

# English to Urdu Machine Translation

**Abdul Moiz Shehzad (22L-7468)**  
**Abdullah Riaz (22L-7489)**

February 2025

## 1 Research Problem and Motivation

Machine translation (MT) has seen significant advancements with the rise of deep learning and neural networks, yet high-quality English to Urdu translation remains an under-explored area. Despite Urdu being spoken by millions, the availability of robust translation models for this language pair is limited. Many state-of-the-art translation models focus on widely spoken languages like English, French, and Spanish, while low-resource languages like Urdu lack sufficient attention. This project aims to bridge this gap by leveraging advanced NLP techniques to develop an accurate and efficient English to Urdu translation model. The model will be trained using parallel corpora and optimized for real-world applications, following industrial standards like MLOps and containerized deployment.

## 2 NLP Techniques to be Used

To achieve high-quality translation, we will explore and experiment with various neural architectures depending on the project timeline and requirements. Possible techniques include:

- **Recurrent Neural Networks (RNNs):** Baseline model using bidirectional RNNs.
- **Long Short-Term Memory (LSTM) Networks:** Handling long-range dependencies in translation sequences.
- **Gated Recurrent Units (GRUs):** A more computationally efficient alternative to LSTMs.
- **Transformer Models:** Attention-based models such as the Transformer architecture, which has outperformed RNNs in translation tasks.
- **Pretrained Language Models & Fine-Tuning:** Leveraging open-source models from Hugging Face, fine-tuned on Urdu-English datasets to enhance translation quality.

- **Sequence-to-Sequence (Seq2Seq) Learning:** Applying encoder-decoder architectures for more fluent translation.
- **Additional Techniques as Needed:** Depending on project scope and progress, other state-of-the-art NLP techniques may be incorporated.

### 3 Datasets and Tools to be Used

For training and evaluation, the project will utilize publicly available parallel English-Urdu datasets:

- **English-Urdu Parallel Corpus** [5]
- **Parallel Corpus for English-Urdu Language** [6]

Tools and frameworks to be used:

- **Deep Learning Frameworks:** TensorFlow, PyTorch
- **Pretrained Models:** Hugging Face Transformers
- **MLOps and Deployment:** Docker, Flask/FastAPI/Streamlit for web deployment
- **Experimentation and Logging:** Weights & Biases, Zenml

### 4 Evaluation Metrics

The performance of the translation model will be assessed using industry-standard NLP evaluation metrics, including:

- **BLEU Score (Bilingual Evaluation Understudy):** Measures how close the generated translation is to the reference translation.
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** Improves upon BLEU by incorporating synonym matching.
- **TER (Translation Edit Rate):** Evaluates the number of edits needed to match the reference translation.
- **ROUGE Score:** Measures recall-oriented translation accuracy.

### 5 Conclusion

This project will contribute to the field of low-resource language translation by developing a state-of-the-art English to Urdu translation model. By leveraging advanced NLP techniques, industrial MLOps standards, and cloud-based deployment, the model aims to provide a scalable and efficient translation system for real-world applications. The exact techniques used will be determined based on the project timeline and progress.

## 6 References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 2017. Available at: <https://arxiv.org/abs/1706.03762>
- [2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *International Conference on Learning Representations (ICLR)*, 2015. Available at: <https://arxiv.org/abs/1409.0473>
- [3] M. Ott, S. Edunov, D. Grangier, and M. Auli, "Scaling Neural Machine Translation," in *Proceedings of the Third Conference on Machine Translation (WMT)*, 2018. Available at: <https://arxiv.org/abs/1806.00187>
- [4] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, and W. Macherey, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," in *arXiv preprint arXiv:1609.08144*, 2016. Available at: <https://arxiv.org/abs/1609.08144>
- [5] M. Anas Mahmood, "English-Urdu Parallel Corpus Dataset," Kaggle, 2022. Available at: <https://www.kaggle.com/datasets/muhammadanasmahmood/englishurdu-parallel-corpus>
- [6] Z. Uddin, "Parallel Corpus for English-Urdu Language," Kaggle, 2021. Available at: <https://www.kaggle.com/datasets/zainuddin123/parallel-corpus-for-english-urdu-language/data>