Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Evaluating Baseline Architectures for English-magahi Machine Translation: Architectural Tuning and the Limits of Generalization on Small Parallel Corpus

Chandan Kumar
School of computer & Systems Sciences
Jawaharlal Nehru University
New Delhi, India

Prof. Piyush Pratap Singh
School of computer & Systems Sciences
Jawaharlal Nehru University
New Delhi, India

Abstract

NMT models for extremely low-resource Indic languages suffer from severe data bottleneck problems. In this paper, we analyze the potential and limitations of architectural fine-tuning approaches for English-Magahi translation under the strict data limit of a parallel corpus of 13840 data pairs. We developed two models from scratch: a simple word-level CNN-BiLSTM and an improved subword BiLSTM with BPE tokenization, label smoothing, scheduled sampling, and attention masking. Despite employing sophisticated regularization techniques, our models exhibit a bizarre problem of training failure. Both our models produce near perfect performance (>90 BLEU) on template-only held-out test set but fail terribly (<1 BLEU) on out-of-distribution examples. We identify the reason for this generalization failure as "template collapse" - a situation wherein the model learns artificial target structures rather than cross-lingual mappings. This work shows the empirical upper bound on the performance of NMT models in data-sparse situations. Neither architectural adjustments nor tokenization changes can replace the need for larger and diverse datasets. Future works on robust Magahi NMT models will have to go beyond from-scratch initialization and adopt multi-lingual transfer learning and synthetic data augmentations.

I. INTRODUCTION

Neural Machine Translation (NMT) has recently become the state-of-the-art for machine translation and is rapidly replacing the old paradigms, primarily statistical MT. Unlike its precursors (which built independent components including phrase tables and alignment models), NMT translates a source text into a target language end-to-end using deep neural networks. Attention mechanisms, and, particularly, Transformer models in NMT have led to more fluent translations that also maintain better contextual coherence. NMT models have attained the best results on many low-resource language pairs, especially with advances in capturing long-range dependencies. Notwithstanding the aforementioned progress, training NMT systems currently requires a vast amount of data, typically billions of parallel sentences which significantly limits the application of high-resource NLP techniques on extremely under-resourced languages [1], [2], [3].

Magahi, which is a part of Indo-Aryan languages is used predominantly in eastern India and has an extremely low number of parallel resources for Machine translation systems and other Natural Language Processing related techniques. It falls under the Category of Low-Resource languages in Indian subcontinent. Despite the extensive speaker base in Indian state such as Bihar, Jharkhand and West Bengal, it remains underrepresented in research publications, in creation of standard lexical or evaluation dataset and creation of publicly accessible parallel corpora between Magahi and English, the language considered more standard as one end of the translation [4], [5], [6]. As the vast, best results from neural machine translate systems are achieved by having enough million parallel sentences, in which the model would learn about lexico-synatactic, however training NMT model from scratch for very few training sentences, results in non-stable convergence and it causes failure of model to incorporate general lexico-syntactic phenomena available in a large corpus, which causes issue of OOV with a high percentage, and poor coverage on novel and real data [7], [8].

Other challenges pertaining to translating under-resourced Indic languages include morphology richness, varied orthography and open word order- characteristics of Magahi also exhibit. Rich word morphological complexity in some Indian language like Magahi and presence of variable orthographies pose severe problems for tokenization and subsequent sequential modeling; standard word level tokenizers on the extremely Zipfian vocabulary generates poor coverage and a substantial portion of Out-Of-Vocabulary (OOV) items [9], [10], [11]. Subword-based tokenization with techniques like Byte-Pair Encoding (BPE) and SentencePiece, widely used to consolidate statistical strength across similar surface form and to exploit morpheme-level similarities, on extremely smaller corpus may not effectively share statistic power across words and also fragment representations to severely large extents. Even with advanced regularisation in place, training on constrained data tends to lead to target-side redundancy and cause a major many-to-one mapping problem where models learn high frequency templated generic phrases instead of translating meaningful content [10], [12], [13], [14].

In this research work we address these very challenges and seek to understand the extreme limits of from-scratch learning on the task of English-Magahi machine translation in an under-resourced setup by analyzing two prominent and relatively efficient baseline machine translation architecture paradigms in terms of their generalizations. We primarily assess to see how advanced architectural tuning in and learning techniques helps avoid generalization collapse with from-scratch NMT without aid of any large-scale pre-trained multilingual models, like mT5 [15], IndicTrans2 [13], and NLLB-200 [16] etc.

