Code Mixed Machine Translation

Problem

The project aims to enhance code-mixed translation, specifically Hinglish to English, to make it more accessible and accurate.

Translating Hinglish to English in code-mixed machine translation faces challenges. Our project aims to enhance accuracy and accessibility, surpassing the baseline's 12.2 BLEU score.

How did we proceed

Dataset exploration:

We used the dataset provided by the task organizers for our systems Since the target sentences in the dataset contain Hindi words in Roman script, We use the CSNLI library (Bhat et al.,2017, 2018) as a preprocessing step.It transliterates the Hindi words to Devanagari and also performs text normalization.

We use the provided train:validation:test split, which is in the ratio 8:1:1

Models trained or fine tuned

Neural machine translation GRU with attention:

The Neural Machine Translation (NMT) model architecture consists of two key components: the encoder and the decoder.

The encoder processes input sentences using a bidirectional GRU structure. It has one GRU that reads the sentence forward and another GRU that reads it backward. The forward and backward hidden states (ht and ht) at time t are computed sequentially. The final encoder hidden state (ht) for each word is formed by concatenating these forward and backward hidden states.

The decoder, on the other hand, utilizes a GRU network to generate translated sentences one word at a time. At each time step t, it calculates the hidden state (st) based on the previous hidden state (st-1) and the context vector (ct).

Models trained or fine tuned

LSTM sequence to sequence

It consists of an Encoder and a Decoder. The Encoder processes input sequences, using LSTM layers after an embedding layer to capture context through bidirectional processing. The final hidden states from the Encoder serve as context. The Decoder takes this context and generates target sequences using unidirectional LSTM layers and an embedding layer. At each step, it predicts an output token, based on the context and prior predictions. This process continues until an end-of-sequence token or a maximum sequence length is reached, allowing translation from source to target language.

Seq to Seq transformer with attention

The Transformer architecture diverges from its LSTM and GRU predecessors by exclusively leveraging self-attention mechanisms to gauge the significance of various input data segments. In the encoder, it comprises multiple layers, each featuring two primary sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Essential to this architecture are residual connections and subsequent layer normalization. The decoder mirrors this structure but introduces a third sub-layer for multi-head attention over the encoder's output. These attention mechanisms enable the model to selectively focus on relevant sections of the input sequence, whether in self-attention within the encoder, encoder-decoder interactions, or self-attention within the decoder.

BERT and its variants

Transformer-based models, especially BERT and its variations, have significantly enhanced machine translation quality. This report delves into their adaptation through fine-tuning to address the intricacies of language translation. These models share a common architecture: a multi-layer bidirectional transformer encoder distinguished by its attention mechanisms, allowing it to emphasize different input sequence elements, and a deep stack of transformer layers.

Fine-tuning is the process of customizing pretrained models for specific tasks by further training them on task-specific datasets. This adaptation enhances their ability to perform machine translation effectively.

BERT infused NMT

Recent advances in pre-training techniques like ELMo, GPT-2, BERT, XLM, XLNet, and RoBERTa have made waves in natural language processing. BERT and its variants shine in language understanding tasks. Neural Machine Translation (NMT) typically involves an encoder and decoder for language translation. Using BERT for NMT has been explored through two strategies: initializing NMT models with BERT and using BERT for context-aware embeddings. The latter approach led to the BERT-fused model, which incorporates BERT representations into all NMT layers and achieves top-tier results in supervised, semi-supervised, and unsupervised translation tasks. This work's components are known as Enc, Dec, and BERT within the NMT module.

. la maison de Léa <end> .

Mistral 7B LLM

The Mistral 7B Language Model, renowned for its few-shot learning ability, shows promise in Machine Translation. This report explores its architecture and usage in translating between languages with minimal example prompts.

