ENG2TEL Transalatron:English to Telugu Translation Project -Bridging Linguistic Gaps

Alokam Gnaneswara Sai & Onteru Prabhas Reddy & Pallekonda Naveen kumar
Indian Institute of Science
Bengaluru, KA, India
{gnaneswaras,kpnaveen,prabhasreddy}@iisc.ac.in
GitHub Repository link: GitHub Repo

Abstract

In our study, we conduct comprehensive experiments, evaluating publicly available multilingual language models on machine translation tasks such as mBART50(1) and NLLB for Telugu language. mBART helps with translation by pretrained with large-scale monolingual corpora in many languages. It was created by Facebook AI researchers and is a sequence-to-sequence(2) denoising auto-encoder pre-trained on lots of different languages, NLLB (3) (which stands for "no language left behind") is a machine translation models published by Meta AI in 2022. We compare the performance over the fine-tuning methodology.

1 Introduction

Machine Translation(MT) coverage for more users speaking diverse languages is limited because the MT methods demand vast amounts of parallel data to train quality systems, which has posed a significant obstacle for low-resource translation. Therefore, developing MT systems with relatively small parallel datasets is still highly desirable.

In this report, we aim to investigate the performance of Fine tuned LLMs on MT tasks on Telugu language, with a particular focus on encoder-decoder based LLMs, Our project focuses on a range of publicly available medium sized LLMs. This includes models pretrained on large corpus, such as mBART,NLLB.We evaluate various versions of these models, with their parameter sizes of 0.61 billion.

In our experiments, we explore fine-tuning standard sequence-to-sequence MT models focus on translating one sentence at a time, overlooking discourse phenomena and the broader context.We demonstrate the effectiveness of fine-tuning on a English-Telugu dataset.

2 Related Work

2.1 LLM Application

Leveraging LLMs across a spectrum of downstream natural language processing (NLP) tasks is now a prevailing approach. However, the optimal strategies for utilizing these models both effectively and efficiently remain an open question. Broadly speaking, there are primary methods to build applications based on LLMs:

- **Fine Tuning:** Fine-tuning involves extending the training of the LLMs using additional, task-specific data.This is particularly beneficial when such tailored datasets are available.

2.2 LLMs for MT

Recent literature has begun to explore the application of LLMs for MT on low resource languages, an area that remained relatively under-explored until now.Their findings suggest that while these decoder-only LLMs are competitive, they still lag behind when compared to the encoder-decoder based multilingual language model NLLB studied the impact of LLM data on MT.

While prior studies have highlighted the potential of LLMs in MT, their focus has been primarily on high resource langagues. A significant gap remains in the exploration of fine-tuning LLMs specifically for MT tasks on low resource languages like Telugu. Recognizing this oversight, the primary objective of this is to address and bridge this research gap.

3 Methodology

The NLLB, mBART50 models are variants of the Transformer architecture specifically designed for multilingual translation tasks. They utilizes a shared vocabulary across multiple languages and are pre-trained on a large corpus of text from various languages. Fine-tuning allows us to adapt these models to specific translation tasks by providing it with domain-specific data.

3.1 NLLB model

We have choosen the NLLB model as one of the base for the project because it is one of the best model for translation, including Telugu. It's like a strong base that can handle complex language stuff well supporting more than 200 languages.

After considering NLLB pretrained model and NLLB pretrained tokinizer we fintuned it on subset of Samanantar dataset by randomly sampling 60K rows.In pre fintuning phase we observed that many sentences in Telugu are containing unknown tokens by which there can be diminish in performance .So we preprocessed the Telugu text accordingly so that no unknown tokens will be there throughout the dataset.

The approach we are looking forward is as follows :

1)Adding more meaningful Telugu tokens into the NLLB tokinizer.This is reasonable because the NLLB tokinizer contained vocabulary size of 256K tokens of which low resource language like Telugu will be dedicated less number of tokens because it has to support more than 200 languages.Also we observed that each Telugu word is getting divided into on a average of 2.65 tokens on Samanantar dataset ,whereas only 1.42 for English.This demonstrates the linguistic gap between them. So the new tokens should be very high in quality and should be able to capture more local dependencies in Telugu.So we will be considering quality datasets like TeluguWiki.

2)To generate new quality tokens we will use tokenizers like Sentence piece and then the NLLB pretrained model embedding layer and softmax layers should be resized accordingly.

3.2 mBART model

Before fine-tuning the mBART50 model, we preprocess the parallel corpus data. This involves tokenization, where we split the sentences into individual tokens using the mBART50 tokenizer. Additionally, we apply any necessary cleaning and normalization steps to prepare the data for training.

From the vast collection of parallel corpora available in the Samanantar dataset, we selected 60,000 samples. We partitioned the collected dataset into three subsets: training, validation, and test sets. The partitioning was performed with a ratio of 70% for training, 20% for validation, and the remaining 10% for testing. This ensures that the model is trained on

a majority of the data while also having separate sets for tuning hyperparameters and evaluating performance.