We propose two different from-scratch training set-ups:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

i) Hybrid CNN-BiLSTM with word level-tokenization; and ii) Enhanced BiLSTM Seq2Seq model with BPE tokenization, masking correction, label smoothing, learning rate warm-up (scheduled decay) and scheduled sampling which will drastically reduce the problem of exposure bias. The main idea of using NLTK based POS and template-based split of the data is to remove the problem of data leakage and structural biases from both validation/test split and train-validation split of the data, which is a very common pitfall on such standard split settings using random assignment.

The current findings on Magahi are equally relevant for several other languages from the Indo-Aryan and Indic language family. These languages are native to South Asia, with potentially significant cultural diversity and socio- linguistic influence but with minimal, fragmented or unavailable digital counterparts.

II. LITERATURE REVIEW

The development of NMT systems for low-resource languages-and particularly Indo-Aryan languages like Magahi-depends on a number of interlinked factors including architectural design, subword/ byte tokenization, transfer learning, and dedicated evaluation metrics. This section gives an overview of some key theories, transfer learning models, tokenization techniques, and evaluation problems related to the scaling of English-Magahi translation using small parallel corpora.

1. Theoretical Background and Tokenization Techniques

Contemporary NMT technology fundamentally depends on the Transformer model, which was proposed by Vaswani et al. [1]. Unlike recurrent networks used in earlier approaches, Transformers allow full parallelization of the training process and better modeling of long dependencies between tokens. Nonetheless, the effectiveness of the Transformer model is significantly limited by the way input sequences are tokenized.

In order to overcome the Out-Of-Vocabulary (OOV) problem and deal with infrequent words, Sennrich et al. [17], [18] developed the Byte Pair Encoding (BPE), which repeatedly merges frequent n-grams to create a subword vocabulary. Although BPE is still used as a primary subword tokenization technique, it often fails in the conditions of low-resourced languages, fragmenting words into semantically unconnected parts. In order to tackle the subword limitations, Sreedhar et al. [19] have suggested Local Byte Fusion (LOBEF) that proves that byte-level modeling, which involves raw bytes' representations, outperforms subwords tokenization in zero-shot transfer learning, multilingual scenario, and domain adaptation by over 6 BLEU points. When speaking of English-Magahi NMT, comparing BPE with character and byte-level tokenization is absolutely necessary to avoid serious vocabulary fragmentation on small training datasets.

2. Transfer Learning and Multilingual Pre-training in Low-Resource NMT

In the case of severe data limitation in parallel datasets, training the NMT architectures from scratch usually results in dataset memorization and complete generalization collapse. In turn, transfer learning became the main way to fill this data gap:

- **Parent-child and sequential transfer:** Nguyen and Chiang [20] pioneered parent-child transfer learning when they trained a model on the high-resource parent languages pair and transferred the parameters to a low-resource child languages pair through sharing BPE vocabulary. This resulted in the improvement of up to 4.3 BLEU.
- **Low-Resource Indic Language Adaptations:** Hujon et al. [21] performed transfer learning in English-Khasi translation through shared sub-word vocabularies, showing that parameter transfer across Indian language pairs improves translation performance over isolated baselines.
- **Massive Multilingual Pre-trained Models:** mBART [22], [23] showed how pre-training of sequence-to-sequence transformers across large-scale monolingual corpora with denoising allows for effective transfer across diverse target languages. On a larger scale, Costa-juss et al. [16] introduced No Language Left Behind (NLLB-200), showing how multilingual transfer across 200+ languages can improve translation performance in underserved dialects.
- **Unified & Parameter-Efficient Frameworks:** Pasupuleti et al. [24] benchmarked multilingual representation models (mBERT, XLM-R, mT5) across a suite of low-resource NLP tasks and confirmed that cross-lingual representations alleviate the problem of data paucity. Liu et al. [23] demonstrated that multilingual denoising pre-training enables effective transfer learning for low-resource neural machine translation by learning generalized multilingual representations that can be fine-tuned for specific language pairs.

