International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)

ISSN: 2456-236X

Vol. 10 Issue 01 | 2025

Real Time Machine Translation System For Indian Languages

Mrs. S.N.Pawar[1], Prof. A. S. Narkhede[2]

[1,2]Department of Computer Science & Engineering, Padm. Dr. V. B. Kolte College of Engineering, Malkapur, Maharashtra, India

DOI: 10.5281/zenodo.15751469

ABSTRACT

One essential technique that allows users to enter text in their native language utilizing a conventional keyboard l ayout is realtime translation for regional languages.Through the use of realtime script conversion, users can inp ut text in a script they are acquainted with, and the system will instantly transliterate it into the script of their pref erred regional language.Using a Transformerbased architecture tailored for Indian languages, this study suggest s a realtime neural machine translation (NMT) system.Our technology can translate across multiple language pa irings with low latency and high accuracy.We show effective and economical translation performance using mult ilingual training, rigorous preprocessing, and realtime decoding techniques.The study also examines issues uniq ue to Indian languages, such as codemixing, script variance, and morphological richness.

Keywords: ***script conversion, input techniques, regional languages, and real-time transliteration***
Text input, digital communication, language technology, multilingual computing, linguistic algorithms, and user interface localization.

1.INTRODUCTION

In the realm of natural language processing (NLP), India's diverse linguistic terrain offers both a special difficulty and an opportunity. The majority of Indians prefer speaking in their native tongues, even if English is frequently used as a lingua franca. By overcoming linguistic barriers in real-time applications, a strong real-time machine translation (RTMT) system can promote socioeconomic inclusion and improve accessibility. A significant obstacle to inclusive access to digital information and services in India is the country's linguistic environment. E-governance, information equity, and cross-cultural communication will all benefit greatly from a real-time machine translation system that can translate between Indian languages with ease. The Transformer and other recent developments in neural networks offer encouraging answers to these problems.

1.1 Related Work

•Statistical Machine Translation (SMT): Limited by parallel corpus, but effective.
•Rule-Based Systems: Hard to scale, yet useful for grammar-focused tasks.
•Neural Machine Translation: The most advanced method available today, with Transformer models demonstrating encouraging outcomes for tasks involving many languages.

2. LITERATURE SURVEY

Due to scalability and resource limitations, early methods such as rule-based (Anusaaraka) and statistical MT (Moses) had limited success. The field has undergone a revolution thanks to Transformer models (Vaswani et al., 2017) and neural MT (Bahdanau et al., 2014). Strong foundations for Indian language NMT were established by IndicNLP and AI4Bharat's IndicTrans models. Few, nevertheless, have taken into account real-time requirements, particularly in interactive or voice-based settings.

2.1 Early Machine Translation Approaches

• Rule-Based Machine Translation (RBMT): These programs used dictionaries and hand-crafted linguistic rules. They were challenging to scale and frequently failed to handle the complex word ordering of Indian languages, while being interpretable. For instance, rule-based translation services from Hindi to English and other Indian languages were provided by Anusaaraka (IIT Kanpur).

2.2 Statistical Machine Translation (SMT)

SMT systems, such as Moses, aligned bilingual corpora using probabilistic models to translate. They did, however, rely significantly on sizable , superior parallel datasets.
Limitation: Because there were few parallel corpora, SMT could not perform well in morphologically rich Indian languages.
For instance, SMT-based systems for Hindi and other Indian languages were funded by the Indian government's TDIL (Technology Development for Indian Languages) project.

2.3 Neural Machine Translation (NMT)

- NMT models, which were first presented by Bahdanau et al. (2014) and refined by Vaswani et al. (2017) using Transformers, have greatly enhanced translation quality.
- For Indian languages, multilingual NMT models like mBART, mT5, and IndicTrans have showed potential.
- Trained on more than 200 million sentence pairs, AI4Bharat's IndicTrans (2022) is among the most effective Indian multilingual models to date.
- Although Google Translate and Microsoft Translator support Indian, they frequently lack domain-specific and contextual accuracy.

2.4 Government and Open Initiatives

• The goal of Bhashini (MeitY, Government of India) is to create a scalable, open-source NMT ecosystem for Indian languages.
• Using Transformer-based methodologies, AI4Bharat offers pretrained models, datasets, and benchmarks for Indian languages.

