# Towards Robust and Efficient Multilingual Neural Machine Translation

Name

Zhi Qu

Laboratory's name

Natural Language Processing Laboratory

Supervisor's name

Taro Watanabe

Abstract

Multilingual neural machine translation (MNMT) seeks to support many language pairs within a single model, simplifying deployment and improving cross-lingual generalization. Encoder–decoder (Enc-dec) systems such as M2M-100 and NLLB-200 have driven recent progress, yet the rise of decoder-only (Dec-only) large language models (LLMs) challenges their relevance. Although LLMs can handle multilingual translation, adapting them for MNMT is costly, and direct Dec-only MNMT models still lag behind Enc-dec performance. This gap raises a key question: what enables Enc-dec architectures to excel in MNMT, and can this advantage be transferred to Dec-only models?

This dissertation examines the mechanism behind Enc-dec success and its implications for Dec-only systems. We introduce the concept of an *identity pair* to analyze Enc-dec representations and show that the encoder maps source sentences into target-language subspaces, aligning multiple source languages semantically. Empirical studies confirm that stronger language-transfer capability correlates with better MNMT performance. We then identify the lack of such transfer as a core limitation of Dec-only models and propose **registering**, which inserts artificial tokens to encode source semantics into the target subspace. Registering establishes a new state of the art for MNMT-specific methods and enables Dec-only architectures to surpass Enc-dec models. Based on this approach, we pre-train **MITRE**, a family of Dec-only MNMT models over 9.3B sentence pairs in 24 languages. With fewer than one billion parameters, MITRE outperforms NLLB-3.3B and approaches commercial LLM performance.

Overall, this work offers a theoretical account of language transfer in MNMT and provides practical methods for building efficient, high-performing multilingual translation models.