3. Indic Language Dynamics in the Low-Resource Setting and Dataset Saturation Boundaries

While transfer learning may help in addressing certain problems, the Extremely Low-Resource Languages (ELRLs) have distinctive typological properties. In their study on the thresholds for transfer learning in Magahi, Bhojpuri, and Braj, Raj and Kumar [25] found that fine-tuning multilingual models on 500 to 1,000 sentences leads to performance saturation in case of simple tasks, while Magahi always shows the worst results due to its distinct typology from Hindi.

Moreover, in surveys on Transformer NMT performed by Banerjee et al. [26], it is noted that all NMT models are fundamentally data-driven. Thus, in case of training of standard NMT models from scratch on very small datasets with template duplication, the severe OOV bottleneck and template memorization are observed, leading to high internal validation scores but zero results on test sets.

4. Evaluation Dilemmas, LLMs, and Synthetic Adaptation

Estimating the performance of NMT systems in sparse data and Devanagari script settings causes considerable metric distortion:

Metric Divergence: Kumar et al. [25] performed the comparative evaluation of BLEU and ChrF++ for Magahi, Bhojpuri, and Chhattisgarhi. It was shown that character-level metrics such as

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

ChrF++ tend to distort the quality measure since they give an unfair boost due to script or source duplication, and BLEU heavily punishes morphological variations. ChrF++ being stable while BLEU dropping is a signal of model hallucination or Devanagari script copying.

LLMs and Synthetic Feedback: According to surveys conducted by Gain et al. [27], LLMs demonstrate great zero-shot capacity through In-context Learning (ICL), and yet the dedicated encoder-decoder models with parallel fine-tuning outperform prompting-based LLMs in extreme low-resource conditions. To cope with the data deficiency without the need for retraining, Yang et al. [28] suggested Translation with Synthetic Feedback (TSF): edit-distance algorithms produce revision instructions to iteratively improve draft translation using domain-specific demonstrations from retrieval. Table 1 below summarizes the key literature across tokenization, architecture, transfer learning, and low-resource evaluation paradigms relevant to English-Magahi NMT:

Table 1: Summarized Review

Ref.	Author(s) & Year	Method	Main Contributions	Limitations
[1]	Vaswani et al. (2017)	Transformer	Introduced self-attention architecture replacing RNNs and achieving state-of-the-art NMT performance.	Requires large-scale parallel data for effective training.
[2]	Bahdanau et al. (2015)	Attention-based Seq2Seq	Introduced attention mechanism to improve alignment and translation quality.	Computationally slower than Transformer models.
[3]	Sutskever et al. (2014)	LSTM Seq2Seq	Proposed end-to-end sequence-to-sequence learning using LSTMs.	Suffers from fixed-length context representation.
[4]	Kumar (2020)	Linguistic Analysis	Studied syntactic characteristics of Magahi.	No machine translation implementation.
[7]	Koehn & Knowles (2017)	NMT Challenges	Identified six major challenges in Neural Machine Translation.	Primarily conceptual discussion.
[8]	Zoph et al.	Transfer Learning	Demonstrated transfer learning improves low-resource NMT. Parent – Child transfer technique	Depends on related high-resource languages.
[9]	Revanuru et al. (2017)	Indian Language NMT	Developed NMT systems for 6 Indian languages.	Limited language coverage.
[10]	Park et al. (2020)	Decoding Strategies	Improved decoding methods for low-resource MT.	Gains depend on model architecture.
[11]	Ranathunga et al. (2023)	Survey	Comprehensive survey of low-resource NMT techniques.	Does not propose a new model.
[12]	Conneau et al.	XNLI	Cross-lingual sentence representation benchmark.	Not a translation model.
[13]	Gala et al.	IndicTrans2	High-quality multilingual translation for 22 Indian languages.	Relies on extensive pretraining and large datasets.
[14]	Joshi et al.	Linguistic Diversity	Analyzed inclusion of low-resource languages in NLP.	Survey-based analysis.
[15]	Xue et al. (2021)	mT5	Multilingual text-to-text pre-trained Transformer.	Requires extensive computational resources.
[16]	NLLB Team	NLLB-200	Large multilingual translation model covering 200 languages.	Requires massive training corpus.
[17]	Sennrich et al. (2016)	BPE	Introduced subword tokenization to reduce OOV words.	Can over-segment words in very small corpora.
[18]	Sennrich et	Back	Improved NMT	Requires

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

	al.	Translation	using monolingual data.	additional monolingual corpora.
[19]	Sreedhar et al. (2023)	Local Byte Fusion	Proposed byte-level tokenization outperforming BPE in several multilingual settings.	Computationally more expensive.
[20]	Nguyen & Chiang	Parent–Child Transfer	Introduced transfer learning across related low-resource languages.	Requires linguistically related parent languages.
[21]	Hujon et al. (2022)	English–Khasi Transfer Learning	Applied LSTM based transfer learning for low-resource English–Khasi MT.	Language-specific evaluation.
[22]	Liu et al. (2024)	NLP-based MT Survey	Reviewed recent machine translation developments.	Survey only.
[23]	Liu et al. (2020)	mBART	Multilingual denoising pretraining for NMT.	Large-scale pretraining required.
[24]	Pasupuleti (2025)	Transfer Learning Review	Discussed multilingual transfer learning for low-resource NLP.	Mostly conceptual.
[25]	Raj & Kumar (2026)	POS Transfer Learning	Investigated required data size for transfer learning in Indian languages.	Focused on POS tagging rather than MT.
[26]	Banerjee et al. (2026)	Transforme-r Survey	Reviewed research challenges in Transformer-based NMT.	No experimental validation.
[27]	Gain et al. (2026)	LLM Survey	Surveyed LLM-based machine translation techniques.	Limited practical evaluation.
[28]	Yang et al. (2026)	Domain Adaptive MT	Proposed synthetic-feedback domain adaptation for LLM-based MT.	Requires pretrained LLMs.

III. METHODOLOGY

Building a strong NMT model for ultra-low-resource languages comes with its own set of architectural and data curation difficulties. In this paper, we will explore the methodology, results, and discussion of English-to-Magahi translation model trained completely from scratch on a highly constrained data set of 13840 parallel data pairs.

In order to determine the best way of modeling this low-resource setup without pretrained multilingual embeddings, two different methodological pipelines have been created and tested. Pipeline 1 uses the hybrid CNN-BiLSTM model architecture using word tokenization. Pipeline 2 uses the BiLSTM Seq2Seq architecture enhanced with BPE subword tokenization, advanced training dynamics, and beam search decoding.

1. Data Curation and Preprocessing

Parallel corpus for English and Magahi languages has been generated based on the template-based approach with the use of LLM models. Special bilingual lexicon of subjects, verbs, objects, time expressions, and adverbs has been used in the predefined grammatical templates. Although such methodology gives us an opportunity to generate parallel corpora with controlled vocabulary coverage and grammar structure for baselines' training in machine translation, they are relatively poor in linguistic diversity compared to the parallel corpora generated by natural means. The models were evaluated on two test sets, first was the Held-Out Test Set (Template-Isolated) and other was the gold-standard test set of 1,436 pairs of manually generated and verified English-Magahi parallel sentences, which had no connection with the training corpus and used only for unbiased model evaluation.

Data cleaning in order to minimize noise in corpora was an important part of the pipeline in both experiments. The texts in both languages were converted to the lowercase and all extra whitespaces were removed; also, the regular expression filtering was applied in order to leave only alphanumeric symbols in the English sentences and Devanagari characters in Magahi sentences. These procedures helped to prevent the models from learning arbitrary punctuation distribution.

After text normalization, the extremely important procedure of deduplication has been performed. Deduplication was performed using only English source sentences. In this way, 6258 duplicated English sentences were excluded from 20,099 input rows. Thus, we were left out with the 13840 valid pairs.

2. Template-Based Dataset Splitting

Using random dataset splitting methods such as train_test_split can

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

lead to data leakage in template heavy datasets due to the same grammatical structure appearing in the training and evaluation datasets. In order to ensure the strictness of the evaluation process of the model's generalization capability, a grammatical template-based splitting process was employed before any tokenization took place on the data frame.

- **Template Creation:** The Natural Language Toolkit was used in order to parse the source sentences of English language. The function words ("the", "a", "is", "in", "he", "she") were kept in their plain text form while the content words were replaced with their respective part of speech (POS). For instance, a sentence such as "she reads history" becomes "she VBZ NN".
- **Isolated Groups:** With the help of Group Shuffle Split, the dataset was split up according to the above-mentioned templates. This ensured that specific grammatical configurations were isolated entirely within their respective splits.
- **Ratio:** The ratio used for dividing the dataset was 80:10:10 for Training, Validation, and Testing.

