A Hybrid Neurosymbolic Approach for Tamil and Malayalam Grammatical Error Correction

Akshay Ramesh, Ratnavel Rajalakshmi

VIT CHENNAI

IJCNLP-AACL 2025

INTRODUCTION

The Grammatical Error Correction (GEC) focuses on automatically detecting and correcting errors in written text, including spelling mistakes, grammatical inconsistencies, punctuation errors, and word choice issues. While significant progress has been made for high-resource languages, GEC for Indic languages face severe challenges, notably data scarcity and morphological complexity**.**

KEY CHALLENGES

Extreme Data Scarcity: IndicGEC provides only 91 training pairs for Tamil, compared to millions for English.
Morphological Complexity: Both languages exhibit agglutinative morphology with rich inflectional systems and complex verb conjugations.
Script Complexity: Unique Unicode challenges, including chillu character variations in Malayalam.

RATIONALE IN THE CHOICE OF THE MODEL: NEUROSYMBOLIC MODEL

Our hybrid neurosymbolic architecture leverages complementary strengths to overcome the limitations of pure neural or rule-based approaches in low-resource settings.

Pure Neural Models

Require millions of training examples, leading to severe overfitting with limited data (e.g., 91 examples for Tamil). Exhibit unpredictable generation behaviours and lack deterministic guarantees.

+

Pure Rule-Based Systems:

Provide perfect accuracy on explicitly encoded patterns but lack generalisation to unseen error types. Cannot correct novel errors not captured in manual rules.

=

Neurosymbolic Solution:

Combines the generalisation of neural models with the precision of symbolic rules, enhanced by augmented data and intelligent ensemble selection.

RELATED WORKS

Recent GEC research, especially for English, has been dominated by neural approaches where Transformer-based models and large pre-trained models like BART and T5 achieved state-of-the-art results (Zhao et al., 2019; Kaneko et al., 2020; Katsumata and Komachi, 2020; Rothe et al., 2021).

Low-resource GEC remains challenging, with researchers exploring synthetic data generation for Czech GEC (Naplava and Straka, 2019) and feedback comment generation for low-resource languages (Flachs et al., 2021). Our work differs from these by combining neural and symbolic approaches with explicit safety mechanisms, specifically, for extremely low-resource settings.

Multilingual models such as mBART and mT5 exhibit promising potential for cross-lingual transfer (Liu et al., 2020; Xue et al., 2021). Complementing this, Rothe et al. (2021) demonstrated that mT5 fine-tuning can achieve competitive GEC performance.

The neurosymbolic approach combines neural learning with symbolic reasoning. Recent works in this area include neural symbolic parsers, hybrid question answering, and rule-augmented neural models (Platanios et al., 2021; Mitra and Baral, 2016). For GEC specifically, Awasthi et al. (2019) combined neural models with rule-based post-editing for English.

Hu et al. (2021) demonstrated that LoRA (Low-Rank Adaptation) enables efficient fine-tuning by injecting trainable low-rank matrices into frozen pre-trained models, reducing trainable parameters by over 99% while maintaining performance.

LANGUAGE MODEL SELECTION FOR SEQUENCE-TO-SEQUENCE GEC

While monolingual BERT-based encoder models exist for both Tamil (l3cube-pune/tamil-bert) and Malayalam (l3cube-pune/malayalam-bert), these models are fundamentally unsuitable for GEC tasks due to their encoder-only architecture. GEC is inherently a sequence-to-sequence task requiring both encoding input sentences and generating corrected outputs, necessitating encoder-decoder architectures like T5 or BART.

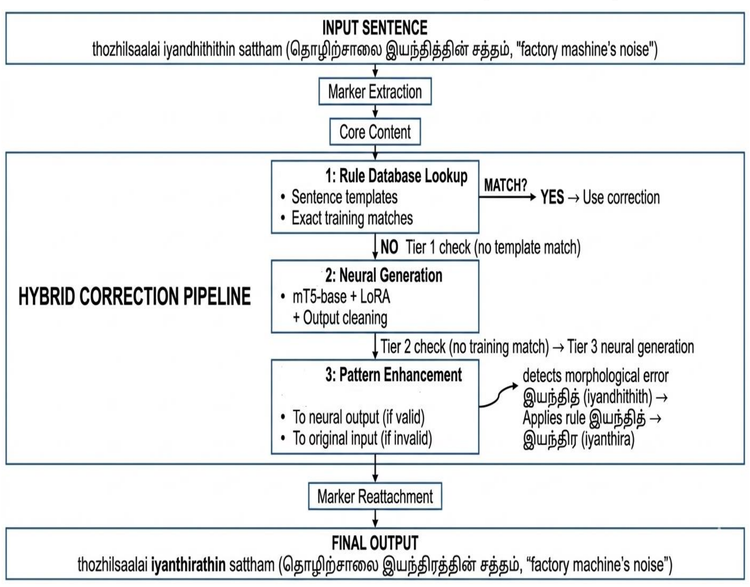
BERT-based models, being encoder-only, can only produce contextual representations and are designed for classification, token labelling, or extraction tasks rather than text generation. Adapting BERT for generation would require adding a decoder component from scratch, essentially reconstructing an encoder-decoder model without the benefits of pre-trained generation capabilities. Furthermore, no production-ready monolingual T5-style encoder-decoder models exist for Tamil or Malayalam in public repositories. While researchers have created language-specific adaptations by pruning multilingual models (e.g., Russian T5), similar efforts for Dravidian languages remain unpublished or unavailable.

Therefore, we leverage mT5, a multilingual T5 variant pre-trained on 101 languages including Tamil and Malayalam, which provides the necessary encoder-decoder architecture for GEC while offering cross-lingual transfer benefits from related languages. The mT5 family's availability in multiple sizes (small, base, large) enables capacity-driven design choices suitable for our low-resource setting, as demonstrated in our ablation studies.

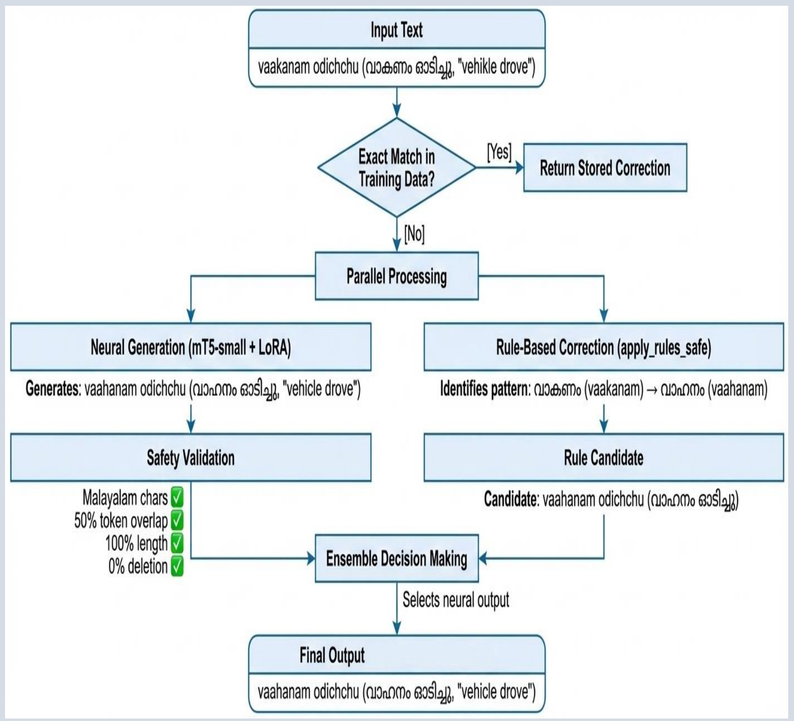
SYSTEM ARCHITECTURE

We present differentiated frameworks for Tamil and Malayalam GEC, reflecting language-specific requirements. This differentiation reflects Tamil's morphological complexity, which requires greater model capacity, and Malayalam's higher observed risk of neural over-correction, requiring conservative safety mechanisms.

