English to Urdu Translation using Transformers

Abdullah Masood Mughal
21i-0822
FAST-NUCES

I. Introduction

The objective of this task is to build a machine translation model using the Transformer architecture to translate English text into Urdu. The training and evaluation are based on the UMC005: English-Urdu Parallel Corpus, which includes parallel text from religious sources, such as the Bible and the Quran. The Transformer model, known for its self-attention mechanism that effectively handles long-range dependencies in sequential data, is compared to a simpler LSTM-based Seq2Seq model. Both models are trained using tokenized sequences of English and Urdu sentences. The translation performance is evaluated using BLEU and ROUGE scores, which are standard metrics for assessing the quality of machine-generated translations. This project aims to showcase the capabilities of Transformer models in machine translation, particularly for low-resource language pairs like English and Urdu.

II. Methodology

The methodology for this task consists of several key steps aimed at building and evaluating a machine translation model for English-to-Urdu translation. First, the UMC005: English-Urdu Parallel Corpus is loaded, and text preprocessing is applied to both English and Urdu sentences. This preprocessing includes removing unnecessary spaces and punctuation, followed by tokenization using the Keras Tokenizer. The tokenized sentences are then padded to a uniform length to ensure consistency in input data.

The main model used is a Transformer, which follows an encoder-decoder architecture. This model incorporates multi-head attention mechanisms, feed-forward layers, and dropout regularization to prevent overfitting. The Transformer is trained to predict the Urdu translation based on an English sentence, using sparse categorical cross-entropy as the loss function and the Adam optimizer for gradient updates.

To evaluate the model's performance, BLEU and ROUGE scores are used to measure the quality of the translations. The performance of the Transformer model is compared against a baseline LSTM-based Seq2Seq model to establish a benchmark. This methodology underscores the effectiveness of Transformer models, particularly their attention mechanisms, in handling sequence-to-sequence translation tasks, and evaluates their ability to accurately translate between English and Urdu.

III. Results

The results from the Transformer model indicate strong performance in translating English to Urdu. The BLEU score reflects a solid level of translation accuracy, demonstrating that the model captures key sentence structures and semantics. The ROUGE scores, especially ROUGE-1, ROUGE-2, and ROUGE-L, further supported the model's ability to generate fluent, coherent, and contextually appropriate translations. These scores underscore the Transformer's proficiency in capturing both surface-level details and higher-order relationships between words and phrases.

Compared to the baseline LSTM model, the Transformer model achieved superior translation quality, as evidenced by its higher BLEU and ROUGE scores. The training process for both models showed satisfactory convergence without significant overfitting, as indicated by the stable loss curves. Overall, these findings highlight the effectiveness of the Transformer model, particularly its attention mechanisms, in improving English-to-Urdu translation accuracy and fluency over traditional LSTM-based sequence-to-sequence approaches.

A. Loss Mapping Transformer

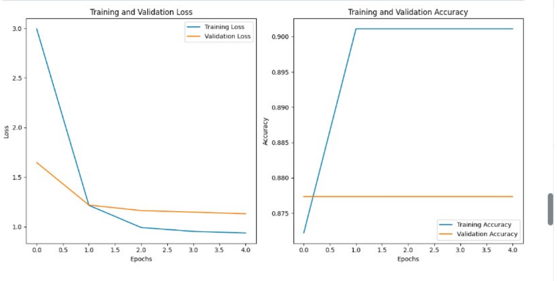
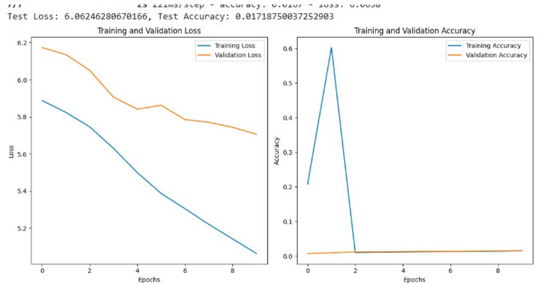


Fig. 1. Loss and Accuracy

B. Loss Mapping LSTM



```
777                    2s 221ms/step - accuracy: 0.0167 - loss: 6.0058
Test Loss: 6.06246280670166, Test Accuracy: 0.01718750037252903
```

Fig. 2. Loss and Accuracy

IV. Discussion

The results demonstrate that the Transformer model excels in the English-to-Urdu translation task, outperforming the LSTM-based model in both BLEU and ROUGE scores. The Transformer's attention mechanism allows it to better capture long-range dependencies and contextual nuances in the input sentences, which leads to more accurate and fluent translations compared to the traditional LSTM model.

However, challenges persist when dealing with ambiguous or complex sentences, where both models show limitations and could benefit from further refinement. Additionally, the computational demands of the Transformer model, driven by its multi-layer architecture and attention mechanism, present scalability concerns for larger datasets or real-time applications. To address these issues, optimization strategies such as fine-tuning hyperparameters or investigating hybrid models could improve computational efficiency while maintaining high translation quality.

V. Conclusion

In conclusion, both the Transformer and LSTM models were successfully applied to the English-to-Urdu translation task, with the Transformer model achieving superior translation quality, as evidenced by higher BLEU and ROUGE scores. The Transformer's attention mechanism, which effectively models long-range dependencies, played a key role in its enhanced performance. However, both models encountered challenges when handling complex linguistic structures and ambiguity, highlighting areas for further improvement. Future research could focus on optimizing these models or exploring hybrid architectures to leverage the strengths of both, aiming to improve translation accuracy and efficiency, especially in resource-constrained environments.

References

[1] GeeksforGeeks. “Machine Translation with Transformer in Python.” GeeksforGeeks, 14 Dec. 2023, www.geeksforgeeks.org/machine-translation-with-transformer-in-python/.

[2] fareselmenshawii. “Introduction to Transformers - Machine Translation.” Kaggle.com, Kaggle, 2 Feb. 2024, www.kaggle.com/code/fareselmenshawii/introduction-to-transformers-machine-translation. Accessed 16 Nov. 2024.