3. Tokenization Approaches

The difference between two different translation pipeline was that they used different approaches for tokenizing the textual information.

- **Pipeline A (Word-Level Tokenization):** The Hybrid CNN-BiLSTM model used a standard Keras Tokenizer fitted exclusively on the training text to prevent vocabulary leakage from the validation and test sets. It relied on whole-word mappings and an Out-Of-Vocabulary (OOV) token <unk> to handle unknown terms.
- **Pipeline B (Subword BPE Tokenization):** The optimized BiLSTM pipeline noted that tokenization at the word level was very inefficient with the extremely limited Zipfian vocabulary of about 13840 pairs. In order to solve this problem of inefficiency of the word-level tokenization, the optimized BiLSTM utilized the Sentence Piece Byte Pair Encoding (BPE) Tokenizer. This type of tokenizer proved to be apt for the parallel data set that was relatively small in size along with having a rich inflectional vocabulary like the Magahi language. The subword vocabulary sizes for English and Magahi languages were limited to 4,000 and 3,000 respectively to ensure a good balance between efficiency and vocabulary coverage.

4. Architectural Frameworks

Both models used a Sequence-to-Sequence (Seq2Seq) encoder-decoder framework along with Bahdanau Attention mechanism, though their internal compositions varied.

- **Hybrid CNN-BiLSTM Network Structure (Pipeline A):** The encoder network in this case took the input embedding (embedding dimension = 256) and processed it using a 1D CNN layer with 256 filters and kernel size = 3, and ReLU activation function. The CNN layer was meant to extract n-gram features from the sequence and then feed it to the subsequent BiLSTM layer having 256 recurrent units. The decoder was built with an LSTM layer having 512 units (equal to the concatenated output of bidirectional encoder).
- **Optimized BiLSTM Network (Pipeline B):** In this pipeline, the CNN layer was removed and only the BiLSTM encoder (256 units) and the LSTM decoder (512 units) were used along with embedding dimension of 256. A high dropout ratio of 0.3 was imposed in the recurrent layers for regularization of small subword vocabulary.
- **Attention Masking Correction:** An important mathematical correction was introduced into the Attention layer by Bahdanau in Pipeline B. Generally, when attention masking is applied, the padding needs to be set aside in order to avoid assigning it attention weights. Pipeline B accomplished this in its own way by adding a large scalar to the padding positions such that S_masked = [S + (1.0 – mask) * - 1 * 10 ^ 9]. This nullified the attention weights of padding tokens after applying the softmax function.

5. Advanced Training Dynamics

To make the learning process in Pipeline B even more stable, the traditional training loop was replaced entirely with a new one that uses the state-of-the-art NMT regularization techniques.

- **Label Smoothing:** The Sparse Categorical Cross-Entropy loss function was modified with label smoothing using a factor of 0.1. As a result, label smoothing introduced penalty for the decoder on the account of generating an overly confident but incorrect output sequence-the key reason behind generic template generation.
- **Learning Rate Scheduling:** The static Adam optimizer was replaced with a learning rate scheduling approach. The learning rate schedule involved linear warming up during the first three epochs with achieving the maximum learning rate of 1e-3 and then used cosine annealing for 60 more epochs.
- **Scheduled Sampling:** “Exposure bias,” in which case a model is trained by using only the ground-truth tokens and during inference, it needs to predict based on its own predictions, which can be wrong. For this reason, the scheduled sampling strategy was employed in the custom tf.GradientTape training loop. In particular, the teacher forcing ratio was decayed from 1.0 (100% feeding of ground-truths) to 0.6 linearly across training.

6. Inference and Decoding

During the inference procedure, Pipeline A used the greedy decoding method by taking the argmax of the softmax distribution at each time step. On the other hand, Pipeline B used a Beam Search decoding method. By keeping a beam width of 5 candidate sequences and a length penalty of 0.7 in order to avoid too short output, the algorithm tried to optimize the global probability of the Magahi translation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

IV. RESULT AND DISCUSSION

The performance of both models was assessed based on exact match percentage, SacreBLEU, chrF++, and Translation Edit Rate (TER). Assessment was done based on two separate boundaries: the template-isolated test set, which is composed of unique grammar found in the original dataset, and the completely separate unseen test set. Table 1 and Table 2 represents the summarized results.