Tamil GEC Architecture
The Tamil system employs a five-stage hierarchical pipeline that combines neural and symbolic approaches strategically. This prioritises correction coverage for complex morphology:
1. Marker Extraction: Regex-based isolation of formatting elements (-, ;-) from linguistic content
2. Rule-Based Priority Checking: Exact matching against sentence templates and training data; immediate return if matched
3. Neural Generation: mT5-base with LoRA adaptation, beam search (width 6, length penalty 0.8, repetition penalty 1.1)
4. Pattern Enhancement: Application of 25+ manual Tamil error patterns to refine neural outputs
5. Marker Reattachment: Deterministic restoration of original formatting



Malayalam GEC Architecture
The Malayalam system employs a conservative parallel processing pipeline with safety-first ensemble selection. This prioritises output reliability and stability:
1. Exact Match Check: Validation against learned corrections dictionary; immediate return if matched
2. Parallel Processing: Simultaneous neural generation (mT5-small + LoRA) and rule-based candidate preparation
3. Safety Validation: Neural outputs undergo comprehensive validation (character presence, token overlap, length ratios, deletion thresholds)
4. Ensemble Selection: Confidence-based choice between neural output, rule-based candidate, or original input fallback
5. Explicit Tracking: Usage statistics (neural used, rule used, exact used, fallback used) for transparency.



RESULTS AND DISCUSSSION

Dataset and Evaluation Setup:

- Tamil: 91 training pairs (augmented to 5,000), 16 validation pairs, 65 test inputs
- Malayalam: Limited training pairs (augmented to 10,000), validation set available, 102 test inputs
- Blind test evaluation: No gold standard outputs provided, simulating real-world deployment
- Primary metric: GLEU (balancing n-gram precision and recall)

Performance Analysis

Blind Test evaluation metrics

Language	GLEU	Overall Rank
Tamil	85.34%	8
Malayalam	95.06%	2

DATA AUGMENTATION AND TRAINING

Data augmentation strategies were designed specifically for each language to mitigate data scarcity through controlled noise injection

Data Augmentation Strategies

Tamil Augmentation (91 → 5,000 examples	Malayalam Augmentation (→ 10,000 examples)
- Vowel dropping: Targeting 12 Tamil vowels (அ, ஆ, இ, ஈ, உ, ஊ, எ, ஏ, ஒ, ஓ, ஐ, ஔ) - Character duplication and deletion - Punctuation perturbation - Word order shuffling - Each sentence underwent 1-2 random transformations (55-fold expansion)	- Vowel sign dropping: Targeting 12 Malayalam vowel signs (ാ, ി, ീ, ു, ൂ, ൃ, െ, േ, ൈ, ൊ, ോ, ൌ) - Safe character duplication/deletion (avoiding first two characters to prevent catastrophic truncation) - Adjacent word swapping (excluding first word to maintain sentence structure) - Comma spacing removal and punctuation normalization - Chillu variation handling: Modern-traditional pairs (ൻ/ന്, ൺ/ണ്, ൽ/ല്, ൾ/ള്, ർ/ര്, ൿ/ക്) - Quality filtering: Similarity filtering (0.6-0.98) and length preservation (≥50% original)

Rationale:

Controlled noise injection mimics natural error patterns while maintaining linguistic validity.

Quality filtering prevents learning spurious noise patterns

Configuration balances training efficiency with model quality for extremely low-resource settings, preventing overfitting while enabling effective adaptation.

ABLATION STUDY: MODEL CAPACITY ANALYSIS

To validate language-specific model selection, we conducted ablation experiments by swapping mT5 variants

Language	Model Configuration	Validation GLEU	Performance Delta
Tamil	mT5-base (proposed)	80.47%	Baseline
Tamil	mT5-small	75.17%	-5.30%
Malayalam	mT5-small (proposed)	55.21%	Baseline
Malayalam	mT5-base	55.03%	-0.18%

Key Findings from the Ablation Study:

1. Reduction from mT5-base to mT5-small resulted in substantial 5.30 percentage point GLEU degradation, demonstrating that Tamil's complex agglutinative morphology with extensive case marking and verb conjugations genuinely benefits from higher representational capacity (580M parameters, 12 layers).

2. Increasing capacity from mT5-small to mT5-base yielded negligible performance difference (0.18 points), with preliminary analysis revealing increased generation instability in larger model. This validates our conservative approach: lower capacity with strict safety validation provides optimal balance.

Thus, the ablation study empirically validates that our model selection was data-driven rather than arbitrary, reflecting fundamental differences in language complexity and dataset-specific generation stability characteristics.

ERROR ANALYSIS

Error analysis on test sets revealed the systems' capabilities across diverse error types.

