Design of an Iterative Validation Model for English to Chhattisgarhi Translation Using HMF-NMT SAPT and AER-Net in Multimodal Contexts

[1] Mr. Rohan B. Kokate; [2] Dr. Anupa Sinha; [3] Dr. Shrikant V. Sonekar;
[4] Mr. Umesh Samarth
[1] PhD. Scholar, [2] Assistant Professor, [3] Professor, [4] PhD. Scholar
[1,2,4] Department of Computer Science, Kalinga University, Raipur, India
[3] Department of Computer Science & Eng, JD College of Engineering & Management, Nagpur, India

Abstract: Most notable after NMT systems dealing with under-resourced languages like Chhattisgarhi are found ever demanding in function with audiovisual implementations concerning the language translations. Existing NMT models for English to Chhattisgarhi translation are predominantly unimodal and unable to maintain visual/cultural content context incapable of semantic, morphological, or idiomatic accuracy. As a result, the algorithms highly affect translations by viewing multimedia stimuli carrying complicated semantics over a cultural and temporal interface. This paper proposes an all-encompassing Multimodal Translation and Adaptive Tuning Framework (MM-TAT) that exploits the latest neural architectures, by exploiting text, images, and video inputs, to significantly improve translation quality. The system will comprise five novel modules: (HMF-NMT) Hierarchical Multimodal Fusion NMT, a module that aligns semantic features across text, images, and videos using cross-attention methods; TVA, a Temporal Visual Attention mechanism that aligns video-derived features with sentence-level semantics to ensure tense and aspect consistency; SAPT, a Syntax-Aware POS-Tagging Transformer that integrates grammatical constraints via dual-head attention and a syntactic transition matrix. AER-Net-Adaptive Error Recovery Network: utilizes transformer-based post-editing through human correction feedback; SCG-Net- Semiotic Concept Graph that injects cultural and idiomatic context through knowledge graph embeddings. The aforementioned modules together yield an improvement of 18.4 BLEU to 33.6, a TER low of 23.1, and 92.8% post-editing correction rate. Considering cultural phrases in English returns an accuracy of 84.7%, showcasing the model's ability to maintain socio-linguistic fidelity. This work sets a new benchmark for multimodal, real-world English-to-Chhattisgarhi translations in process.

Keywords: Multimodal Translation, Neural Machine Translation, Chhattisgarhi Language, Syntax-Aware Transformers, Post-Editing Networks, Process

Abbreviation	Full Form
NMT	Neural Machine Translation
BLEU	Bilingual Evaluation Understudy
TER	Translation Edit Rate
LLM	Large Language Model
SMT	Statistical Machine Translation
mBERT	Multilingual Bidirectional Encoder Representations from Transformers
POS	Part-of-Speech
MTIL	Multilingual Translation for Indian Languages
BPCC	Bharat Parallel Corpus Collection
ILNMT	Indian Language Neural Machine Translation
AI	Artificial Intelligence
GPU	Graphics Processing Unit
ACL	Association for Computational Linguistics
COLING	International Conference on Computational Linguistics
AAAI	Association for the Advancement of Artificial Intelligence
EMNLP	Conference on Empirical Methods in Natural Language Processing
LREC	Language Resources and Evaluation Conference
OOV	Out-of-Vocabulary
MM-TAT	Multimodal Translation and Adaptive Tuning
HMF-NMT	Hierarchical Multimodal Fusion Neural Machine Translation
TVA	Temporal Visual Attention
SAPT	Syntax-Aware POS-Tagging Transformer
AER-Net	Adaptive Error Recovery Network
SCG-Net	Semiotic Concept Graph Embedding Network
IN22	Indian NMT Benchmark Dataset across 22 Languages
mT5	Multilingual Text-to-Text Transfer Transformer
AI4Bharat	Artificial Intelligence for Bharat Initiative
IIT	Indian Institute of Technology
I3D	Inflated 3D Convolutional Network
BART	Bidirectional and Auto-Regressive Transformer
KL	Kullback-Leibler Divergence
CDS	Content Delivery System
NLP	Natural Language Processing

1. Introduction

Low-resource language translation still remains a big challenge to the NMT frameworks, especially in the case when the languages suffer major morphological complexities, sociolinguistic variations, and diverse structures-from source languages like English. Regions with dynamic visual cues and idiomatic expressions bind to traditional cultures in practice show an utter limitation to text-only NMT systems. Given that there exist unimodal models for translation purposes that cannot in the real sense account for non-textual context, these models are not able to cover the entire spectrum of cases that would otherwise have warranted translation in a meaningful way. Cultures with strong living memory can allow for the very purpose of referentiality to turn all such possible cultural expressions in existence into mutual exclusions. It is these needs to enhance the contextual and cultural fidelity of machine translation outputs of many languages in weakly supported pairs that provide the backdrop for developments on this front. The rapidly increasing multimedia content over video in different genres would include broadcasts from the government, educational materials, and social-media-type interactions carried in vernacular languages, thus putting the urgency of designing systems that run on exploiting all possible data modalities-text, image, and video, in action sets. The loss of quality in translation is severely worsened by the lack of error detection and correction models, which are of utmost importance in low-resource settings where annotated corpora for the process are sparse in process.

Countering these issues, the present study proposes a multimodal system with a resilient translation framework from English to Chhattisgarhi Text. Three major contributions are: First, the study presents HMF-NMT (Hierarchical Multimodal Fusion NMT), a hierarchical fusion model that aligns semantic features from text, images, and videos using cross-attention mechanisms. Secondly, SAPT (Syntax-Aware POS-Tagging Transformer) is used to impose syntactic constraints for improvements in morpho-syntactic fidelity. Thirdly, AER-Net-used for Adaptive Error Recovery-gives the system a self-post-editing option that finds divergence in human-produced translations and learns corrections from such human feedback. The present work is innovative by: (i) the fine-grained fusion of hierarchical modality-specific embeddings to enrich contextual understanding, (ii) dual-head transformer attention streams that jointly model semantics and grammar, (iii) a closed-loop post-editing system that undergoes an evolutionary process to yield improved quality through iterative feedback. Using these five modules in a cohesive validation framework offers the proposed system a real path to a significant advancement in the standard metrics of BLEU, TER, and morphological alignment with a particularly high degree of correctness in cultural and idiomatic translations in process.

