Molding Models' Mother Tongues: Fine-Tuning for Marathi-English Translation

Sanika Nandpure and Melissa Huang

Project GitHub: https://github.com/sanikanandpure/coe379_project3

I. Introduction and Problem Statement

In an increasingly globalized world, translation has become an integral part of connection and communication. Tools like Google Translate and various AI models have mastered this for common languages, such as Spanish or even Mandarin. However, for some smaller diasporas or communities, these tools are far less ubiquitous. Many state-of-the-art LLMs might demonstrate great general translation capabilities, but fail when addressing languages like Navajo, Yoruba, or Marathi, where pretraining resources and corpora are sparse.

Thus, this project aims to investigate the performance of a fine-tuned English-based model on translations to a lesser-known language, specifically Marathi, with its unique writing style that does not rely on the Latin or Roman alphabet. The objective is to enable the model to produce accurate translations from Marathi to English without having to pretrain an entirely new model, which would be computationally and time-expensive.

We chose Marathi for a few reasons. First, a co-author in this project is fluent in Marathi, allowing us some baseline knowledge of the language in case manual inspections are ever needed. Secondly, we wanted to leverage a lesser-known dialect/language, but needed to balance the amount of proper and cohesive datasets or resources available. We originally wished to conduct our research on indigenous American or African languages, but could not find enough datasets to satisfy the fine-tuning of the model, nor would we be able to verify their accuracy effectively.

II. Data Sources

For our project, we will be using this dataset from the Hugging Face Hub. It provides around three million entries of simple English-Marathi translations of short sentences/phrases from various sources. This provides enough samples to pull from for both the fine-tuning and testing phases. Upon an initial inspection of the dataset, it seems to sample quotes from fictional stories (possible television shows or novels), first-person dialogue, news reports, and other online media, providing a large variety of quotations.

III. High-Level Methods

We will first split the English-Marathi dataset above into test, training, and validation sets. Our approach uses supervised fine-tuning, applying LoRA (Low-Rank Adaptation) to the base Llama 3 model. Our dataset consists of paired English and Marathi phrases/sentences, which the model will learn to translate during training (Marathi → English).

After fine-tuning on the training set, we will evaluate translation quality using the BLEU (Bilingual Evaluation Understudy) metric, which evaluates the quality of machine-generated

language translation by comparing against human-created/ground-truth translations. We will compare the BLEU scores of the base model (without fine-tuning) and the model fine-tuned with LoRA to determine whether language understanding can be instilled into a model via supervised fine-tuning.

To save and present the final fine-tuned model, we will publish our LoRA adapters to HuggingFaceHub.

IV. Products To Be Delivered

The primary deliverable would be the fine-tuned LLM on the English-Marathi dataset. We hope to be able to deliver a model that can robustly handle subsequent translations from Marathi to English, which we aim to have deployed/published to be able to handle external requests or accesses.

Alongside the model, we will be tracking performance statistics through the aforementioned methods, and we hope to provide documentation for our test prompts, the translated results, and the BLEU scores before and after the fine-tuning. If time allows, we can even try to have a front-facing UI to interact with the translator, but this is dependent on how well the model actually performs and is not guaranteed.