Error Analysis: Tamil and Malayalam

Input Sentence	Hybrid Output	Correction Type
-தொழிற்சாலை இயந்தித்தின் சத்தம் *thozhilsaalai iyandhithithin sattham* / "factory mashine's noise"	தொழிற்சாலை இயந்திரத்தின் சத்தம் *thozhilsaalai iyanthirathin sattham* / "factory machine's noise"	Morphological இயந்தித் → இயந்திர *iyandhithith → iyanthira*
-போக்குவரத்து வாகணங்களின் ஹாரன் *-pokku varatthu vaakanangalin haaran* / "traffic vehikles' hron"	போக்குவரத்து வாகனங்களின் ஹார்ன் *pokku varatthu vaahanangalin haarn* / "traffic vehicles' horn"	Multi-token . ஹார்ன் -- ஹாரன் ; வாகணம் → வாகனம், *haaran → haarn, vaakanam → vaahanam*
இரயில் பயனத்தில் களைத்துப் போன எங்களுக்கு irayil payaṇattil kaḷaittup pōṉa eṅkaḷukku / "train journey in tired gone for us"	ரயில் பயணத்தில் களைத்து போன எங்களுக்கு rayil payaṇattil kaḷaittu pōṉa eṅkaḷukku / "train journey in tired gone for us"	Multiple Errors இரயில் → ரயில்,பயனம் → பயணம், irayil → rayil, payaṉam → payaṇam
വാകണം ഓടിച്ചു *vaakanam odichchu* / "vehikle drove"	വാഹനം ഓടിച്ചു *vaahanam odichchu* / "vehicle drove"	Spelling correction വാകണം → വാഹനം / *vaakanam → vaahanam*
ധ്വനി മലിനീകരണത്തിന് കാരണങ്ങൾ *dhvani malineekaraṇaththinu kāraṇaṅṅaḷ* / "noise pollution's reasons"	ധ്വനി മലിനീകരണത്തിന് കാരണങ്ങൾ *dhvani malineekaraṇaththinu kāraṇaṅṅaḷ* / "noise pollution's reasons"	Token-level preservation

DISCUSSIONS

Extremely low resource GEC requires hybrid approaches with optimal balance between the neural and symbolic rule-based components depending on the language characteristics, dataset size, and deployment priorities. While direct comparison remains limited due to novel datasets, our results surpass reported low-resource GEC performance. Czech GEC with synthetic augmentation achieved approximately 60-70% accuracy with similar data constraints (Naplava & Straka, 2019), while our hybrid approach achieved 85.34% GLEU for Tamil and 95.06% GLEU for Malayalam, demonstrating viability for extreme low-resource scenarios.

LIMITATIONS AND FUTURE WORK

Limitations include:

1. Statistical Confidence: Small training (91 examples for Tamil) and validation datasets (16 examples for Tamil) limit statistical confidence in generalization.
2. Pattern Coverage Gaps:Manual patterns for Tamil (25+) and automated phrase-level extraction for Malayalam (6-token limit) cannot exhaustively address all possible grammatical errors.
3. Architectural Limitations: Ablation experiments conducted only with mT5 variants; alternative multilingual encoder-decoder architectures (mBART, ByT5) unexplored.

Future Research Directions

1. Adaptive Safety Mechanisms**:** Develop dynamic threshold adjustment based on input characteristics
2. Cross-Lingual Transfer: Investigate knowledge transfer between related Dravidian languages.
3. Automated Pattern Discovery: Explore grammar induction or constituency parsing for automated pattern discovery
4. Monolingual Model Development: Address resource gap by developing production-ready monolingual T5-style encoder-decoder models for Dravidian languages

CONCLUSION

The Neurosymbolic systems prove that combining modern pre-trained models, parameter-efficient fine-tuning, aggressive augmentation, and linguistic rule engineering provides a powerful practical approach for GEC.

REFERENCES

- Zhao, W., Wang, L., Shen, K., Jia, R., & Liu, J. (2019). Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In Proceedings of NAACL-HLT, (pp. 156-165).
- Kaneko, M., Mita, M., Kiyono, S., Suzuki, J., & Inui, K. (2020). Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In Proceedings of ACL, (pp. 4248-4254)
- Naplava, J., & Straka, M. (2019). Grammatical error correction in low-resource scenarios. In Proceedings of the 5th Workshop on Noisy User-generated Text, (pp. 346-356).
- ACL-IJCNLP, (pp. 702-707).