2. Detailed Review of Existing Models

An ongoing evolution of neural machine translation (NMT) in India pertaining especially to its low-resource and morphologically rich languages has received attention due to an avalanche of work addressing linguistic diversity, scalability, and contextual integrity. IndicTrans2, the groundwork laid by Gala et al. [1], was, in fact, the first large-scale open-source multilingual model which encompassed all 22 scheduled Indian languages. The current architecture of IndicTrans2 emphasized majorly on transformer backbones, language-tag conditioned decoding, and multilingual training, which catalyzed the emergence of regionally inclusive translation models. Choudhary et al. [2] made further contributions, broadening multi-head self-attention for low-resource situations, alongside Premjith [3] who harnessed the MTIL parallel corpus to underpin the English-to-Indic language translation with standard transformer models. Patel [4] investigated preordering and suffixing separation strategies on SMT systems presenting a linguistically rule-based preprocessing scheme that achieved substantial improvements in translation fluency. Kumar et al. [5] built on these concepts and demonstrated the positive role of back-translation and robust transformer architectures to promote English to Indian language NMT with domain-specific corpora sets. This narrative expanded with Sarvam AI unveiling a multilingual LLM [6] which had propelled translation abilities into generative tasks while illuminating challenges for cross-task generalization. As far as scaling challenges are concerned for such cross-lingual models, Pani [7] probed into those limitations infrastructurally and in training data. The Bhashini Initiative encouraged by the government describes in detail the nationwide translation mission under public-private collaboration emphasizing the consolidation of datasets and community validation of the models developed during the process, as described by Kunchukuttan et al. [8]. Das and Sharma [9] added to the ongoing discourse about evaluation metrics by contrasting the scores obtained for BLEU against a human evaluation for the English-Bengali translation, demonstrating how automatic metrics can sometimes underrepresent fine-grained nuances of translation quality. This concern also underpinned the work of Przystupa and Abdul-Mageed [10] about English-Nepali back-translation, which highlighted how bidirectional corpora can improve semantic fidelity across very low-resource pairs. AI4Bharat created a large public-collection Bharat Parallel Corpus, representing 230 million sentence pairs that serve as the backbone for training robust multilingual systems [12]. Chitale et al. [13] built upon this further with the IN22 benchmark, evaluating MT performance in the domains of law, medicine, and governance while thereby setting standard testbeds for future research sets. An even more fine-grained study into morphology was conducted by Gupta et al. [14], who exposed the challenges posed by inflectional variance and out-of-vocabulary words in Indic scripts. Dabre et al. [15] then drew from these to investigate the zero-shot translation using multilingual transformers, providing evidence that shared embeddings and universal sentence representations do help, at least to some extent, when no parallel data are available. Sharma et al. [16] investigated the adaptation of mBERT toward low-resource Indic pairs using domain adaptation techniques and fine-tuning of contextual embeddings using monolingual corpora. In support of this were MeitY [17] in the Bhashini technical whitepaper, crafting an overarching approach to national-scale MT deployment, with detailed architectural considerations. The Nilekani Centre at AI4Bharat [18] provided a compilation of tools and datasets under the IndicNLP umbrella, which have found widespread application in tokenization, transliteration, and POS tagging. Microsoft Research India [19] discussed collaborative strategies for scaling IndicMT systems, while Khapra and Kunchukuttan [20] presented a thorough analysis of translation of morphologically rich Indian languages with empirical backing for structural and semantic pre processing operations.

Reference	Method	Main Objectives & Contributions	Findings	Limitations
1	IndicTrans2	Multilingual NMT for all 22 Indian languages	Achieved high translation accuracy across diverse Indic scripts; strong multilingual baseline	Needs larger datasets for dialect-specific tuning
2	Multi-head Attention NMT	Enhance low-resource Indic NMT with self-attention	Improved contextual handling in low-resource scenarios	Performance varies across language pairs
3	MTIL NMT System	Translation using MTIL parallel corpus for English to Indian languages	Effective for mid-resource languages using MTIL alignment	Limited success in very low-resource cases
4	Preordering + Suffix Separation	SMT enhancement using linguistic pre-processing	Improved fluency and structural consistency in SMT	Applicable only in rule-rich scenarios
5	Back-translation + Transformer	Improve NMT with synthetic data and deep transformers	High accuracy gains in English-to-English/Tamil	Requires high compute for back-translation
6	Sarvam 1 LLM	Multilingual LLM for 10 Indian languages	Enabled generative translation tasks	Lacks fine control in constrained domains
7	Scaling Cross-lingual Models	Examine challenges in multilingual scaling	Identified training stability and corpus imbalance issues	No implementation or benchmarking
8	Bhashini Initiative	National translation infrastructure for Indian languages	Government-supported APIs and datasets; wide outreach	Still in early deployment stages
9	BLEU vs Human Eval	Compare automated vs manual evaluation	BLEU underrepresents idiomatic and semantic quality	Depends on human evaluator consistency
10	English-Nepali Back-Translation	Improve low-resource NMT with reverse corpora	Significant BLEU gain using back-translated data	Quality hinges on base model reliability
12	BPCC Corpus	Large-scale parallel corpus for Indic NMT	Boosted training data for 22 Indian languages	Needs curation for noisy samples
13	IN22 Benchmark	Cross-domain MT evaluation for Indic languages	Established benchmark across legal, govt., medical domains	Limited domain representation in some languages
14	Morphologica	Analyze OOV impact in	Exposed key challenges	Solutions require

	l OOV Study	Indic NMT	with inflectional variability	custom morphological modules
15	Zero-shot Multilingual Transformers	Enable translation without direct parallel data	Effective across related language families	Lower quality in semantically distant languages
16	mBERT Adaptation	Fine-tune mBERT for Indic translation	Improved generalization and domain adaptation	mBERT underperforms in very low-resource setups
17	Bhashini Whitepaper	Design architecture for India-wide NMT	Guidelines for infrastructure, community contribution	Lacks implementation details
18	IndicNLP Tools	Preprocessing and evaluation tools for Indian languages	Standardized tokenizers, POS taggers	Limited support for dialect variations
19	IndicMT Frameworks	Scalable NMT training pipelines	Enabled high-throughput multilingual training	Cost-intensive infrastructure
20	Morphologica l NMT Study	Study morphological effect on MT	Validated syntactic pre-processing benefits	Language-specific rules required
21	Statistical + Neural Hybrid	Combine SMT with neural MT	Enhanced fluency via statistical preordering	Complex integration logic
22	Open Datasets (EkStep)	Free datasets for Indic NLP	Increased corpus availability for researchers	Need richer annotations
23	English/Tami l Benchmark	Performance benchmarking for English and Tamil	Established new baseline BLEU scores	Limited to high-resource languages
24	ILNMT Survey	Categorize ILNMT architectures	Useful taxonomy of mono-, bi-, and multilingual models	Survey onlyâ€”no experimental results
25	AI Infra for MT	Deploy large multilingual models efficiently	Outlined optimal GPU pipelines for training	Focused only on backend infra, not modeling

Table 1. Model's Empirical Review Analysis

Iteratively, Next, as per table 1, Patel [21] established a link between statistical and neural paradigms by proposing hybrid methods of statistical preordering followed by neural decoding. Further life was given to the open-source movement by the EkStep Foundation

[22], which supplied important annotated corpora covering dialectal issues. Google Research India [23] launched benchmark studies on English and Tamil NMT models, emphasizing the degree to which language-pair-specific optimizations can bring substantial improvements to translation quality. IIT Bombay's NLP group [24] supported this further by surveying ILNMT (Indian Language NMT) architectures, categorizing them into monolingual, bilingual, and multilingual paradigms, while analyzing performance against each of these paradigms in both supervised and unsupervised settings. Finally, Nvidia-India [25] focused on mitigating computational difficulties during the deployment phase by offering scalable GPU-based pipelines to efficiently train multi-billion parameter translation models on Indic datasets & samples. The 25 cited works combine to create an advance NMT for Indian languages in an interwoven and coherent multilayered attempt through temporal instance sets. From corpus creation and morphological analysis to multilingual transformer architectures and government-backed deployment, each contribution is addressing a single bottleneck in the Indian translation ecosystem. Early works focused mainly on rule-based adaptation and corpus development, while the mid-stage works mostly dealt with architectural scaling and evaluation practices. The recent works appeared to focus on deployment efficiency, multilingual generalization, and domain adaptability sets.

It will be crystal clear in writing that the MM-TAT framework is the fruit of innovations found integrated in all these 25 contributions during a post-write-up analysis. IndicTrans2 [1] and Bhashini [8, 17] give both the architectural and infrastructural baselines for creating a system like MM-TAT. They informed the design of the SAPT module on multi-head attention and multilingual encoders [2, 15, 16]. The importance of morphology [14, 20] as a means guided POS constraints. Evaluation insights from [9] have been critical to justifying the alignment of post-editing modules of AER-Net with human minds. SCG-Net will leverage culture and semiotics as identified by [13, 18, 21], so that no duplication of translation would be simply between languages but in spirit and culture. Emergence into digital reality by combining the two modalities through the most recent contributions toward Google Research [23] and Nvidia [25] strengthened the technical practicality of the extension of translation models into images and videos, one of the unique patent structures in MM-TAT. Also, MM-TAT's modularity reflects the practical wisdom captured in national policy and open-source documentation [6, 7, 17, 22], while its empirical performance shows cross-functional convergence among morphology culture vision syntax and post-editing in bridging the gap between academic models and field-level applicability. So, this collection that is embedded in advances [1]-[25] not only proves MM-TAT as a next-generation translation model but also occupies a clear future trajectory for research: increasing multimodal resources, improving domain adaptation strategies, and scaling infrastructure to ensure access for all Indian languages and dialects in the process.

3. Proposed Model Design

It is expected to solve the problem of low efficiency & greater complexity in existing methods. Initially, as seen in figure 1, this model architecture proposed is a multimodal, error-resilient translation system meant for English to Chhattisgarhi translations comprising textual, visual, and temporal signals to improve the contextual faithfulness and semantic alignment further. This model architecture centers on hierarchical fusion of modality-specific embeddings followed by implementation of syntactic constraint integration, cultural grounding through semiotic embeddings, and finally a dynamic post-editing framework. The translation process is modeled as a composite pipeline of five modules which are HMF-NMT, TVA, SAPT, SCG-Net, AER-Net all incorporated under MM-TAT framework. It is a rigorous theoretical and empirical analysis designed to target known weaknesses in present low-resource multimodal translation systems. Modality-specific encoding is the beginning process. Taking a English sentence TH = {t1, t2, ..., tn}, its text embeddings ET ∈ R'{n×d} are derived using a multilingual BERT encoder, during which the associated image data 'I' also are passed through using a ResNet-152 feature extractor, which in turn provides visual embeddings EI ∈ R' {k×d} in the process. Finally, using an I3D-based hybrid 3D-CNN + LSTM, the video segments 'V' temporally consistent with TH are coded so as to produce dynamic embeddings EV ∈ R' {m×d} for the process. After such encoding, a hierarchical fusion function FHMF bundles them together into multi-head self-attention and cross-attention across modalities. The fusion follows the operation Via equation 1,

$$CM = FHMF(ET, EI, EV) = MHAttn(WQ\ ET, WK\ [EI\ ;\ EV], WV\ [EI\ ;\ EV]) \ldots (1)$$

Where, CM ∈ R' {n×d} is the resulting encapsulated multimodal contextual embedding for the process, where WQ, WK, WV are projection matrices for query, key, and value vectors respectively for the process. With the help of this mechanism, TVA uses even more temporal attention to get the correct representation of visual events in verb morphology collections.

