Polyglot Pal: A Context-Aware Neural Machine Translator for Indian Languages Using MBart50

Abstract

This paper presents Polyglot Pal, a lightweight neural machine translation (NMT) system designed to facilitate real-time phrase-based translation between English and five widely spoken Indian languages: Hindi, Telugu, Tamil, Bengali, and Kannada. The system leverages MBart50, a multilingual transformer-based architecture, as its core translation engine. With a special emphasis on low-resource linguistic scenarios and cultural relevance, this project demonstrates the potential of contextual translation in multilingual environments like India.

1. Introduction

India's vast linguistic diversity includes 22 scheduled languages and hundreds of dialects, making multilingual communication both vital and challenging. While global platforms like Google Translate offer support for Indian languages, they often lack cultural fluency and contextual relevance, especially in informal or conversational settings.

Polyglot Pal was developed to overcome these limitations by providing accurate, phrase-level translations that consider both linguistic structure and sociocultural context. The tool caters especially to real-world scenarios such as tourism, inter-state migration, and education, where accurate and sensitive communication is crucial.

2. Related Work

Early machine translation techniques were rule-based, evolving into statistical models, and now predominantly neural-based systems. Transformer architectures, particularly those designed for multilingual training like MBart, have revolutionized the field by enabling zero-shot translation capabilities.

Recent efforts such as IndicNLP, AI4Bharat, and IndicTrans have emphasized the importance of Indian language translation in low-resource settings. These initiatives contribute valuable corpora and pretrained models but are often optimized for full-document or sentence-level translation rather than everyday, phrase-based interactions.

3. Methodology

3.1 Data Collection

We curated a dataset of 200+ commonly used phrases in domains like travel, emergency, greetings, and shopping. Phrases were translated from English into Hindi, Telugu, Tamil, Bengali, and Kannada by native speakers. Each translation was validated for accuracy, naturalness, and contextual appropriateness.

3.2 Model Architecture

We use MBart50, a multilingual transformer encoder-decoder model pretrained with denoising objectives. MBart50 enables many-to-many translation by prepending language-specific tokens. Its architecture consists of:

- 12 encoder & decoder layers
- Self-attention with positional embeddings
- Shared vocabulary via SentencePiece tokenization

3.3 Application Interface

We developed a simple web application using Flask. The user can:

- Select source and target languages
- Enter a phrase
- View output translation instantly

A unique feature, Context Mode, allows users to select between "formal" and "informal" translation tones, altering the phrasing accordingly.

4. Challenges

1. Idiomatic Expressions: Common idioms often produce awkward or literal translations.

2. Word Order Differences: English follows Subject-Verb-Object (SVO), whereas most Indian languages follow Subject-Object-Verb (SOV).

3. Gender and Politeness Forms: Indian languages require adjustments based on gender and social hierarchy.

4. Script Variations: Rendering text in native scripts requires proper Unicode handling and font support.

5. Results and Discussion

5.1 Human Evaluation

Native speakers evaluated 50 randomly selected phrases using metrics such as Fluency, Contextual Accuracy, and Cultural Relevance. Polyglot Pal outperformed Google Translate significantly.

5.2 Performance

- Translation Time: ~1.3 seconds per phrase on standard CPU

- Model Size: 1.5 GB

- UI Load Time: < 500ms

6. Conclusion and Future Work

This project demonstrates the viability of context-sensitive neural machine translation tailored for Indian

languages. Polyglot Pal proves that even a lightweight system can yield high-quality translations in culturally complex settings.

Future Directions:

- Expand Dataset through community contributions
- Add speech integration
- Create offline-capable mobile app
- Include auto language detection features

References

1. Liu, Yinhan, et al. "Multilingual Denoising Pre-training for Neural Machine Translation." arXiv:2001.08210 (2020).

2. Kunchukuttan, Anoop, et al. "AI4Bharat IndicNLP Corpus." AI4Bharat (2020).

3. Vaswani, Ashish, et al. "Attention is All You Need." NeurIPS (2017).

4. Wolf, Thomas, et al. "Transformers: State-of-the-art Natural Language Processing." EMNLP 2020.

5. Flask Documentation - https://flask.palletsprojects.com

6. Hugging Face MBart50 - https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt