Team Horizon at BHASHA Task 1: Multilingual IndicGEC with Transformer-based Grammatical Error Correction Models

Manav Dhamecha, Gaurav Damor, Sunil Choudhary, Pruthwik Mishra
{u24ai034, u24ai026, u24ai063, pruthwikmishra}@aid.svnit.ac.in

Abstract

This paper presents Team Horizon's approach to the BHASHA Shared Task 1: Indic Grammatical Error Correction (IndicGEC). We explore transformer-based multilingual models — **mT5-small** and **IndicBART** — to correct grammatical and semantic errors across five Indian languages: Bangla, Hindi, Tamil, Telugu, and Malayalam. Due to limited annotated data, we develop a synthetic data augmentation pipeline that introduces realistic linguistic errors under ten categories, simulating natural mistakes found in Indic scripts. Our best submissions achieve competitive performance with GLEU scores of **86.03** (Tamil, 5th rank), **84.36** (Malayalam, 8th rank), **82.69** (Bangla, 6th rank), **80.44** (Hindi, 7th rank), and **72.00** (Telugu, 6th rank) on the official test sets. We further analyze the impact of dataset scaling, multilingual fine-tuning, and training epochs, demonstrating that linguistically grounded augmentation significantly improves grammatical correction accuracy in low-resource Indic languages.

1 Introduction

Automatic Grammatical Error Correction (GEC) aims to detect and correct errors in text while preserving its intended meaning. Although modern GEC systems for English have achieved remarkable success through large-scale pre-training and high-quality datasets, their extension to Indic languages remains challenging due to linguistic and data-related constraints. Indic languages exhibit high morphological complexity, rich inflectional patterns, free word order, and diverse orthographies. Available annotated corpora for languages such as Bangla, Tamil, and Malayalam are extremely small, often only a few hundred examples, making traditional supervised learning insufficient for robust correction.

The BHASHA 2025 Shared Task 1: IndicGEC introduces a multilingual benchmark for grammatical error correction in five major Indian languages: Bangla, Hindi, Tamil, Telugu, and Malayalam. Team Horizon adopted a hybrid approach combining:

- Synthetic data augmentation through linguistically motivated error injection.
- Multilingual transformer fine-tuning using mT5-small (Xue et al., 2021a) and IndicBART (Dabre et al., 2022).

We deliberately selected these two models because they represent the two dominant pre-training paradigms for Indic languages—general multilingual (mT5-small) and Indic-specific (IndicBART)—while remaining lightweight ($\leq$ 300M parameters), publicly available, and fast to fine-tune on standard academic hardware. This choice ensures fair comparison under identical conditions and establishes strong, reproducible baselines for future low-resource IndicGEC research.

We created a controlled error generation pipeline introducing mistakes across 10 linguistic categories. This expanded training data from less than 1k to over 10k high-quality pairs per language. Our main contributions are as follows:

- Introduce a linguistically informed synthetic error-injection framework for Indic GEC data augmentation.
- Evaluate and compare two multilingual transformer models: mT5-small and IndicBART.
- Provide empirical analysis of dataset scaling, training epochs, and their effects on generalization.
- Release insights into error-type distributions, cross-language transfer, and limitations in multilingual setups.

Error Class	Sub-class	Example (Hindi)
Spelling	Non-Dictionary	मैं कारखानाया काज करु। → मैं कारखाने में काम करता हूँ। (I work in a factory.)
	Dictionary	मैं कल बारी जाऊँगा। → मैं कल घर जाऊँगा। (I will go home tomorrow.)
Word	Tense	मैं कल पढ़ेगा। → मैं कल पढ़ूँगा। (I will study tomorrow.)
	Person	मैं स्कूल जाती है। → मैं स्कूल जाता हूँ। (I go to school.)
	Number	वे किताबें पढ़ता है। → वे किताबें पढ़ते हैं। (They read books.)
	Gender	सीमा सोया। → सीमा सोई। (Seema slept.)
	Case	राम को किताब पढ़ीं। → राम ने किताब पढ़ी। (Ram read the book.)
	Parts-of-Speech	हिमालय सुंदर आश्चर्यजनक है। → हिमालय आश्चर्यजनक रूप से सुंदर है। (The Himalayas are remarkably beautiful.)
	Missing	मैं कल जाऊँगा। → मैं कल घर जाऊँगा। (I will go home tomorrow.)
	Extra/Structure	राम ने ने खाना खाया। → राम ने खाना खाया। (Ram ate food.)
Punctuation	—	क्या तुमने खाना खाया → क्या तुमने खाना खाया? (Did you eat food?)
Semantic	—	राम आकाश को खा रहा है। → राम आम खा रहा है। (Ram is eating mango, not the sky.)

Table 1: Synthetic error categories with detailed sub-classes and examples. Wrong text is shown in violet, correct text in blue.

Note: The meaning and correctness of some error examples, such as वे किताबें पढ़ता है। and वे किताबें पढ़ते हैं।, can depend on the intended context. Both sentences may seem grammatically plausible, but only the correct form accurately conveys plural subject-verb agreement in typical usage. Such distinctions are essential in grammatical error correction, as surface correctness may not always reflect the intended meaning.

The rest of this paper is organized as follows: Section 2 details dataset collection and augmentation. Section 3 presents model architecture and training setup. Section 4 describes evaluation and results. Section 5 provides detailed error analysis. Section 6 concludes the paper.

2 Dataset Preparation and Augmentation

2.1 Overview

The official IndicGEC datasets released by BHASHA 2025 (Bhattacharyya and Bhattacharya, 2025) contains relatively small language-specific corpora as shown in Table 2, each consisting of a few hundred annotated pairs. To mitigate the data limitation, we develop a synthetic data augmentation pipeline that generates realistic grammatical errors based on predefined linguistic categories. This allow us to scale the dataset size to approximately 5k–10k pairs per language for both mT5-small and IndicBART experiments.

2.2 Data Sources

- **BHASHA GEC Data:** The official shared task dataset containing human-written and expert-corrected essays in five Indian languages.
- **Supplementary Corpora:** Clean sentences were additionally sourced from the AI4Bharat IndicCorp v2 (Doddapaneni et al., 2023) dataset and Indic Wikipedia dumps to expand the data coverage.

Language	#Train	#Dev	#Test
Hindi	599	107	236
Bangla	598	101	330
Malayalam	300	50	102
Tamil	91	16	65
Telugu	599	100	310

Table 2: Language Wise Data Statistics

Each clean sentence from these and supplementary sources is treated as a gold reference and trans-

formed into a synthetically "incorrect" version using our controlled error injection framework.

2.3 Synthetic Error Injection

We design a rule-based error generator that introduces one or more grammatical or orthographic errors per sentence (full list of categories and subclasses with examples in Table 1). In total, we implemented 42 linguistically motivated rules (2–8 per main class). Representative rules include:

- Spelling (non-dictionary): random मात्रा swaps (ा↔ि,े↔ै,ु↔ू, ो↔ौ), visually similar consonant substitution (क → ख, त → थ, न → ण, स → श), or insertion of typographically adjacent keys.
- Spelling (dictionary): replacement with real-word homophones/misspellings from a hand-curated list (e.g., बारिश → बारीश).
- Word (all sub-classes): morphological inflection mutations using pattern lists (e.g., है → थी for gender mismatch, हैं → है for number disagreement, ने → को for wrong case).
- Parts-of-Speech and Missing/Extra/Structure: random omission, duplication, or insertion of postpositions (ने, को, में, से, का/की/के) and adverbs (बहुत ↔ थोड़ा).
- Punctuation: removal or wrong placement of । / ? / ,.
- Semantic: semantically incorrect postposition or adverb choice (में → पर, आज → कल).

The number of errors per sentence follows the distribution 60% (1 error), 25% (2 errors), 10% (3 errors), 5% (4+ errors). Each clean sentence generates five synthetic noisy variants (three heavy with 2–4 errors, two light with 1–2 errors), yielding approximately 10k–12k high-quality parallel pairs per language after deduplication.

2.4 Language-specific Adaptation

Each Indic language exhibits distinct structural patterns and error tendencies:

- **Hindi, Bangla:** Primarily grammar and spelling inconsistencies.
- **Tamil, Telugu, Malayalam:** Morphological, tense, and word-order errors.

3 Model Architecture and Training Setup

3.1 Transformer Models

We experimented with two models:

- **mT5-small** (Xue et al., 2021a): 300M parameters, pre-trained on mC4 (Xue et al., 2021a).
- **IndicBART** (Dabre et al., 2022): Pretrained seq2seq model for Indic languages.

3.2 Input-Output Formatting

- Input: "correct this: <incorrect sentence>"
- Output: <correct sentence>
- Language tags (e.g., [HI], [BN]) are prepended for multilingual fine-tuning. The language tags are two lettered identifiers for the languages defined under ISO 639-1 [1] standards.

3.3 Training Setup

The hyper-parameters used in training are detailed in Table 3.

Parameter	Setting
Optimizer	AdamW
Learning Rate	5e–5 (mT5-small), 3e–5 (IndicBART)
Batch Size	16–32
Epochs	10–15
Loss Function	Cross-entropy
Early Stopping	Based on GLEU score (dev set)

Table 3: Training setup and hyperparameter configuration.

4 Evaluation and Results

Model	Bn	Hi	Ta	Te	Ml
mT5-small	82.69	80.44	86.03	72.00	84.36
IndicBART	73.50	72.33	76.45	66.10	74.84

Table 4: GLEU scores (%) per language. Bn: Bangla, Hi: Hindi, Ta: Tamil, Te: Telugu, Ml: Malayalam.

Previous studies (Taunk and Varma, 2023) have often observed comparable or even superior performance of IndicBART over mT5-small in Indian language tasks, particularly in summarization and machine translation. IndicBART, being specifically pre-trained on Indic languages, tends to better capture linguistic nuances. However, in our experiments, we found that mT5-small slightly outperformed IndicBART for certain languages (most notably Tamil and Malayalam), possibly due to more effective parameter tuning or differences in the data augmentation scheme. Nonetheless, our findings are consistent with the observation that

[1] https://en.wikipedia.org/wiki/ISO_639-1

model performance is sensitive to task, data size, and fine-tuning strategy.

GLEU (Mutton et al., 2007) is used for evaluation because of its robustness in short corrections and small datasets. The results are shown in Table 4.

Ablation Studies

- **Dataset size:** Training on larger augmented datasets improved GLEU by 4–5 points.
- **Number of epochs:** Performance plateaued at 8–10 epochs; overfitting observed beyond this.

5 Error Analysis

Errors are grouped into different categories across different languages in the validation set and are presented in Table 5.

Error Type	Corrected (%)	Missed (%)
Spelling	95	5
Grammar	88	12
Punctuation	92	8
Word Choice	85	15
Semantic	78	22
Structural	80	20
Duplication	90	10

Table 5: Error-type performance across dev sets

Language-specific Observations

Bangla/Hindi: High agreement errors corrected effectively.
Tamil/Telugu/Malayalam: Morphological and word-order errors were more challenging.
Cross-lingual transfer observed between related Dravidian languages.

6 Limitations

This study has several key limitations. First, our synthetic error generation may not fully reflect the diversity and complexity of real-world errors, reducing ecological validity. Second, we evaluated only two multilingual models (mT5-small and IndicBART), excluding stronger language-specific alternatives such as BanglaT5 (Bhattacharjee et al., 2023) or ByT5-based models (Xue et al., 2021b). Third, the rule-based error injection, while linguistically motivated, may miss rare or highly context-dependent phenomena (e.g., dialectal variations or code-mixing).

Additionally, the BHASHA datasets are small and limited to five Indic languages, constraining generalizability. Deeper cross-lingual transfer opportunities were not fully explored, and evaluation relied solely on automatic metrics (GLEU (Napoles et al., 2015)) without human assessment of fluency, meaning preservation, or practical usability.

Future work should incorporate real learner corpora, test language-specific pretrained models, extend augmentation to more Indic languages, perform human evaluations, and investigate advanced cross-lingual and few-shot approaches for ultra-low-resource settings.

7 Conclusion

We demonstrate that linguistically guided synthetic data augmentation, combined with multilingual fine-tuning of transformer models such as mT5-small and IndicBART, can significantly bridge the low-resource gap in Indic grammatical error correction. Our approach yields competitive performance across Bangla, Hindi, Tamil, Telugu, and Malayalam on the BHASHA 2025 Shared Task benchmark, highlighting the effectiveness of controlled error injection in scaling limited annotated data. These results underscore the potential of scalable, language-informed augmentation strategies for advancing GEC in morphologically rich, low-resource Indic languages.

References

Abhik Bhattacharjee, Tahmid Hasan, Anindya Sahu, Kumar Shikhar Deep Anand, Md. Rokibul Hasan, Rakesh Sreenivas, Roshni Ghosal, Asif Salekin, and Mohammad Mahbubur Rahman. 2023. Banglat5: A suite of pre-trained language models for Bangla. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12345–12356.

Pramit Bhattacharyya and Arnab Bhattacharya. 2025. Leveraging LLMs for Bangla grammar error correction: Error categorization, synthetic data, and model evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8220–8239, Vienna, Austria. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL*

2022, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 588–593.

Dhaval Taunk and Vasudeva Varma. 2023. Summarizing indian languages using multilingual transformers based models. In *Proceedings of the 15th Forum for Information Retrieval Evaluation (FIRE 2023)*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021a. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mt5: A massively multilingual pre-trained text-to-text transformer.