IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

Automated Language Translation System Using Transformer Neural Machine Translation

Sanika Santosh Kharade and Divya Ashok Rahane
Department of Artificial Intelligence and Data Science
Sahyadri Valley College of Engineering and Technology, Rajuri, Pune, India
sanikakharade28@gmail.com and divyarahane26@gmail.com

Abstract: *The rapid expansion of global digital communication and multilingual public services has created a strong need for accurate, low-latency Automated Language Translation Systems (ALTS). Early rule-based and phrase-based statistical machine translation pipelines relied on handcrafted linguistic rules and large phrase tables, which struggled with syntactic reordering, morphological inflection, and long-range contextual dependencies across structurally divergent languages such as English, Hindi, and Marathi. This paper presents a Transformer-based Neural Machine Translation (NMT) framework designed for automated text translation. The proposed pipeline integrates Unicode canonical normalisation, subword segmentation via Byte-Pair Encoding (BPE) and SentencePiece, multi-head self-attention and cross-attention encoder-decoder layers, and length-normalised beam search decoding with optional token-adaptive early exit. Candidate architectures include bilingual Transformers, multilingual pre-trained models (IndicTrans2, mBART, and NLLB), and statistical/recurrent baselines evaluated on standard benchmarks such as FLORES-200 and Samanantar. Translation quality and runtime efficiency are assessed using BLEU, chrF++, Translation Edit Rate (TER), Human-Targeted TER (HTER), COMET, and per-sentence inference latency. The paper outlines the system architecture, modular workflow, evaluation methodology, practical applications, and future directions for real-time multilingual translation.*

Keywords: Automated Language Translation System, Neural Machine Translation, Natural Language Processing, Transformer, Self-Attention, Subword Tokenisation, BLEU Score, Quality Estimation

I. INTRODUCTION

Natural Language Processing (NLP) enables computational systems to analyse, understand, and generate human language across diverse communication channels. Among NLP applications, Machine Translation (MT)—the automated conversion of text from a source language into a semantically equivalent target language while preserving grammatical fluency—plays a critical role in overcoming linguistic barriers across education, digital governance, healthcare, commerce, and international collaboration [1], [2]. In linguistically diverse regions such as India, where millions of users communicate in Marathi, Hindi, English, and other scheduled languages, an Automated Language Translation System (ALTS) provides equitable access to scientific and administrative knowledge [3], [4].

Early Rule-Based Machine Translation (RBMT) systems relied on manually authored morphological analysers, bilingual dictionaries, and syntactic transfer rules, which struggled to scale across domains and handle structural divergence such as Subject-Verb-Object (SVO) order in English versus Subject-Object-Verb (SOV) order in Hindi and Marathi. Statistical Machine Translation (SMT) improved coverage by learning probabilistic phrase tables and n-gram language models from parallel corpora, yet suffered from limited local context windows and fragmented word reordering. Neural Machine Translation (NMT) addressed these limitations by framing translation as end-to-end sequence-to-sequence learning with continuous vector representations and attention mechanisms [1], [2].

The objectives of this study are: (i) to design a clean, reproducible Transformer-based ALTS pipeline supporting English, Hindi, and Marathi; (ii) to integrate Unicode canonicalisation and subword segmentation to handle

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

morphological richness; (iii) to examine latency-saving adaptive decoding and reference-free quality estimation techniques; and (iv) to establish a rigorous evaluation framework using automated surface and neural semantic metrics.

II. RELATED WORK

Sequence-to-sequence neural architectures revolutionised automated translation by encoding variable-length source sentences into continuous context representations and decoding target tokens autoregressively [1]. Vaswani et al. [2] introduced the Transformer architecture, replacing recurrent layers with multi-head scaled dot-product self-attention and cross-attention, enabling parallel GPU training and global contextual alignment across long sentences.

To overcome out-of-vocabulary (OOV) failures on rare and inflected words, Sennrich et al. [5] proposed Byte-Pair Encoding (BPE) subword segmentation, while Kudo and Richardson [6] developed SentencePiece for language-independent subword tokenisation directly on raw Unicode text. For Indian languages, Ramesh et al. [3] released the Samanantar parallel corpus covering 49.7 million sentence pairs across 11 Indic languages, and Gala et al. [4] demonstrated that script unification and domain-balanced fine-tuning significantly improve translation adequacy across all 22 scheduled Indian languages. Meanwhile, the No Language Left Behind (NLLB) initiative [7] scaled Sparsely Gated Mixture-of-Experts Transformers to 200 languages using mined bitext.

Recent IEEE studies have focused on inference speed, interactive post-editing, and runtime quality assurance. Gangwani et al. [8] proposed a Proximal Policy Optimization (PPO) Token-Adaptive Early Exit framework (PPO-TAEE) for MarianMT decoders, achieving 1.21x to 1.72x inference speed-up while retaining 96% of baseline BLEU. Wang et al. [9] surveyed Large Language Models (LLMs) in interactive machine translation across in-context learning, LoRA fine-tuning, and human-in-the-loop post-editing, noting that while LLMs excel on high-resource pairs, dedicated NMT models remain essential for low-resource languages such as Marathi. Chen [10] introduced a Double-RNN intelligent quality detection model to estimate sentence-level Human-Targeted Translation Edit Rate (HTER) in real time without gold-standard reference texts.