3.3 mT5 Transformer model

The mT5 transformer is an immense multilingual text-to-text pre-trained transformer.To finetune the transformer to our needs(English-Telugu translation) we make use of samanantar dataset in which we use 50000 sentence pairs.The sentences are divided into three sets, namely test, train and dev sets.The mT5 is pre-trained on immense amounts of data already. mT5 supports a maximum token length of 20 tokens. This is what we have used for our model. Longer sentences have been truncated to 20 tokens and Out-Of-Vocabulary (OOV) tokens are discarded. Besides, it has also been shown how backtranslation can significantly hurt the model rather than improve it, if the quality of the translations is poor.We trained our model employing the AdamW optimizer, which shows better results than the default Adam optimizer. The parameters set by us are: number of epochs to run = 1, size of each batch = 4.

4 Experiemental Setup

4.1 Datasets

In this study, we focus on the translation direction from English to Telugu due to its significant demand for high-quality translation and the availability of substantial parallel data. Our fine tuning set includes the commonly used Samanantar which consists of 4.95 million english-telugu sentence pairs. The dev and test sets are the FLORES-22 Indic dev set,IN22 test set,These datasets are constructed from documents, thus enabling a natural evaluation of sentence-level translation. Table 1 summarizes the statistics of the datasets used in project.

Table 1: Dataset Statistics

	Datasets	**#sentences**
train	Samanantar	5M
dev	Flores-22	1K
test	IN22 Test set	1.5K

4.2 Pretrained LLMs

We investigate a varied collection of pretrained LLMs accessible on HuggingFace all based on the transformer architecture (4). This collection comprises two distinct LLMs, each trained on either English-centric or multilingual data and available in multiple versions with varying parameter sizes.This results in a comprehensive assortment of models, with parameter sizes 0.61 billion. Table 2 summarizes the models included in our study.

Table 2: Overview of evaluated LLMs

Model	**Data**	**Size(Billions)**
NLLB	Multilingual	0.61
mBART	Multilingual	0.61

4.3 Fine-tuning Setup

We configure the learning rate to 5e-4 and employ the AdamW optimizer (5) for the training process. A batch size of 4 is used. Models are with less than 1 billion parameters are trained on a single NVIDIA A100 GPU with 16GB of memory. The tokenizer is configured with a **max-input-length** of 128 for the source language (English) and a **max-target-length** of 128 for the target language (Telugu).

4.4 Evaluation Metrics

We use BLEU (6) as evaluation metrics to assess the performance of our models.For BLEU we use the SacreBLEU (7) (Post, 2018) implementation, which standardizes tokenization and facilitates reproducibility.By employing BLEU, we can ensure that our evaluation is robust and comprehensive, accounting for not only the lexical similarity between the translations and the references but also the overall quality and preservation of meaning in the translations

5 Results

- Fine-tuning demonstrates significant promise for sentence-level translations, enhancing the SacreBLEU scores of 6.7. The notable improvement is seen in Flores-22 dataset, which witnesses a SacreBLEU increment of 2.2 (from 4.5 to 6.7).
- We had achieved the SacreBLEU score of 7.16 for pretrained NLLB,and after fine tuning pretrained NLLB with Samanantar dataset with 60000 samples of which 80% of training and 20% of testing , we achieved a SacreBLEU score of 8.81 and further preproccessing of the unknown words of finetuned NLLB achieved the SacreBLEU score of 9.034.
- Table 3 presents the final results for sentence-level translations.

Table 3: Results

Model	**Training Size**	**#epoch**	**BLEU**
mBARt	35000	1	7.8
NLLB	48000	1	9.03

- We also observed that training for more epochs could increase the quality of translation.

Conclusion

For this report we have exploited the transformer models like mBART50 and NLLB and have fine tuned the models using 60000 training samples taken from samanantar dataset and was able to acheive decent BLEU score.For the training we trained the model with 1 epoch.We had tried to implement the mT5 transformer model.as mT5 model is not pretrained with the telugu corpus which results in drop of BLEU score.

For future work, we plan to delve deeper into the exploration of additional pre-trained large language models (LLMs), such as LLaMA2(8), coupled with LoRA(9) fine-tuning techniques. This approach aims to further enhance the performance of our translation models by exploiting the strengths of diverse pre-trained architectures.

Additionally, we intend to experiment with training our models using custom tokenization strategies specially designed to the characteristics of the English to Telugu translation task. By adapting tokenization of the target language, we anticipate improvements in translation accuracy and efficiency. Furthermore, we aim to augment our training dataset by increasing the number of samples, thereby providing the model with a richer and more diverse set of contexts to learn from. We will try to train the model with more no of epochs for increasing the performance of the model. We will also consider different metrics like ChrF++ (10),ROUGE score (11) for evaluation the quality of translation.

Contributions

Alokam Gnaneswara Sai fine-tuned the mBART50 model by setting different hyper parameters and read research papers relevant to the implementation of that model,and attempted to implement a transformer from scratch for training on the English-Telugu parallel corpus.

Onteru Prabhas Reddy preprocessed Telugu text accordingly to the NLLB model and fine-tuned the NLLB model by varying different hyper parameters on Samanantar dataset and read related research papers for further idea of optimizations like increasing language specific tokens in tokenizer and model layers reshaping.

Pallekonda Naveen Kumar Fine-tuned the mT5 model helped in reading research papers. Additionally, contributed to the literature review process by reviewing and summarizing relevant research papers.attempted to implement a LSTMs on the English-Telugu parallel corpus.

References

[1] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. 2020.

[2] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

[3] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[7] Matt Post. A call for clarity in reporting bleu scores, 2018.

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[10] Maja Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[11] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.