Table 1: Evaluation on Held-Out Test Set (Template-Isolated)

Metric	Hybrid CNN-BiLSTM (Pipeline A)	Optimized BiLSTM (Pipeline B)
Exact Match	80.10%	83.09%
SacreBLEU	93.83	95.12
chrF++	96.79	97.73
TER	4.0	3.86

Table 2: Evaluation on Completely Unseen Test Set

Metric	Hybrid CNN-BiLSTM (Pipeline A)	Optimized BiLSTM (Pipeline B)
Exact Match	0.28%	3.48%
SacreBLEU	0.20	0.35
chrF++	7.42	7.53
TER	93.81	95.89

1. Analysis of the Generalization Gap

The assessment of the English-to-Magahi translation systems shows the presence of significant gap between internal validation scores and practical performance of these systems. Two neural network architectures have been assessed in two different environments – on the template-free held-out test set created from the Training_file.csv file, and on the totally unseen test set which includes completely new English inputs.

In case of internal held-out test set, both models showed an outstanding performance on internal test set of 1573 samples. Pipeline A demonstrated the BLEU score of 93.83 and the chrF++ score of 96.79. Pipeline B showed the BLEU score of 95.12 and chrF++ of 97.73. These numbers mean perfect performance of translation system, close to the level of human performance.

What I've discovered though was that testing the two systems on the set of 1,436 unseen samples made me notice the catastrophic failure of generalization ability. Both systems failed: Pipeline A scored the BLEU of 0.20 and the Exact Match of 0.28%, while Pipeline B got the BLEU of 0.35 and Exact Match of 3.48%. The Translation Edit Rate for both models exceeded 90 in case of unseen test set, which means that all generated outputs were completely useless and need to be retranslated.

2. The Illusion of Translation and Template Collapse

The striking difference between the internal test metrics and the unseen test metrics gives the key insight into the mechanisms behind extremely low-resource NMT: highly BLEU scores on small homogenous corpora usually denote the memorization of the datasets rather than the understanding of language.

Even though we have applied the elaborate template isolation technique, using NLTK POS tagging to guarantee no repetitions of the same English sentence structure in the training and testing processes, the models have been able to "cheat" the internal evaluation, which was caused by the extreme redundancy on the target side of Training_file.csv: roughly 27.4% of Magahi target sentences had identical copies (which corresponded to the different English source sentences). Consequently, the models learned a limited number of frequent Magahi templates. So, when translating real world unseen English sentences, the decoders in the CNN-BiLSTM and in the subword BiLSTM were falling back to these overused Magahi templates. Thus, even though we have applied such advanced techniques as BPE, label smoothing, and scheduled sampling which helped to stabilize the training process, the models failed to produce the semantic variety which was missing in the fundamental 13840 sentence pairs.

V. CONCLUSION AND FUTURE DIRECTIONS

In this research paper, we have compared two baseline architectures of Neural Machine Translation (NMT): the word-level CNN-BiLSTM and subword BiLSTM, trained and tested on 13840 valid pairs for English-Magahi translation. The results of our experiments demonstrated the existence of an interesting paradox in extremely low-resource NMT: although both of the pipelines gave near-perfect internal validation metrics (90+ BLEU), they experienced complete generalization collapse (sub-1 BLEU) on unseen data. The models memorized artificial redundant templates rather than acquiring real cross-lingual semantics, which we call "template collapse."

Conclusions derived from the aforementioned methodologies strongly highlight the fact that no magic can create linguistic data which is not present in the training set. In order to obtain truly high-quality translations in future Magahi research, the above methodology should be changed completely and from-scratch initialization should no longer be considered as the default approach. We recommend pursuing the following key areas:

1. Transfer Learning Implementation

The most common next step for further research would be transfer learning implementation. It is extremely important to fine-tune huge multilingual Seq2Seq models such as IndicTrans2 or Meta's NLLB-200 (already containing Magahi in the Devanagari script) as they would provide the translation system with a huge repository of common vocabulary, syntax, and morphology inherited from the sister languages like Hindi, Bhojpuri, or Maithili.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