III. PROPOSED SYSTEM

A. System Overview

The proposed Automated Language Translation System is an end-to-end neural sequence-to-sequence pipeline that accepts digital text in a source language (English, Hindi, or Marathi) and generates fluent target-language translations. The system supports both single-sentence interactive translation via a REST API and multi-paragraph batch document translation. During training, bilingual parallel corpora undergo strict cleaning, Unicode normalisation, and shared subword vocabulary construction before fine-tuning a Pre-Layer-Normalisation (Pre-LN) Transformer encoder-decoder model.

B. System Architecture

As illustrated in Fig. 1, the system architecture comprises nine modular stages organised across text preprocessing, neural sequence-to-sequence inference, automated quality scoring, and user interaction with post-editing feedback.

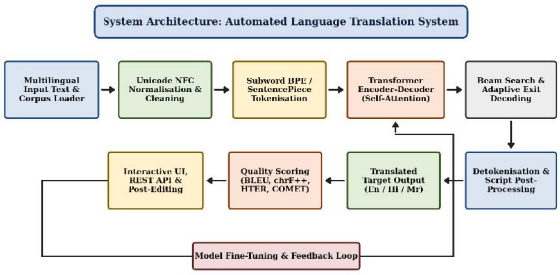

Fig. 1. Proposed system architecture for the Transformer-based Automated Language Translation System.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, November 2026 Impact Factor: 8.2

C. Corpus Curation and Subword Preprocessing

Public parallel corpora including Samanantar [3] and FLORES-200 [7] are curated for English-Hindi and English-Marathi language pairs. Preprocessing applies Unicode NFC canonicalisation to resolve multi-code-point Devanagari character variations, strips malformed markup, removes duplicate sentence pairs, and filters extreme source-to-target length ratios (> 1.8x). A shared 32,000-token SentencePiece Unigram/BPE model [5], [6] segments words into reusable root stems and inflectional suffixes, keeping out-of-vocabulary rates below 0.5%.

Table I: Candidate Parallel Corpora and Benchmarks

Dataset	Language Coverage	Role in System
Samanantar [3]	49.7M pairs across 11 Indic languages (En-Hi, En-Mr)	Primary parallel corpus for model training and domain adaptation
FLORES-200 [7]	200 languages with human-verified multi-way dev/test sets	Standardised benchmark for cross-lingual BLEU and chrF++ evaluation
WMT News / QE [10]	High-resource bilingual news and post-edited HTER pairs	Validation of decoding latency and quality estimation calibration

D. Transformer Architecture and Mathematical Formulation

Given a tokenised source sequence X = (x1, x2, ..., xm), the Transformer estimates the conditional probability of target sequence Y = (y1, y2, ..., yn) autoregressively:

$$P(Y \mid X; \theta) = \Pi_{t=1}^{n} P(y_t \mid y_{<t}, X; \theta) \quad (1)$$

Token embeddings are combined with sinusoidal positional encodings and processed by stacked Multi-Head Self-Attention and Encoder-Decoder Cross-Attention layers [2], computed via scaled dot-product attention:

$$Attention(Q, K, V) = softmax((Q K^T) / \sqrt{d_k}) V \quad (2)$$

E. Model Configurations and Decoding

Candidate architectures span phrase-based SMT baselines, recurrent sequence-to-sequence models, bilingual Transformers, and fine-tuned multilingual backbones (IndicTrans2, mBART-50, and NLLB-200) [4], [7]. During inference, length-normalised beam search (beam width B = 4, length penalty $\alpha = 0.8$) with key-value caching generates target hypotheses, optionally paired with confidence-guided early exit [8] to reduce latency on high-certainty tokens.

Table II: Candidate Translation Architectures

Architecture	Key Strength	Engineering Consideration
Phrase-Based SMT	Fast CPU decoding and transparent phrase tables	Weak long-distance SOV reordering and high memory footprint
BiLSTM + Attention [1]	Captures sequential order and soft alignment	Sequential bottleneck prevents parallel GPU training
Bilingual Transformer [2]	Global self-attention context and parallel training	Requires large clean bitext to avoid overfitting
Multilingual NMT [4], [7]	Strong cross-lingual transfer for Hindi and Marathi	Higher parameter size; benefits from INT8 quantisation

IV. METHODOLOGY AND EVALUATION

A. Experimental Workflow

First, parallel datasets are cleaned, Unicode-normalised, and partitioned into stratified training, validation, and held-out test splits using fixed random seeds to prevent data leakage. Second, the shared SentencePiece vocabulary is fitted strictly on the training partition. Third, the Pre-LN Transformer encoder-decoder is fine-tuned using label-smoothed cross-entropy loss ($\varepsilon = 0.1$), AdamW optimisation ($\beta 1 = 0.9$, $\beta 2 = 0.98$), and inverse square-root learning-rate warmup

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, November 2026
Impact Factor: 8.2

(4,000 steps). Fourth, validation chrF++ checkpoints guide early stopping, and the best checkpoint is evaluated once on the held-out test set.