Mistral 7B, a transformer-based model, optimizes efficiency and performance. It leverages self-attention mechanisms for parallel data processing, enhancing speed and scalability. With 7 billion parameters, Mistral 7B effectively stores extensive linguistic knowledge, making it ideal for few-shot learning applications in Machine Translation.

mT5-small

The mT5 model is a groundbreaking multilingual text-to-text framework that transforms machine translation by leveraging a unified approach to process over 100 languages. By fine-tuning mT5 with diverse datasets and specific translation tasks, it significantly enhances translation accuracy and fluency. The fine-tuning process involves adjusting the pre-trained model with a focus on translation datasets and optimizing hyperparameters to achieve peak performance. Evaluated using benchmark datasets and BLEU score metrics, the fine-tuned mT5 demonstrates high levels of translation quality, indicative of its practical application potential. Despite this success, challenges remain in maintaining consistent performance across languages, especially those with fewer resources. Overall, fine-tuning mT5 represents a substantial leap forward in the realm of machine translation, paving the way for more seamless and effective cross-language communication.

IndicBard

IndicBART is a specialized transformer-based model designed to address the complexities of translating Indic languages, boasting an encoder-decoder structure fine-tuned for the linguistic patterns and scripts unique to this language group. It utilizes a comprehensive pre-training on a vast corpus of Indic texts, followed by fine-tuning with parallel corpora to enhance its translation capabilities. The training methodology is carefully crafted, with a focus on text normalization and adaptive learning strategies to handle the computational challenges presented by Indic scripts. In evaluations, IndicBART outperforms existing models, delivering higher BLEU scores and demonstrating a deep understanding of context and idiomatic nuances. This report underlines IndicBART's role as a benchmark in Indic language translation, with future research aimed at expanding its scope to include speech-to-text features and support for lesser-spoken dialects, further advancing the field of Machine Translation.