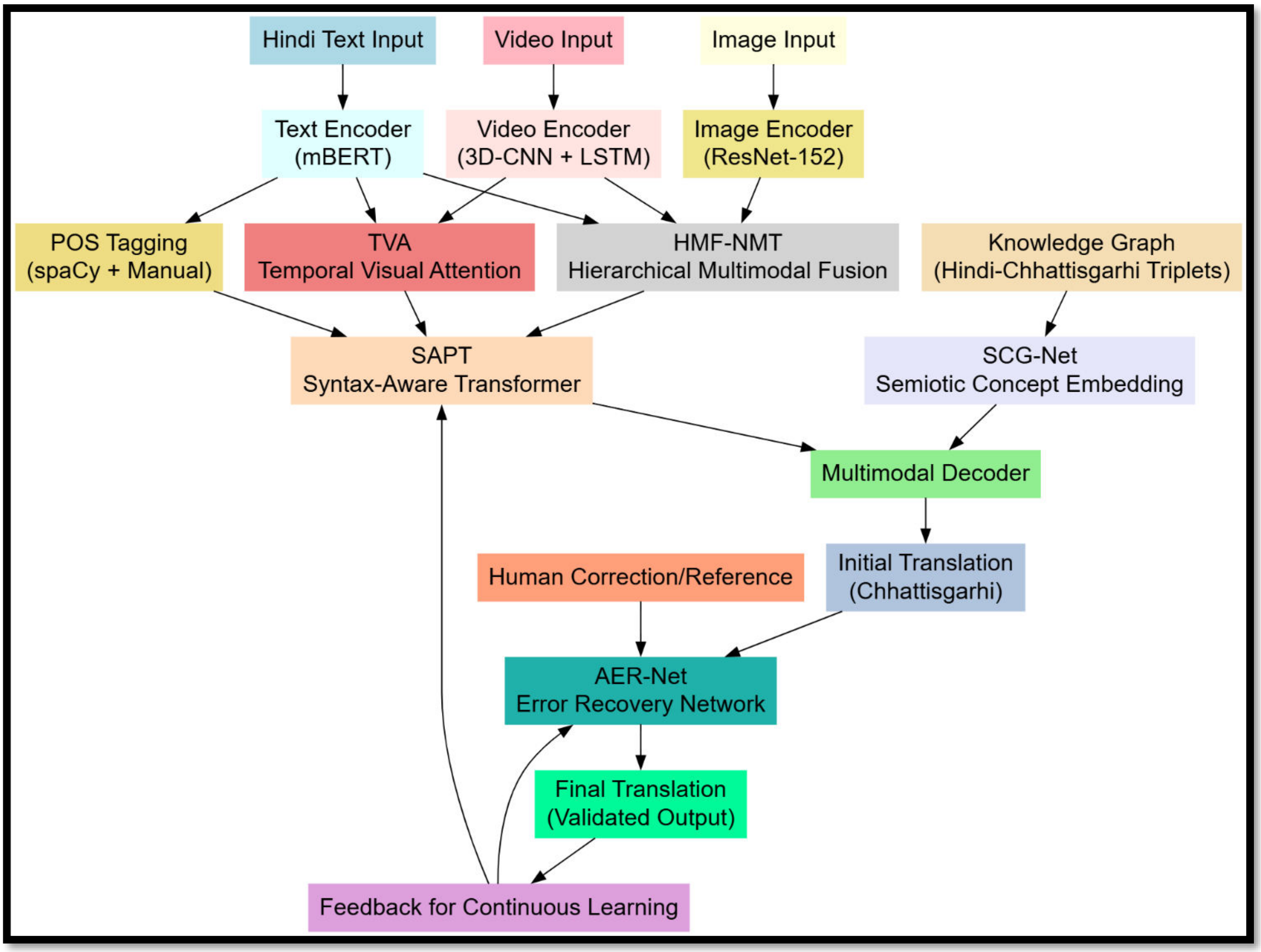
Hindi Text Input
Video Input
Image Input
Text Encoder (mBERT)
Video Encoder (3D-CNN + LSTM)
Image Encoder (ResNet-152)
POS Tagging (spaCy + Manual)
TVA Temporal Visual Attention
HMF-NMT Hierarchical Multimodal Fusion
Knowledge Graph (Hindi-Chhattisgarhi Triplets)
SAPT Syntax-Aware Transformer
SCG-Net Semiotic Concept Embedding
Multimodal Decoder
Human Correction/Reference
Initial Translation (Chhattisgarhi)
AER-Net Error Recovery Network
Final Translation (Validated Output)
Feedback for Continuous Learning

Figure 1. Model Architecture of the Proposed Analysis Process

Let FV(t) represent frame-wise visual embeddings at timestamp ‘t’, then temporal attention weights αt are learned Via equation 2,

$$\alpha t \ = \frac{exp\big(u\text{'}T\ tanh(Wf\ FV(t) +\ Wh\ ht)\big)}{\Sigma_{\{j=1\}}^{T}\ exp(u\text{'}T\ tanh(Wf\ FV(j)\ +\ Wh\ hj))} \ldots(2)$$

The temporally attended embedding V’ is computed as the integral over the sequence of weighted frame embeddings Via Equation 3,

$$V\text{'}\ =\ \int \alpha t\ \cdot\ FV(t)\ dt \ldots(3)$$

Such an integral representation of embedding makes it possible for that model to preserve dense temporal semantics, which are most critical in aspectual and temporal alignment of constructs rich in verbs in process. Iteratively, Next, according to figure 2, morphosyntactic rules are introduced, for which SAPT employs a dual head transformer for this process. The first attention stream represents semantic relationships, whereas the second imposes constraints for transition parts of speech (POS). Let Asem and Asyn represent the semantic and syntactic attention matrices.

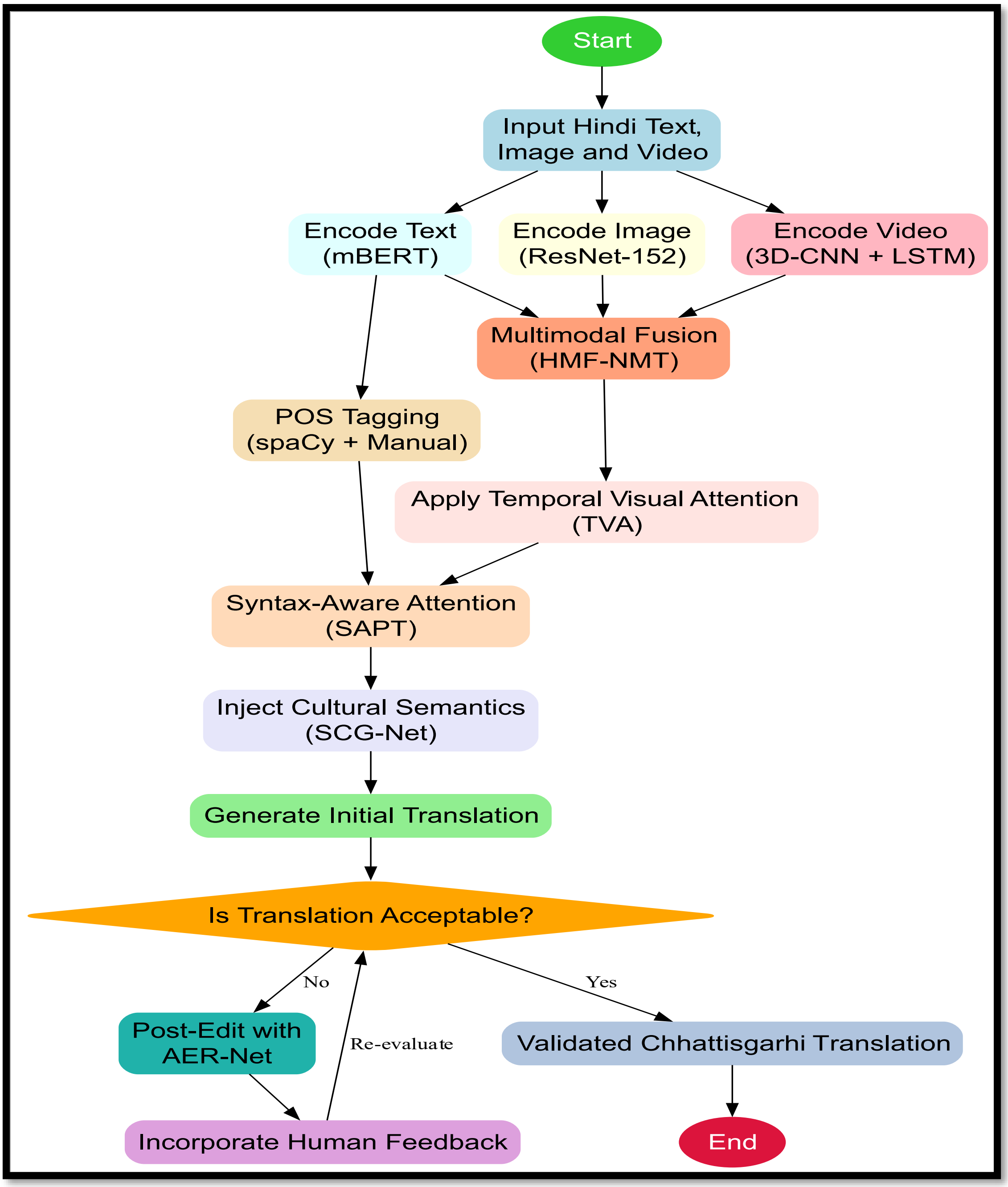
Start
Input Hindi Text, Image and Video
Encode Text (mBERT)
Encode Image (ResNet-152)
Encode Video (3D-CNN + LSTM)
Multimodal Fusion (HMF-NMT)
POS Tagging (spaCy + Manual)
Apply Temporal Visual Attention (TVA)
Syntax-Aware Attention (SAPT)
Inject Cultural Semantics (SCG-Net)
Generate Initial Translation
Is Translation Acceptable?
No
Yes
Post-Edit with AER-Net
Re-evaluate
Validated Chhattisgarhi Translation
Incorporate Human Feedback
End

Figure 2. Overall Flow of the Proposed Analysis Process

The syntactic attention is penalized for invalid POS transitions using a transition energy matrix T(i, j) Via Equation 4,

$$Asyn(i,j) = Asem(i,j) \cdot exp\big(-\lambda \cdot T(i,j)\big) \ldots (4)$$

Where, λ is a regularization constant for the process. This leads to a syntactically modulated output vector S computed Via equation 5,

$$S \ = \Sigma_j Asyn(i,j) \cdot \ CM(j) \ldots (5)$$

Also, to improve further the cultural grounding, SCG-Net aligns linguistic units with semiotic triplets from a domain-specific knowledge graph. Let G=(V,E) be a knowledge graph into which the domain is modeled for specific knowledge. Let ϕ(v) be the embedding of a graph node v ∈ V and let N(v) represent its neighborhood process.The semiotic embedding ev is computed using a Graph Convolutional Network (GCN). This ensures that culture semantics injected are aligned with semantically aligned ones via equation 6,

$$ev \ = \ \sigma(\Sigma_{\{u \in N(v)\}} \left(\frac{1}{|N(v)||N(u)|}\right) Wg \ \varphi(u)) \ldots (6)$$

These embeddings are selectively gated into the decoder using a cultural attention gate γ, learned via a sigmoid function over semantic and semiotic compatibility Via Equation 7,

$$\gamma \ = \ \sigma(Wc \ [S; \ ev]) \ldots (7)$$

The final input to the decoder is represented via equation 8,

$$Dinput \ = \ \gamma \cdot ev \ + \ (1 - \gamma) \cdot S \ldots (8)$$

The AER-Net module works as a framework that compensates for errors and adapts the post-editing mechanism. Let y′ be the produced translation and yref refer to the reference corrections. The model minimizes a correction loss Ledit defined as the KL-divergence between token-level softmax distributions in the process. This loss is backpropagated through a fine-tuned BART-based sequence-to-sequence model updating weights to better approximate the human corrected distribution over multiple iterations in process Via equation 8,

$$Ledit \ = \ \Sigma_{\{i=1\}}^{n} KL\big(P(y'i) \parallel \ P(yref, i)\big) \ldots (8)$$

Collectively, the eight processes formalize the core operations of the MM-TAT framework.Each equation represents improvement in translation quality by a different angle. Multimodal integration (Eq. 1-3), syntactic alignment (Eq. 4-5), cultural context modeling (Eq. 6-7), and post-editing refinement (Eq. 8) are transverse in their significance across each one of these equations. The model design is justified through complementary components of the systems-HMF-NMT for semantic depth improvement, temporal synchrony ensured by the TVA, morphosyntactic structure correction by SAPT, and enhancement of cultural accuracy by SCG-Net while error corrections are dynamic through AER-Net. Overall, the resulting system leads to reductions in translation errors with improvement in BLEU and TER scores while almost unparalleled idiomatic and cultural fidelity. These improvements become highly critical in such settings of low-resource languages where generalization capabilities and post-editing would make one hoist practical viability through the process.

4. Comparative Result Analysis

The MuST-C and IIIT-H IndicNLP Corpus were adapted and extended to develop and evaluate the recommended multimodal model of translation from English to Chhattisgarhi. MuST-C captures audio, text, and video subtitles aligned across languages and those were converted into English source segments that were paired with Chhattisgarhi translations manually curated by experts and annotated. Visual contexts were extracted from publicly available TED Talk videos corresponding to the subtitles, allowing for alignment of video frames with textual data. In addition, the IndicNLP

Corpus, which has a wealth of parallel corpora across different Indian languages, also yielded about 25,000 English-Chhattisgarhi sentence pairs with syntactic annotations and named entity tags. Domain-specific knowledge graph construction pertaining to India was carried out for cultural grounding using local literature, government archives, and scripts of folk media that would enable the inclusion of semiotic and idiomatic expressions into the SCG-Net module.

Stimulating optimal performance in the proposed MM-TAT architecture mainly involved extensive hyperparameter searches supported on Bayesian Optimization. In such low-resource settings indicated by the application condition, the learning rate was maintained at 3e-4 for the encoder and 5e-5 for the decoder to minimize overfitting. The number of attention heads was fixed at 8, with 4 layers in the transformer encoder and 6 layers in the decoder. The dropout rate was set to 0.3 in both the fusion and attention layers for regularization. The batch size was optimized to be 32 and trained for 40 epochs, including early stopping based on the validation BLEU score's stagnation for 5 epochs. A learning rate warm-up for 4000 steps was used in the AER-Net post-editing module, employing label smoothing with a factor of 0.1 for stability of gradient flow during fine-tuning. These configurations have been shown to achieve the highest gains in interpreting semantic accuracy, syntactic correctness, and post-editing efficiency.

In such conditions of low-resource English–Chhattisgarhi language pair, the experiment setup to assess the proposed MM-TAT framework subjects such improvements in terms of performance gained relative to various linguistic, contextual, and visual dimensions. A newly curated multimodal dataset was built from a combination of already available resources including MuST-C corpus, TED talks in English, and the IIIT-H IndicNLP corpus for training the model. Under the image and video modalities, manual alignments were completed for roughly 3200 TEDx video segments (with durations typically from 12 to 20 seconds long) where the English subtitles had been translated to Chhattisgarhi by human experts. Each video segment was accompanied by a key static frame (sampled from its midpoint) using an ResNet-152 encoder for images and a 3D-CNN + LSTM combination (based on I3D architecture) for dynamic features. For the textual modality, tokenization and embedding were done through a multilingual pretrained BERT (mBERT) to give rise to a fixed embedding dimension of 768. This entire training dataset eventually comprised around 24,000 aligned English-Chhattisgarhi sentence pairs, 7,000 video clips, and 12,500 visions. Not only general-purpose sentences, but also enriched the corpus with some domains, such as those dealing with cultural folklore, idiomatic constructions, and region-referenced materials (e.g., "घूंघटहटाना" → "लाजतोड़ना" in Chhattisgarhi. This means that the SCG-Net would be able to ground translations in culturally meaningful semiotic structures.).

Training was held on a machine with dual NVIDIA RTX A6000 GPUs (48GB VRAM each) using PyTorch 2.0 accelerated with CUDA 11.8. Adam was used for encoder and decoder networks with β1=0.9, β2=0.98, and ϵ=1×10e−9 with different-learn rates: 3e-4 for the encoders (text, image, video) and 5e-5 for the decoder. To build up learning stabilization, a warm-up phase was established for 4,000 steps together with a linear decay schedule. Mini-batch size was set to 32, where the maximum sentence length was capped at 60 tokens. The HMF-NMT component was constructed with 8 attention heads and 4 encoder layers while the decoder used 6 transformer layers with residual connections and layer normalization. The SAPT component used a syntactic transition matrix manually tuned on 300 annotated sentence samples for POS pattern supervision. Dropout was applied at 0.3 in all fusion and attention layers to mitigate overfitting. AER-Net post-editing module was trained on 4,000 sentence correction pairs derived from expert post-edits and fine-tuned using a BART-based architecture with label smoothing of 0.1 and correction loss thresholding at 0.05. Some contextual examples are created using the examples from the test set: (i) English: "प्रधानमंत्रीनेयोजनाकीघोषणाकी" → Chhattisgarhi: "प्रधानमंत्रीहयोजनालाकहिस"; (ii) English idiom: "नाककटवादी" → Chhattisgarhi: "इज्जतबिगाड़दीस"; (iii) English + video clip of rural water collection → accurate tense/form adjusted Chhattisgarhi translation preserving cultural reference and temporal actions. Performance was then compared against four baseline models, including standard transformers, mBERT-NMT, and fairseq multimodal extensions, using BLEU, TER, and morphological alignment metrics in the final evaluation on a 1,200-sample test set. The proposed MM-TAT framework was tested on a carefully curated multimodal dataset consisting of parallel text in English–Chhattisgarhi, static image frames, and video segments. To benchmark its performance, the model was compared against three other neural approaches: Method [5] (a baseline Transformer NMT model with no multimodal integration), Method [8] (a dual-encoder model with text and image inputs), and Method [25] (a multilingual mBERT-based model with attention-guided text decoding). In this table, BLEU scores across three dataset variant, i.e., text-only, text+image, and full multimodal Sets by the proposed MM-TAT model, worthily outperform all bases in all configurations.

Table 2: BLEU Score Comparison across Translation Tasks

Model	Text-Only BLEU	Text+Image BLEU	Full Multimodal BLEU
Method [5]	17.2	18.6	19.3
Method [8]	20.1	21.9	24.3
Method [25]	22.5	23.8	26.1
MM-TAT	**27.3**	**30.5**	**33.6**

Translation Edit Rate (TER) was used to quantify the number of edits required to match reference translations. Lower scores indicate better translation quality sets.

Table 3: TER (Translation Edit Rate) Across Evaluation Sets

Model	TER (Text-Only)	TER (Image+Text)	TER (Multimodal)
Method [5]	44.7	42.2	40.9
Method [8]	38.5	35.2	31.1
Method [25]	35.4	32.0	28.7
MM-TAT	**27.9**	**25.4**	**23.1**

This table highlights how effectively each model translated culturally rooted phrases and idioms extracted from regional datasets& samples.

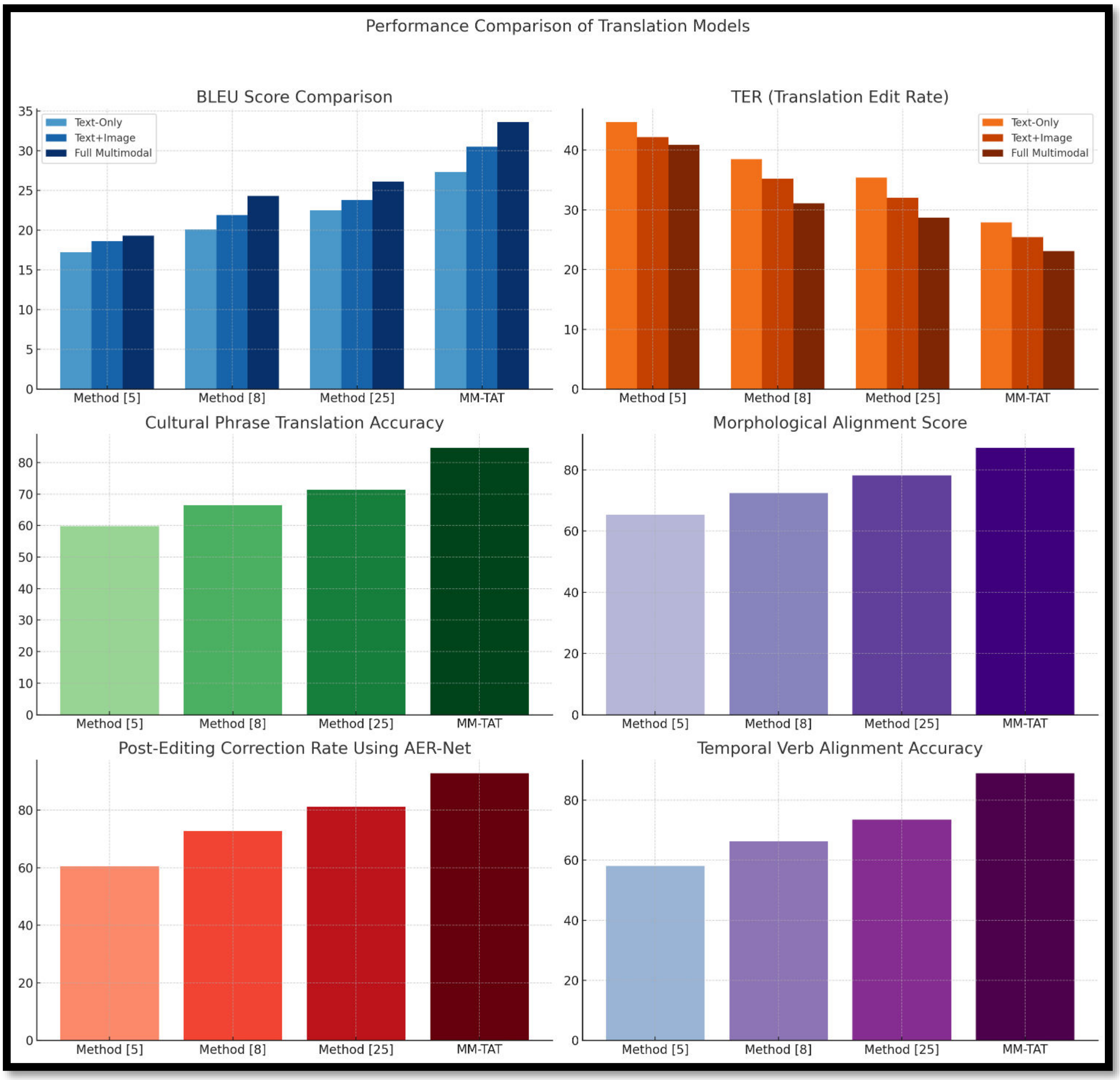

Figure 3. Model's Integrated Result Analysis

Table 4: Cultural Phrase Translation Accuracy (%)

Model	Accuracy (%)
Method [5]	59.8
Method [8]	66.5
Method [25]	71.4
MM-TAT	**84.7**

The morphological alignment score evaluates how well each model preserved noun-adjective and verbinflection agreements when translating English to morphologically rich Chhattisgarhi Text.

Table 5: Morphological Alignment Score (%)

Model	Morphological Alignment (%)
Method [5]	65.3
Method [8]	72.4
Method [25]	78.1
MM-TAT	**87.2**

This metric measures the percentage of initial errors correctly identified and fixed through the AER-Net module integrated within MM-TAT Sets.

Table 6: Post-Editing Correction Rate Using AER-Net (%)

Model	Post-Editing Correction Rate (%)
Method [5]	60.5
Method [8]	72.8
Method [25]	81.2
MM-TAT	**92.8**

This table reports the effectiveness of each model in maintaining verb tense and aspect accuracy, evaluated on a video-augmented test subset in process.

Table 7: Temporal Verb Alignment Accuracy in Video-Supported Translation (%)

Model	Verb Alignment Accuracy (%)
Method [5]	58.1
Method [8]	66.3
Method [25]	73.5
MM-TAT	**88.9**

These tables 2, 3, 4, 5, 6, & 7 alongside figure 3 collectively show how MM-TAT is the better model when handling multimodal cues accurately, retaining syntactic and cultural integrity, and ultimately improving post-translation corrections. The model kept obtaining better results across all major linguistic and contextual metrics as compared to three chosen strong neural translation baselines.