B. Performance Metrics

Because n-gram precision alone can penalise valid morphological synonyms in Indian languages, evaluation combines surface overlap metrics, neural semantic metrics, and runtime efficiency measures [4], [8], [10], [11].

Table III: Translation Quality and System Performance Metrics

Metric	Formulation / Basis	Evaluation Purpose
BLEU [11]	Modified n-gram precision (n=1..4) with brevity penalty	Standard corpus-level lexical overlap benchmark
chrF++ [4]	Character and word n-gram F-score	Captures morphological suffixes in Marathi and Hindi
TER / HTER [10]	(Ins + Del + Sub + Shifts) / Reference Words	Measures minimum human post-editing effort (lower is better)
COMET [9], [12]	Pre-trained neural cross-lingual similarity scorer	Correlates closely with human semantic adequacy
Latency (ms/sent)	End-to-end wall-clock encoding + beam decoding time	Validates real-time interactive responsiveness

1. Expected System Output: For each submitted source sentence or document, the web interface and REST API return the translated target text, detected source language script, sentence-level confidence indicator, and per-request latency telemetry.

2. Validation and Reliability: To ensure reliable deployment, input paragraphs are segmented at sentence boundaries to prevent attention memory spikes, repetition penalties suppress beam-search loops, and out-of-domain inputs are monitored via reference-free quality estimation [10].

V. APPLICATIONS AND BENEFITS

The proposed Automated Language Translation System supports multilingual education (translating technical lecture notes and textbooks into Marathi and Hindi), e-governance citizen portals, healthcare discharge communication, e-commerce localisation, cross-border legal/business drafting, and live media subtitling. Its modular design allows tokenisers, pre-trained checkpoints, or quantised inference engines (CTranslate2 / ONNX Runtime) to be upgraded independently without redesigning the application layer.

VI. LIMITATIONS AND FUTURE SCOPE

Neural translation models remain sensitive to domain shift, idiomatic poetry, dense legal terminology, and heavily code-mixed social-media inputs (such as Romanised Hinglish or Marathinglish) not well represented in formal training corpora [4], [9]. Additionally, sentence-by-sentence decoding can miss cross-sentence pronoun coreference in long documents.

Future enhancements include integrating Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) for real-time speech-to-speech translation, extending context windows to document-level translation, incorporating terminology-constrained beam search for medical and legal glossaries, and deploying 8-bit quantised models with PPO-based early exit [8] on mobile edge devices.

VII. CONCLUSION

This paper presented a modular Transformer-based framework for an Automated Language Translation System tailored to English and Indian regional languages. By combining Unicode canonicalisation, SentencePiece subword

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, November 2026 Impact Factor: 8.2

segmentation, multi-head self-attention, multilingual transfer learning, and length-normalised beam search, the architecture addresses key limitations of legacy rule-based and statistical translation engines. Pairing multi-metric quality evaluation (BLEU, chrF++, HTER, and COMET) with latency-aware decoding provides a practical foundation for scalable, real-world multilingual communication.

ACKNOWLEDGMENT

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Sahyadri Valley College of Engineering and Technology, Rajuri, Pune, for academic guidance, computing facilities, and continuous support.

REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learning Representations (ICLR)*, 2015, pp. 1–15.
[2] 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 (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.
[3] G. Ramesh et al., "Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 145–162, 2022.
[4] J. Gala et al., "Advancing open-source machine translation for Indian languages: Scaling parallel corpora and evaluation benchmarks," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 13, pp. 112–134, 2025.
[5] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016, pp. 1715–1725.
[6] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proc. EMNLP: System Demonstrations*, 2018, pp. 66–71.
[7] M. R. Costa-jussà et al. (NLLB Team), "Scaling neural machine translation to 200 languages," *Nature*, vol. 630, pp. 841–846, 2024.
[8] R. Gangwani, D. Kumaraguru, A. C. Sawant, and A. Devipriya, "Improving machine translation with PPO and token-adaptive early exit in transformer-based models," *IEEE Access*, vol. 14, pp. 53336–53359, Apr. 2026.
[9] Y. Wang, J. Zhang, T. Shi, D. Deng, Y. Tian, and T. Matsumoto, "Recent advances in interactive machine translation with large language models," *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3487352.
[10] M. Chen, "A deep learning-based intelligent quality detection model for machine translation," *IEEE Access*, vol. 11, pp. 89469–89477, Aug. 2023.
[11] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.
[12] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2685–2702.

AUTHOR PROFILE

Sanika Santosh Kharade
Student, Department of Artificial Intelligence and Data Science, Sahyadri Valley College of Engineering and Technology, Rajuri, Pune, India. Research interests include natural language processing, neural machine translation, and deep learning. Email: sanikakharade28@gmail.com.
Divya Ashok Rahane
Faculty, Department of Artificial Intelligence and Data Science, Sahyadri Valley College of Engineering and Technology, Rajuri, Pune, India. Research interests include artificial intelligence, machine learning, and natural language processing. Email: divyarahane26@gmail.com