A SURVEY OF NEURAL MACHINE TRANSLATION ON LOW-RESOURCE LANGUAGE PAIRS

EXPLORING LINGUISTIC JUSTICE AND DATA EQUITY IN AI

An Undergraduate Thesis Portfolio
Presented to the Faculty of the
School of Engineering and Applied Science
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Anusha Choudhary

May 12, 2023

SOCIOTECHNICAL SYNTHESIS

In the field of Machine Learning and Artificial Intelligence, fine-tuning refers to the method of iteratively adjusting several parameters to achieve better performance of a system. Through the State-of-the-Art technical report and the STS research paper presented in this portfolio, we explore the idea of fine-tuning in both the technical field of low-resource machine translation and to the wider societal world surrounding it. In order to achieve higher quality translation in languages for which there exists a lack of training data, several methodological advancements have been made. There exists a need to summarize these advancements. Thus, to benefit the Neural Machine Translation community, we provide a State-of-the-Art review of the most recent methodologies being applied to achieve higher translation quality on multilingual Neural Machine Translation models and analyze the emerging trends within those methods. Moreover, the resulting linguistic injustice caused by the poor quality of low-resource translation tools is previously unexplored from a Science, Technology, and Society perspective. Thus, the STS research paper provides a sociological framework for the technical community to keep linguistic justice at the forefront in the development of future machine translation tools. The problem of lower quality of low-resource machine translation is inherently both a sociological and technical problem; thus, to achieve the best improvements, both sociological and technical fixes must be applied.

Although the problem of low-resource machine translation had been identified as early as 2017, in 2023, low-resource machine translation is a more active field of research than ever before due to the meteoric rise in computational power that supports Neural Machine Translation models. With increased volume of research, there is an increasing need to summarize the methodologies being used to improve low-resource machine translation. The rationale behind the

State-of-the-Art technical report is to provide an updated review for the Neural Machine Translation community, specifically on the improvements being made to multilingual Neural Machine Translation models as multilingual models are better-performing in low-resource conditions.

The literature review done as part of the technical report discovered a general qualitative trend in the literature of moving away from learning universal language-agnostic representations and instead focusing on language families for better low-resource translation quality in multilingual NMT models. The results of the technical report reemphasize that the NMT community is currently focused on making finer low-level improvements to the high-level optimization techniques discovered in the past.

The goal of the STS research paper has been to formulate a framework for engineers building Neural Machine Translation models and tools that promotes linguistic and social justice. In formulating this framework, a preliminary question of why the topic has been previously unexplored was answered. Then, the current state of low-resource machine translation was further explored using the Social Construction of Technology. The major actors wielding power over the development of Neural Machine Translation and the relationships between them were identified using the framework of Actor-Network Theory. The framework of Linguistic Justice, originally formulated for all tasks within Natural Language Processing, was then adapted to the field of Neural Machine Translation.

The final frame of reference formulated using the frameworks of Social Construction of Technology, Actor-Network Theory, and Linguistic Justice focused on two main areas of equity and inclusivity: 1) inclusivity of semantic and syntactic language structure across all languages, and 2) inclusivity of low-resource language speakers in all stages of model development, not only as third-party consultants but as the engineers and researchers developing the machine translation models and tools.

An overarching conclusion that can be drawn from the results of the technical report and the STS research paper is that linguistic justice and better-quality low-resource machine translation are complementary and tightly coupled. Therefore, there exists no need for a tradeoff between the two; i.e., improvements toward linguistic justice inherently promote improvements in machine translation and vice-versa. Thus, it is possible to achieve a socio-technical synergy within the field of low-resource machine translation.

TABLE OF CONTENTS