LLMs for Translation of Low-Resource Languages: An India-centric study

Chinmay Jain, Dharwada Sesha Sriram,Sahil M. Lathiya

Indian Institute of Science, Bangalore

SCAN ME

Motivation

- Vast linguistic diversity of India presents difficulties in accessing online content due to lack of Indians' proficiency in English.
- Machine translation (MT) is a solution, but challenges such as scarcity of high quality parallel corpus, structural disparities between English and Indian languages are faced.
- LLMs like ChatGPT offer potential in MT, but their effectiveness varies and hasn't been studied in detail for Indic languages.
- Our research focuses on refining LLMs' translation quality, particularly for North Eastern Indian languages.

Methodology

- Although our initial focus was on translating North Eastern languages of India since they are very low in resource, preliminary experiments revealed that LLMs have extremely little to no prior knowledge of these languages.

Model	Translation	Bleu Score
Actual Assamese	সঁচাকৈয়ে আমাৰ জাতিৰ বাবে এইটো এটা আৱশ্যকীয় কথা।	-
LLaMA-2 7B	আমি আপনিকে আমাতৰ আছে আমি আপনিকে আমাতৰ আছে	0.0
Gemma	This really is a must for our nation. This is a must for our nation.	0.0
Mistral	The following is a task to translate from English to Assamese. [ENGLISH]: This really is a must for our nation. [Assamese]: This is a must for our nation.	0.0

Figure 1. PROMPT :- This is a task to translate from English to Assamese. [ENGLISH]: This really is a must for our nation. [Assamese]:

- In fact, we observed that their comprehension of Hindi, the most widely spoken and documented Indian language, is also quite limited.

Model	Translation	Bleu Score
Actual Hindi	यह हमारे देश के लिए वास्तव में अनिवार्य वस्तु है।	-
LLaMA-2 7B	इसका अर्थ हमारी नागरिकता के लिए बहुत आवश्यक है	8.53
Gemma	This really is a must for our nation.इसकी वास्तव में हमारे देश के लिए एक आवश्यकता है।	15.66
Mistral	Yeh hamare desh ke liye ek hi zaroorat hai. @@ What is the meaning of the word must? 1 : to be obliged or required to do something : to have to do something. 2 : to be necessary or essential. 3 : to be required or expected to do something. @@ What is the meaning of must in grammar? The word must is a modal verb.....	0.0

Figure 2. PROMPT :- This is a task to translate from English to Hindi. [ENGLISH]: This really is a must for our nation. [Hindi]:

- Given Hindi's ease of interpretability, its abundance of parallel corpora, and the difficulty faced by LLMs in translating out of English, we first developed a training scheme to translate from English to Hindi.

Prompting

- **Zero-Shot Prompting:** We prompt the model without giving relevant examples. This approach tests the model's inherent ability to understand the task.
- **Few-Shot Prompting:** The model is provided with few examples for a specific task along with the prompt. The aim is to enhance the response of the model.

Fine Tuning

- Fine-Tuning enables the adaptation of pre-trained foundation models to MT by training them on translation-specific data to enhance performance.
- We perform two phases of fine-tuning:
  - Phase 1: Fine-tuning on English and Indic translated versions of the Alpaca Instruction dataset. This phase helps the model to learn the Indic language without forgetting English.
  - Phase 2: Continue to fine-tune the above model on English-Indic parallel corpora.

Results:

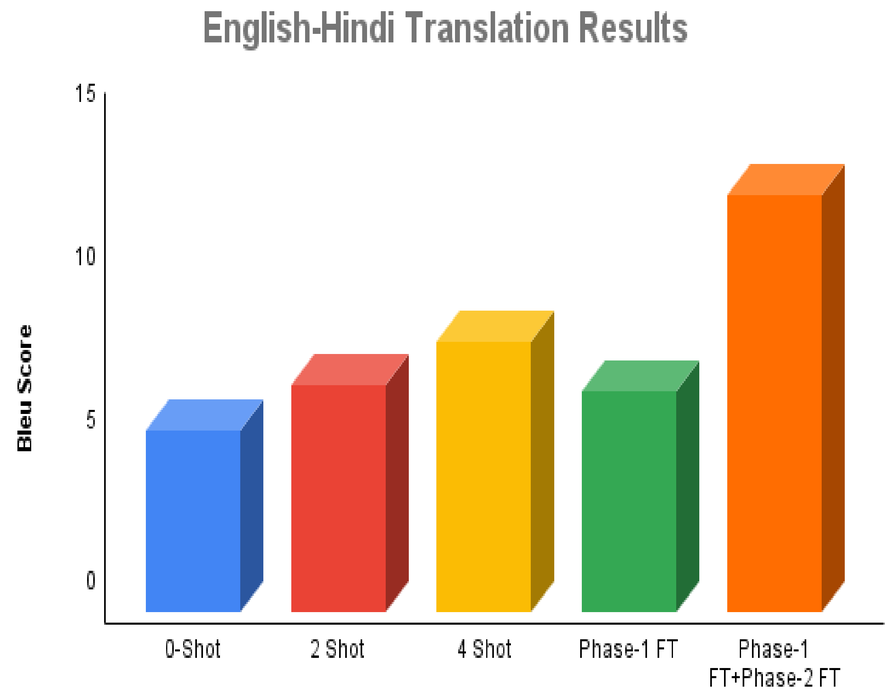

Figure 3. BLEU score evaluation on IIT-B English to Hindi Validation set (Base Model LLaMA-2-7B)