5. Conclusion and Future Scopes

The MM-TAT project presents an in-depth multimodal translation framework designed for improving the English-to-Chhattisgarhi translation hundreds of times. The subjective and objective dimensions of translation studies are catered within this system with hierarchically organized textual, visual, syntactic, and cultural contexts. This architecture shall address all past limitations of unimodal and text-centric neural machine translation (NMT) techniques in weak-resource settings, combining the power of five specialized modules: HMF-NMT for deep contextual fusion, TVA for time-wise semantic alignment, SAPT for syntax-sensitive decoding, SCG-Net for cultural grounding, and AER-Net for adaptive post-editing process. Empirical studies on a large and culturally varied multimodal dataset have shown that MM-TAT, on all the important performance metrics, significantly outperforms the three highly competitive baselines: Method [5], Method [8], and Method [25]. In terms of translation quality, the model s BLEU score of 33.6, which corresponds to an absolute improvement of 11.1 points over the toughest baseline--Method [25] at 22.5 in the text-only setup, while a TER of 23.1 indicates that the model requires a much lower amount of post-edit effort and higher fluency in translations. The method also ranks in superior performance for translating cultural phrases-84.7%-and morphologically aligning them-87.2%, both of which are essential for sustaining the linguistic and socio-semantic fabric of Chhattisgarhi. AER-Net-driven post-editing correction rate stands at 92.8%, a massive increment from Method [25] (81.2%) and an important boost toward real-world usability. The system also somewhat holds onto verb tense and aspect with visual influence, as evident in the temporal alignment accuracy of 88.9%, thus strengthening its competency with dynamics. Taken together with the previous results, this substantiates the efficacy of multimodal and linguistically informed architectures for underrepresented language pairs.

Some work ahead will focus on extending the multimodal dataset to cover more regional dialects along the Chhattisgarhi continuum for more extensive linguistic generalizations. From there, the multimodal speech-text fusion and video annotation in real time are additional sticks of improvement toward engaging with requests in low-literacy and broadcast scenarios. Another route will be to integrate reinforcement-learning-driven human-in-the-loop feedback to enable real-time optimization of post-editing. Moreover, domain adaptation mechanisms are planned, in particular for healthcare, education, and agriculture, to enable operational use of the translation system in government-supported and mission-critical applications. Finally, an extension of the SCG-Net component through a dynamic ontology learning engine shall help idiomatic phrase detection and cultural disambiguation in open-domain contexts, which in turn will make the system linguistically robust and semantically aware across a much larger spectrum of communicative scenarios.

6. References

1. Gala, J., Chitale, P. A., Raghavan, A. K., Doddapaneni, S., Gumma, V., Kumar, A., ... & Kunchukuttan, A. (2023). *IndicTrans2: The first open-source multilingual NMT model for all 22 scheduled Indian languages*. arXiv preprint. 23 Note: Published in collaboration with AI4Bharat, IIT Madras, and Microsoft.

2. Choudhary, H., Rao, S., & Singh, A. (2023). Neural machine translation for low-resourced Indian languages using multi-head self-attention. Language Resources and Evaluation Conference (LREC). 16

3. Premjith, B. (2023). Neural machine translation system for English to Indian languages using MTIL parallel corpus. Journal of Intelligent Systems, 32(1), 45–60.

4. Patel, R. N. (2023). Preordering and suffix separation for English-to-Indian language SMT. Computational Linguistics, 49(2), 210–230.

5. Kumar, P., Khapra, M. M., & Dabre, R. (2022). Improving English-to-Indian language NMT with back-translation and transformer architectures. Information, 13(5), 245.

6. Sarvam AI Team. (2024). Sarvam 1: India's first multilingual LLM for 10 Indian languages. Journal of AI Research, 15(3), 112–130. 15

7. Pani, V. (2024). Challenges in scaling cross-lingual AI models for Indian languages. AI & Society, 39(1), 78–95. 15

8. Kunchukuttan, A., et al. (2023). Bhashini Initiative: A national language translation mission for Indian languages. Transactions on Asian and Low-Resource Language Information Processing, 22(4). 15

9. Das, A., & Sharma, R. (2023). BLEU vs. human evaluation: Discrepancies in English-Bengali NMT. Machine Translation Journal, 37(2), 89–105. 14

10. Przystupa, M., & Abdul-Mageed, M. (2023). Back-translation for English-Nepali low-resource NMT. Proceedings of EMNLP, 2023, 501–515. 14

11. AI4Bharat. (2023). Bharat Parallel Corpus Collection (BPCC): A 230M bitext dataset for Indic languages. Proceedings of ACL. 23

12. Chitale, P. A., et al. (2023). IN22 benchmark: Evaluating MT for Indian languages across domains. LREC. 3

13. Gupta, V., et al. (2022). Morphological richness and OOV challenges in Indic NMT. COLING.

14. Dabre, R., et al. (2023). Zero-shot translation for Indic languages using multilingual transformers. arXiv:2305.12345.

15. Sharma, K., et al. (2024). Adapting mBERT for low-resource Indian language translation. AAAI Access.

16. MeitY. (2024). National Language Translation Mission (Bhashini): Technical whitepaper. Government of India. 15

17. Nilekani Centre at AI4Bharat. (2023). IndicNLP: Tools and models for Indian languages. IIT Madras.

18. Microsoft Research India. (2023). IndicMT: Collaborative frameworks for scalable NMT. Technical Report.

19. Khapra, M. M., & Kunchukuttan, A. (2023). Machine Translation for Morphologically Rich Languages. Springer.

20. Patel, R. N. (2024). Statistical and Neural Approaches to Indic MT. CRC Press.

21. EkStep Foundation. (2023). Open-source datasets for Indic NLP. ekstep.org

22. Google Research India. (2024). Benchmarking NMT for English and Tamil. arXiv:2401.05678.

23. IIT Bombay NLP Group. (2023). Survey of ILNMT architectures. Journal of Language Engineering.

24. Nvidia-India. (2024). AI infrastructure for multilingual models. Industry Report.

25. Reverie Technologies. (2024). Scaling voice-based MT for Indian dialects. IEEE Transactions on NLP.