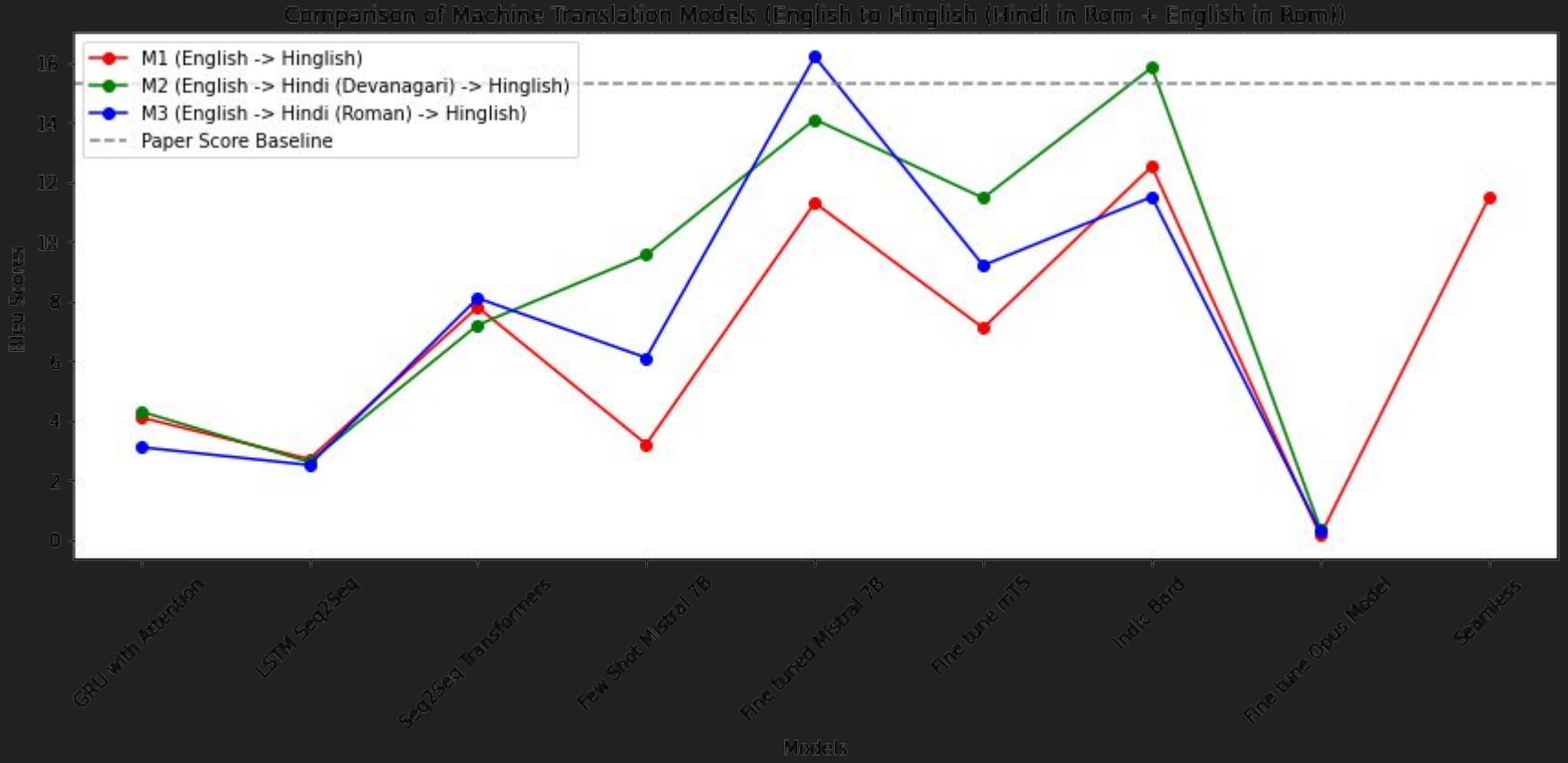
Opus Model

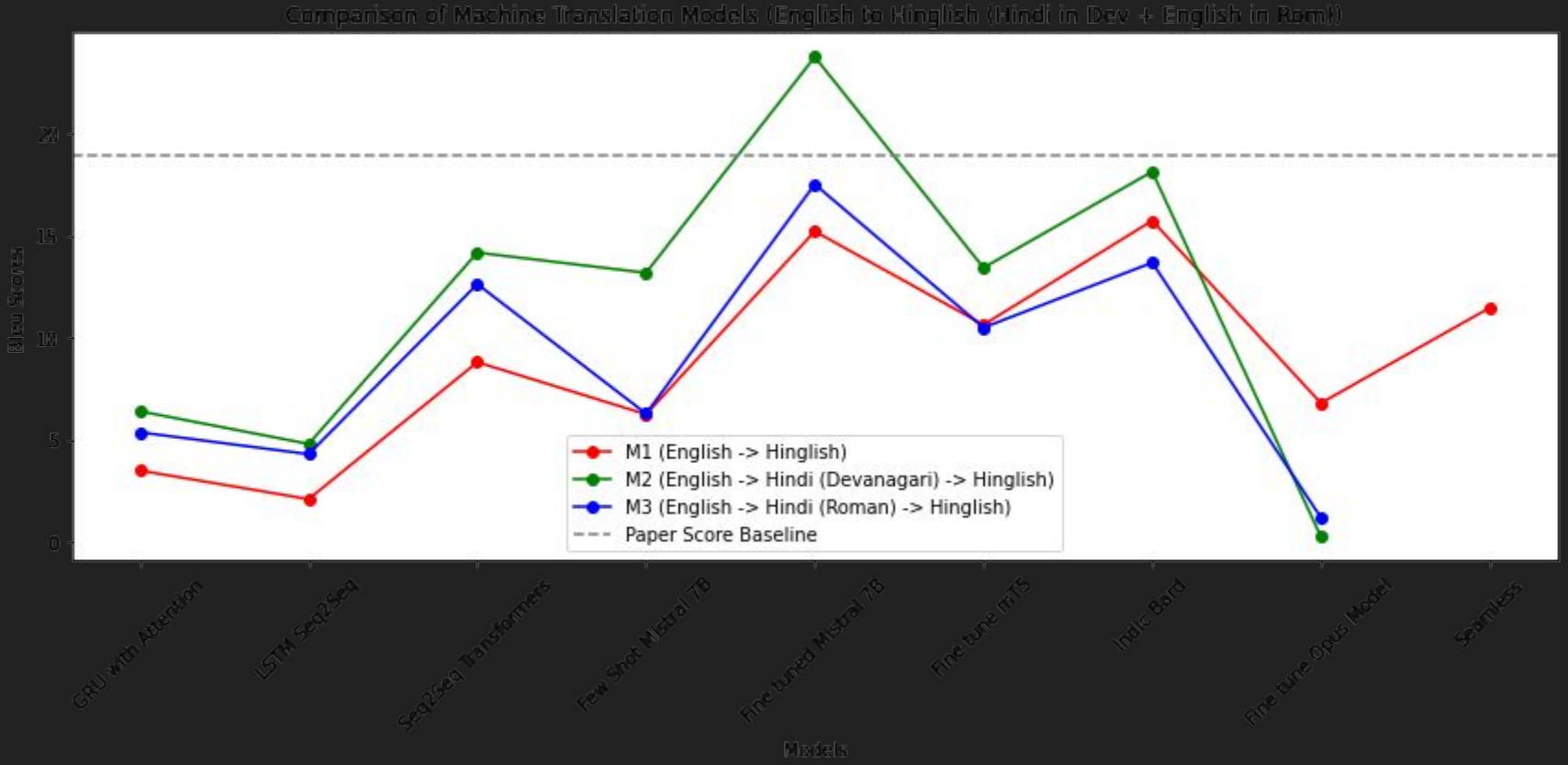
This report delves into the fine-tuning of the Opus Model, a leading neural machine translation framework, to enhance its translation accuracy for specific linguistic tasks and domains. Fine-tuning involves customizing the model's parameters to address unique language nuances, jargon, and stylistic preferences pertinent to particular fields, like legal or scientific texts. The Opus corpus, a vast collection of texts in various languages, serves as the training foundation for the model, enriched by domain-specific datasets during fine-tuning. Techniques such as transfer learning, layer re-training, and hyperparameter optimization are key to this process. The performance of the fine-tuned model is rigorously evaluated against industry benchmarks using metrics like BLEU and METEOR, with results showing marked improvements in translation quality. The report concludes with the positive impact of fine-tuning on the Opus Model's precision and its potential for future applications in NMT technologies.

Seamless

The Seamless machine translation model is an advanced NMT system that epitomizes the evolution of machine translation, offering exceptional fluidity and accuracy. It builds upon previous encoder-decoder architectures, introducing innovative components like sophisticated attention mechanisms and contextual understanding. The encoder uses deep transformer layers or advanced recurrent units for nuanced linguistic representation, while the decoder employs complex language models for generating target text. What sets Seamless apart is its contextual integration, considering beyond sentence-level information to paragraph or document themes, enhancing the relevance and coherence of translations. It's trained on extensive datasets with advanced algorithms and substantial computing resources. Performance is measured against top metrics like BLEU and includes thorough qualitative analyses. The results underscore its proficiency across diverse language pairs and contexts. The report concludes with the Seamless model's potential to transform global communication and future research directions.

