Language-group-specific datastores match the performance of full multilingual datastores while reducing size by up to 75% and improving inference speed by up to 5.3×.

Preserving LLM capabilities during fine-tuning: Fine-tuning on parallel data improves general translation quality but degrades capabilities such as formality control, few-shot domain adaptation, and document-level translation. The degradation appears with as few as 18K fine-tuning examples and is consistent across model scales. Mixed-data fine-tuning preserves these capabilities while achieving higher overall translation quality than parallel-only fine-tuning, resolving the fine-tuning paradox.

Effects of language diversity: Scaling language diversity during fine-tuning improves translation quality for both seen and unseen language pairs, resolving conflicting evidence from prior work. Models fine-tuned on more diverse language sets outperform those trained on narrower sets, even for pairs explicitly included in specialized training. Our analysis shows that broader language coverage produces more cross-lingual overlap between related languages in middle layers, while increasing specialization in language-specific layers.

Taken together, these findings clarify how cross-lingual knowledge transfer operates and yield methods for extending machine translation to under-resourced languages. We release the IdiomsInCtx-MT evaluation dataset, our code, and model checkpoints to support reproducibility and follow-on work. The thesis shows that careful choices in modeling and data extend multilingual NLP beyond well-resourced languages.

Acknowledgements

The work described in this thesis has been carried out at the Language Technology Lab of the University of Amsterdam and in part during an internship at Amazon AGI, Berlin. The research carried out at the University of Amsterdam was funded in part by the Association of Collaborating Dutch Universities (VSNU) under project Onderzoek Digitale Samenleving (DiSa) and in part by the Netherlands Organization for Scientific Research (NWO) under project numbers VI.C.192.080 and 2023.017. We thank SURF (`www.surf.nl`) for the support in using the National Supercomputer Snellius.

References

[Stap and Monz2023] Stap, David and Christof Monz. 2023. Multilingual k-nearest-neighbor machine translation. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9200–9208, Singapore, December. Association for Computational Linguistics.

[Stap and Monz2025] Stap, David and Christof Monz. 2025. The effect of language diversity when fine-tuning large language models for translation. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4199–4211, Suzhou, China, November. Association for Computational Linguistics.

[Stap et al.2023] Stap, David, Vlad Niculae, and Christof Monz. 2023. Viewing knowledge transfer in multilingual machine translation through a representational lens. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14973–14987, Singapore, December. Association for Computational Linguistics.

[Stap et al.2024] Stap, David, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6189–6206, Bangkok, Thailand, August. Association for Computational Linguistics.

[Stap2025] Stap, David. 2025. *Analyzing and Improving Cross-lingual Knowledge Transfer for Machine Translation*. Ph.D. thesis, University of Amsterdam. `https://dare.uva.nl/id/9302c107-d675-496d-8019-60a0e8eda304`.