2. Fulfillment of a sturdy Dataset Requirement

In addition to above, developing a robust NMT system requires fulfillment of the sturdy dataset requirement. The current limitation of about 13,840 highly redundant template-based sentence pairs is a very serious bottleneck making models memorize the output, instead of learning semantic alignments. It is important for the future research to create a large-scale quality parallel corpus of the language in question. A sturdy dataset should:

- **Scale Significantly:** Go from tens of thousands to hundreds of thousands of unique parallel sentences to fully map the cross-lingual vector space.
- **Guarantee Linguistic Diversity:** contain diverse morpho-syntactic complexity, different sentence lengths, multi-clauses and other conversational nuances missing in synthetic/template-based datasets.
- **Take care of strict orthography normalization:** ensure strict orthographic normalization to address issues of dialectal variations in the devanagari script causing an increase in vocabulary sizes and making subword tokenization ineffective.

3. Aggressive Data Augmentation

In case from-scratch training has to be maintained, any algorithmic tuning should be abandoned in favor of aggressive data augmentation to meet the sturdy dataset requirement. The methods of such augmentation may include back-translation (with the use of a Magahi-to-English reverse model to create synthetic pairs using monolingual data) or large-scale crowdsourcing of authentic texts by native speakers.

VI. VALIDITY THREATS

In order to have a reliable and thorough evaluation of the research, several validity threats in these findings need to be discussed.

1. Internal Validity

Internal validity relates to the possibility that the experimental design has introduced some biases that would have distorted the results.

- **Target-side Deduplication:** Honestly, the biggest threat to internal validity is the data deduplication procedure. Although Training_file.csv has been extensively deduplicated on the English source side (with more than 6,258 pairs being removed), the targets remained highly repetitive. Such a many-to-one mapping heavily skewed the loss function and made the attention mechanism prioritize frequent target templates rather than source-target mappings.
- **Template Split Accuracy**: In the process of creating the training/test split via Group Shuffle Split, the use of the NLTK English POS tagger in order to identify sentence templates might have led to an erroneous splitting of sentence templates that resulted in some structural leakage between the training and test datasets, and hence caused an overestimated held-out BLEU score.
- **Hyperparameters' Restrictions:** The architectures used in this research were heavily regularized (e.g., with Dropout of 0.3 and label smoothing of 0.1) in order to avoid overfitting due to a small vocabulary. However, this restriction might have over-constrained the models to learn the more complex dependencies, which in turn caused the poor performance of the models on unseen data.

2. External Validity

External validity relates to the generalizability of the findings beyond the specific setting of this particular experiment.

- **Dataset Size & Domain Restriction:** First of all, the study uses the extremely limited dataset consisting of roughly 13,840 sentence pairs obtained from training file. However, such size is completely insufficient for training the NMT model from scratch. Furthermore, if the dataset is domain-specific (e.g., conversational or generic phrases), the model cannot be expected to generalize on literary, technical, or colloquial Magahi.
- **Application to Other Low-Resource Languages:** Although the failure to train a language model from scratch is a common issue in low-resource NLP studies, the specific inflectional/agglutinative nature of Magahi means that these very failure modes and extent of the template collapse may not be applicable to other languages with different morphology.

3. Construct Validity

Construct validity deals with the appropriateness of the used metrics in measuring the translation capability of the system.

- **Limitations of Metrics:** Using exact match percentage as well as BLEU based on corpora can be very misleading for languages having high morphology, such as Magahi language. BLEU calculates exact n-grams overlap and severely penalizes those translations which may be semantically correct but use different synonyms, different order of words, or different morphologies.
- **chrF++ Applicability:** Although chrF++, which calculates n-grams overlap on character level, is a more stable metric for highly inflected languages compared to BLEU, it could not calculate any kind of semantic transfer in the unseen data set, thus proving that the problem is fundamental.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

References

[1] A. Vaswani *et al.*, "Attention Is All You Need."

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," May 2016, [Online]. Available: http://arxiv.org/abs/1409.0473

[3] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," Dec. 2014, [Online]. Available: http://arxiv.org/abs/1409.3215

[4] C. KUMAR, "NP/DP Parameter and Classifier Languages: A Case of Magahi," *The Journal of Indian and Asian Studies*, vol. 01, no. 01, p. 2050005, Jan. 2020, doi: 10.1142/s2717541320500059.