Results


Comparison of Machine Translation Models (English to Hinglish (Hindi in Rom + English in Rom))
M1 (English -> Hinglish)
M2 (English -> Hindi (Devanagari) -> Hinglish)
M3 (English -> Hindi (Roman) -> Hinglish)
Paper Score Baseline
BLEU Scores
0
2
4
6
8
10
12
14
16
GRU with Attention
LSTM Seq2Seq
Seq2Seq Transformers
Few Shot Mistral 7B
Fine tuned Mistral 7B
Fine tune mT5
Indic Bard
Fine tune Opus Model
Seamless
Models


Comparison of Machine Translation Models (English to Hinglish (Hindi in Dev + English in Rom))
Bleu Scores
0
5
10
15
20
M1 (English -> Hinglish)
M2 (English -> Hindi (Devanagari) -> Hinglish)
M3 (English -> Hindi (Roman) -> Hinglish)
Paper Score Baseline
GRU with Attention
LSTM Seq2Seq
Seq2Seq Transformers
Few Shot Mistral 7B
Fine tuned Mistral 7B
Fine tune mT5
Indic Bard
Fine tune Opus Model
Seamless
Models

Bleu scores comparison table

Model	M1	M2	M3	
GRU with attention	4.1	4.3	3.1	15.3
LSTM seq2seq	2.7	2.6	2.5	15.3
Seq2Seq Transformers with attention	7.8	7.89	8.1	15.3
Shot prompting on mistral 7B LLM	3.2	9.56	6.1	15.3
Mistral 7B llm fine tuning	11.3	14.1	16.2	15.3
Fine tuning mT5	7.1	11.46	9.2	15.3
Indic bard	12.52	15.84	11.5	15.3
Finetuning opus	0.14	0.37	0.27	15.3
Seamless	11.46			15.3

Bleu scores (normalised) comparison table

Model	M1	M2	M3	
GRU with attention	3.5	6.4	5.37	18.9
LSTM seq2seq	2.1	4.8	4.3	18.9
Seq2seq transformers with attention	8.8	14.16	12.62	18.9
shot prompting on mistral 7B llm	6.25	13.17	6.28	18.9
Fine tuning Mistral 7B llm	15.2	23.73	17.5	18.9
Fine tune mT5	10.63	13.42	10.49	18.9
Indic bard	15.72	20.5	13.67	18.9
Fine tune opus model	6.8	0.26	1.2	18.9
Seamless	11.46			18.9

Challenges encountered during implementation

Limited Data Resources: Obtaining high-quality and extensive datasets for Hinglish proves to be a challenging task, impeding the training of models with robust generalization capabilities.

Computational Limitations: The fine-tuning process of large language models (LLMs) demands substantial computational power, often surpassing the capabilities of standard GPUs. This makes it economically unfeasible without access to high-end computational resources.

Code-Switching Complexity: Hinglish commonly involves code-switching, where speakers seamlessly alternate between Hindi and English within a single conversation or even a sentence. This intricate linguistic phenomenon adds a layer of complexity for models required to comprehend and generate text in such a hybrid context.

Cultural Nuances: Grasping the cultural subtleties embedded within Hinglish expressions poses a significant challenge for machine translation systems, which may lack the finesse needed for precise translation.

Transliteration Inconsistencies: The absence of a standardized transliteration scheme from Devanagari (Hindi script) to the Roman alphabet results in inconsistencies when representing Hinglish words across different datasets and systems.

Ambiguity and Polysemy: Hinglish, like many natural languages, contains words with multiple meanings based on context. This ambiguity presents a substantial challenge for natural language understanding and generation.

Evaluation Metrics: Conventional evaluation metrics for language tasks may not be entirely suitable for assessing Hinglish due to its distinctive characteristics, necessitating the development of specialized evaluation criteria.

Annotated Corpus Quality: The quality of annotations in Hinglish corpora can be variable, affecting the training of models reliant on high-quality annotated data for tasks such as part-of-speech tagging, named entity recognition, and sentiment analysis.

Navigating Translation Challenges & Baseline Model Insights

- **Baseline Model Evaluation:**
  - Model Used: mBART with a BLEU score of 15.3.
  - Strengths: Efficiency, consistency, and support for multiple languages.
  - Weaknesses: Struggles with capturing nuances and contextual understanding.

Thank You