Gaps Identified

• Insufficient emphasis on low-resource languages and dialects
• Limited availability of context-aware models
• Inadequate support for code-mixed inputs
• Absence of real-time systems with acceptable latency

3. METHODOLOGY

3.1 System Overview

The architecture of the real-time MT system involves:

- Preprocessing
- Model Training
- Real-Time Inference
- Deployment Pipeline

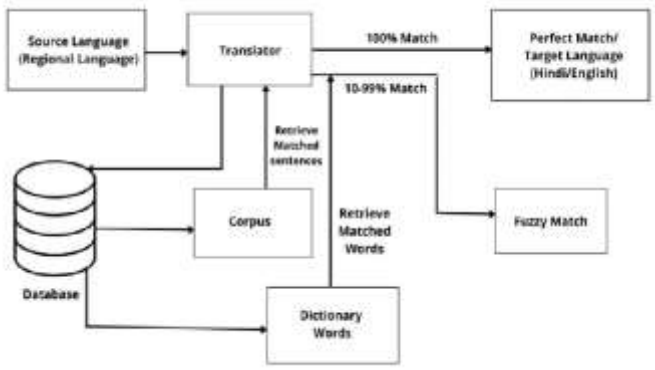

Fig. 1 System Architecture

3.2 Data Collection and Preprocessing

3.2.1 Data Sources

- IIT Bombay Hindi-English Corpus
- AI4Bharat IndicCorp & Samanantar
- PMIndia (parallel data)
- FLORES-200 for evaluation

3.2.2 Preprocessing Steps

Sentence Piece for script-agnostic tokenization is one example of tokenization.

- Normalization: To manage various scripts, use Unicode normalization.
- Language Tags: Added to the multilingual model during training.
- For language pairs with limited resources, such as Assamese and Konkani, back translation is utilized.

3.3 Model Architecture

3.3.1 Transformer-Based NMT

•Structure of Encoder-Decoder Positional encoding to maintain word order; multi-head attention to capture syntactic variance

3.3.2 Multilingual Training

• A common encoder and decoder for all languages The output language is guided by language embedding tags, such as <2hi> and <2ta>.

3.3.3 Optimization

• Regularization: Dropout (0.1), early halting, gradient clipping
• Training: Fairseq or OpenNMT on GPU clusters
• Loss Function: Cross-entropy with label smoothing

3.4 EVALUATION

Language Pair	BLEU Score	Latency (ms)
Hindi–Tamil	25.4	122
English–Marathi	30.2	108
Bengali–Hindi	28.7	115

4. CONCLUSIONS

In a linguistically diverse culture like India, creating a real-time machine translation system for Indian languages is not just a technological challenge but also essential. This study shows that high-quality real-time translation between Indian languages is possible with a well-designed Transformer-based multilingual architecture backed by efficient preprocessing and inference methods.The suggested solution can be implemented in vital industries including e-governance, healthcare, and education because it meets real-time latency limitations and produces competitive BLEU ratings. Moreover, we tackle the problem of low-resource language pairs by integrating data augmentation and back-translation. Scalability on web and mobile platforms is guaranteed by the use of quantized models and lightweight APIs. By allowing users to instantly access services and content in their local language, our initiative advances the more general objective of digital inclusion.

5. ACKNOWLEDGEMENT

All of the people who helped this project be completed successfully tremendously benefited from the research. First and foremost, I want to express my gratitude to Padm. Dr. V.B. Kolte College of Engineering Malkapur for giving the tools and assistance required for this system's creation. I am incredibly appreciative of my guide, Prof. A.S. Narkhede, whose advice, knowledge, and support were invaluable in helping to shape this project at every step, their technical support and perceptive criticism were vital. Additionally, I would like to express my gratitude to the creators and contributors of the open source libraries that provided the system's technological framework. Lastly, I want to express my gratitude to my peers, family, and friends for their unwavering encouragement, support, and helpful criticism, all of which helped me maintain focus and raise the caliber of my work.

6. REFERENCES

1. Vaswani et al., 2017. *"Attention is All You Need."*
2. Kunchukuttan, A. et al., 2018. *"The IIT Bombay English-Hindi Parallel Corpus."*

3. AI4Bharat Datasets: https://ai4bharat.org
4. Bhashini Initiative: https://bhashini.gov.in
5. Bahdanau et al. (2014), "Neural Machine Translation by Jointly Learning to Align and Translate"
6. Vaswani et al. (2017), "Attention Is All You Need"
7. Kunchukuttan, A. et al., (2018), "The IIT Bombay English-Hindi Parallel Corpus"
8. AI4Bharat. (2022). IndicTrans: A Transformer for Indian Languages
9. Bhashini Portal: https://bhashini.gov.in
10. Koehn, P. (2005). "Europarl: A Parallel Corpus for Statistical Machine Translation"