[5] "Preservation of Magahi Language in India: Contemporary Developments | Springer Nature Link." Accessed: Jun. 13, 2026. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-18146-7_11

[6] "Magahi | 13 | The Indo-Aryan Languages | Sheela Verma | Taylor & Franc." Accessed: Jun. 13, 2026. [Online]. Available: https://www.taylorfrancis.com/chapters/edit/10.4324/9780203945315-13/magahi-sheela-verma?context=ubx&refId=74b91005-042e-4979-8cba-281fdbd56bf7

[7] P. Koehn and R. Knowles, "Six Challenges for Neural Machine Translation," 2017. [Online]. Available: http://www.statmt.org/wmt17/

[8] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer Learning for Low-Resource Neural Machine Translation."

[9] K. Revanuru, K. Turlapaty, and S. Rao, "Neural machine translation of Indian languages," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Nov. 2017, pp. 11–20. doi: 10.1145/3140107.3140111.

[10] C. Park, Y. Yang, K. Park, and H. Lim, "Decoding strategies for improving low-resource machine translation," *Electronics (Switzerland)*, vol. 9, no. 10, pp. 1–15, Oct. 2020, doi: 10.3390/electronics9101562.

[11] S. Ranathunga, E. S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural Machine Translation for Low-resource Languages: A Survey," *ACM Comput. Surv.*, vol. 55, no. 11, Nov. 2023, doi: 10.1145/3567592.

[12] A. Conneau *et al.*, "XNLI: Evaluating Cross-lingual Sentence Representations," Association for Computational Linguistics.

[13] J. Gala *et al.*, "IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages." [Online]. Available: https://openreview.net/forum?id=vfT4YuzAYA

[14] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." [Online]. Available: https://microsoft.github.io/linguisticdiversity

[15] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2010.11934

[16] N. Team *et al.*, "No Language Left Behind: Scaling Human-Centered Machine Translation." [Online]. Available: https://github.com/facebookresearch/fairseq/tree/nllb.

[17] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units."

[18] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data." [Online]. Available: http://www.statmt.org/wmt15/

[19] M. N. Sreedhar, X. Wan, Y. Cheng, and J. Hu, "Local Byte Fusion for Neural Machine Translation," Long Papers. [Online]. Available: https://github.com/

[20] T. Q. Nguyen and D. Chiang, "Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation." [Online]. Available: https://cis.temple.edu/

[21] A. V. Hujon, T. D. Singh, and K. Amitab, "Transfer Learning Based Neural Machine Translation of English-Khasi on Low-Resource Settings," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 1–8. doi: 10.1016/j.procs.2022.12.396.

[22] Y. Liu, Y. Ma, S. Zhou, and X. Luo, "A Survey of Research and Application of NLP-based Machine Translation," in *2024 6th International Conference on Natural Language Processing, ICNLP 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 315–319. doi: 10.1109/ICNLP60986.2024.10692780.

[23] Y. Liu *et al.*, "Multilingual Denoising Pre-training for Neural Machine Translation", doi: 10.1162/tacl.

[24] M. K. Pasupuleti and R. Director, "( ) Research Paper Title :Multilingual NLP for Low-Resource Languages Using Transfer Learning," *International Journal of Academic and Industrial Research Innovations(IJAIRI)*, [Online]. Available: www.nationaleducationservices.org

[25] M. Raj and R. Kumar, "How Much Data in Low-resource Indian Languages is 'Sufficient'' for Transfer Learning: A Comparative Study for POS Annotation,'" *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 25, no. 1, Jan. 2026, doi: 10.1145/3783981.

[26] A. Banerjee, B. K. Singh, V. Kumar, and D. Banik, "Research challenges and future directions in transformer-based neural machine translation," May 05, 2026, *Elsevier Ltd*. doi: 10.1016/j.eswa.2025.131062.

[27] B. Gain, D. Bandyopadhayay, A. Ekbal, and T. N. Singh, "Bridging the linguistic divide: a survey on leveraging large language models for machine translation," *Lang. Resour. Eval.*, vol. 60, no. 2, Jun. 2026, doi: 10.1007/s10579-026-09919-7.

[28] X. Yang *et al.*, "Domain Adaptive Machine Translation with Synthetic Feedback for Large Language Models," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 25, no. 3, Feb. 2026, doi: 10.1145/3787498.