# Neural Machine Translation from English to Urdu using MarianMT: A Transformer-Based Approach

**Author:** **Faiz Azam (BAI-22S-010)**

Submitted to: **Syed Mohammad Hassan**

Department of Artificial Intelligence

Sindh Madressatul Islam University (SMIU)

**Abstract:**

The MarianMT model, which is based on the Transformer architecture, is used in this research to propose a neural machine translation (NMT) system for translating English to Urdu. Because Urdu is a low-resource language with a rich morphology, translation is extremely difficult. To improve translation accuracy, this study makes use of a pretrained MarianMT model (Helsinki-NLP/opus-mt-en-ur) that has been refined on parallel corpora. We examine earlier research in Urdu MT, point out any shortcomings, and demonstrate how the MarianMT model enhances fluency and performance. Human judgment and BLEU scores are used for evaluation.

## I. Introduction

The goal of machine translation (MT) is to translate text between languages automatically. Due to syntactic complexity and a lack of parallel corpora, translating from English to Urdu is still understudied, even though machine translation (MT) systems for high-resource language pairs (such English and French) have attained near-human accuracy.

This study investigates the use of MarianMT, a Transformer-based model trained by Helsinki-NLP, to translate text from English to Urdu with high accuracy and little setup. The need for more easily available Urdu language resources for web content, government, and educational reasons is the driving force.

## II. Related Work

### 2.1 Rule-Based & Statistical MT

The morphological richness of Urdu could not be captured by early attempts that relied on rule-based algorithms, such as Google's SMT models. Standard statistical MT tools like Moses and GIZA++ were employed, but they lacked context awareness.

2.2 Neural Machine Translation

End-to-end NMT was first introduced by Facebook's FAIR system and Google's GNMT system. The quality of translation greatly increased with the use of Transformer models (Vaswani et al., 2017).

**2.3 English-Urdu Translation**

Recent Urdu translation works include:

- **UET's Urdu-Corpus (2012)** – First large-scale dataset.
- **Saman et al. (2019)** – Used LSTM-based Seq2Seq models.
- **Helsinki-NLP (2020)** – Released pretrained MarianMT for 100+ language pairs including en-ur.

However, few have evaluated MarianMT specifically for English-Urdu translation in low-resource academic or real-world contexts.

## III. Methodology

3.1 MarianMT Model

We employ Hugging Face Transformers to implement the Helsinki-NLP/opus-mt-en-ur model. MarianMT is an encoder-decoder model that was trained on OPUS parallel corpora and is based on the Transformer architecture.

3.2 Preprocessing

- Tokenization with MarianTokenizer
- Use of PyTorch tensors (CPU inference)

3.3 Translation Function

```python
def translate(text):
    tokens = tokenizer(text, return_tensors="pt", padding=True, truncation=True)
    translation = model.generate(**tokens)
    return tokenizer.decode(translation[0], skip_special_tokens=True)

```

### 3.4 Sample Output

Input: *"How are you?"*

Output: *"آپ کیسے ہیں؟"*

## IV. Evaluation

### 4.1 Qualitative Analysis

- Fluency: Translations are contextually appropriate and natural.
- Grammar: Sentence structure and verb agreements are mostly accurate.

### 4.2 Quantitative Metrics

Due to hardware limitations, BLEU scores were approximated via human-labeled translations.

| Sentence | BLEU-like Match | Human Rating (1-5) |
|---|---|---|
| **“I am learning NLP.”** | **High** | **5** |
| **“My name is Faiz.”** | **High** | **5** |
| **“The weather is nice today.”** | **Moderate** | **4** |

## V. Advancements Made

**GUI Integration:** A clean interface for easy text input/output.

**Error Handling:** Integrated feedback when PyTorch was missing.

**Updated Tokenization:** Replaced deprecated methods with modern Hugging Face practices.

**Real-Time Testing:** Enabled translation with instantaneous feedback.

## VI. Limitations

Limited user testing due to system dependency on PyTorch.

Model not trained on domain-specific (e.g., legal, medical) Urdu.

## VI. Conclusion

Our use of MarianMT for English-to-Urdu translation yields encouraging outcomes. This approach can be implemented for governmental or educational platforms to close the language gap in multilingual cultures like Pakistan with minor adjustments and extensive user testing.

## VII. Future Work

Integrating BLEU/METEOR auto evaluation.

Adding voice input (ASR) for speech-to-translation.

Deploying as a web app for wider accessibility.

## References


1. Vaswani, A., et al. (2017). *Attention Is All You Need*. NeurIPS.
2. Tiedemann, J. (2020). *The OPUS-MT Project*. Helsinki-NLP.
3. Koehn, P. (2007). *Moses: Open-Source Toolkit for SMT*. ACL.
4. Saman, S. et al. (2019). *Urdu Machine Translation using Seq2Seq*. IJCNLP.
5. Hugging Face. (2023). *Transformers Library Documentation*.
6. UET Lahore. (2012). *Urdu-English Parallel Corpus*.