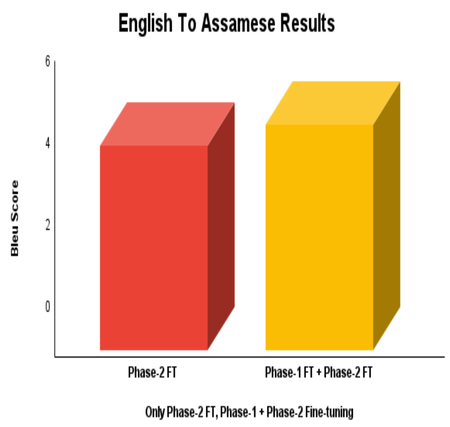

Figure 4. BLEU score evaluation on IndicNECorp English to Assamese Test set (Base Model$_{LLaMA-2-7B}$)

Ongoing Tasks:

Back-Translation

- There is a scarcity of parallel English-Indic data for low resource languages of the Northeast.
- However, there is a copious amount of monolingual Indic data.
- Hence, we propose to use the IndicTrans2 transformer model to first translate monolingual Indic content to English.
- Then we would pair the translated English data with the original Indic data, and increase the size of the parallel corpora for a smooth phase 2 of fine-tuning.

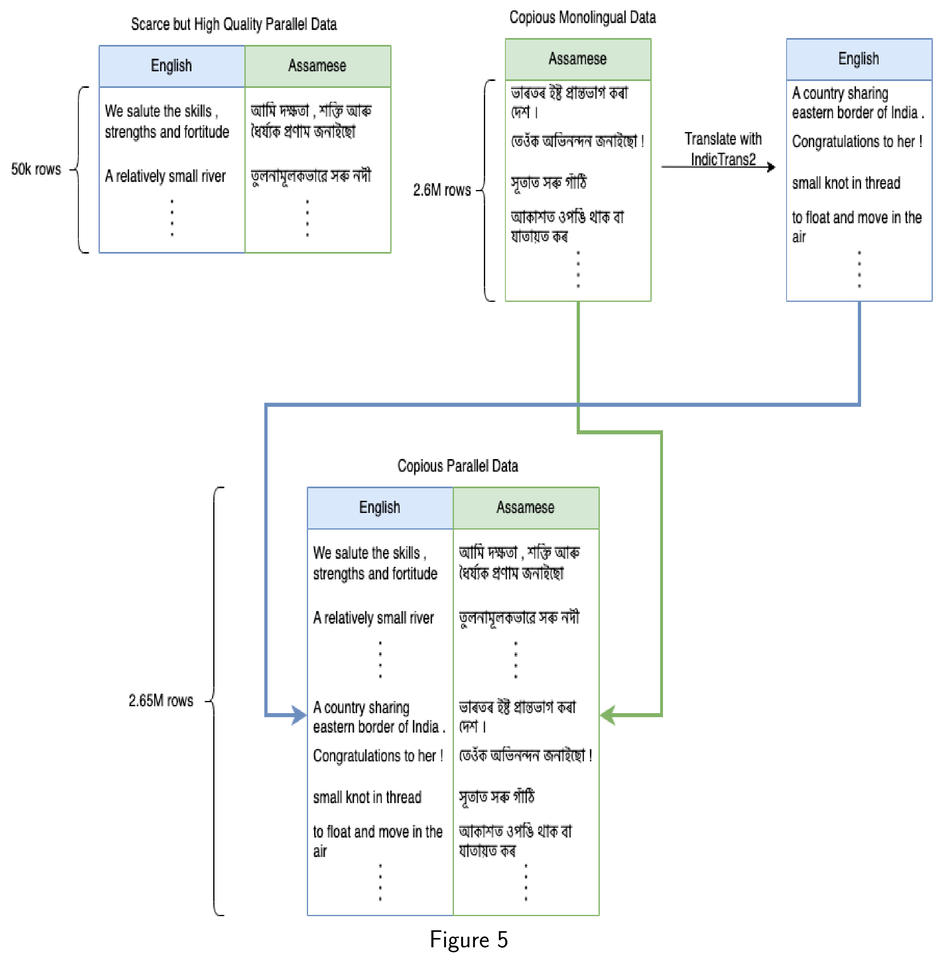

Figure 5

This technique called back-translation has been extensively used for transformer MT, but never before for LLM MT.

- The English sentences translated by the transformer will be paired with their high quality Indic original sentences.
- Given their strong understanding of English, even if the translated English sentences are not up to the mark, the LLMs can easily infer their true meaning.
- Back-Translation has potential to boost model performance even more significantly in the LLM world than it did for transformers.