A3-108 at BHASHA Task1: *Asymmetric BPE configuration* for Grammar Error Correction

Saumitra Yadav Manish Shrivastava

LTRC, KCIS, IIIT-Hyderabad

saumitra.yadav@research.iiit.ac.in, m.shrivastava@iiit.ac.in

Main Contributions

Low-resource GEC for Indic languages: GEC systems for Bangla, Hindi, Malayalam, Tamil, and Telugu developed for BASHA Task 1 under strict low-resource settings.

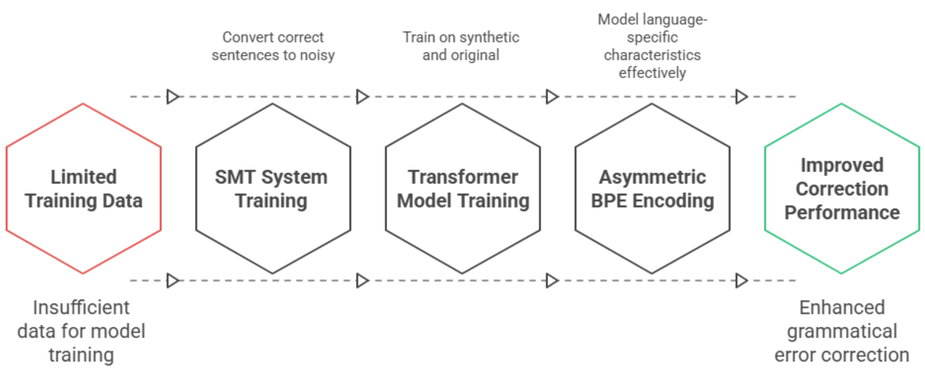
Two-step data-efficient training:
SMT trained on minimal parallel data
Large-scale synthetic noisy-to-clean data generated from monolingual corpora (clean).

Asymmetric subword modeling: Different BPE merge operations for source (erroneous) and target (corrected) text improve correction in morphologically rich languages.

Competitive shared-task results: Strong performance in BASHA Task 1 (e.g., Rank 4 for Malayalam, Rank 5 for Tamil).

Generalizable framework: A scalable, language-agnostic, and cost-effective approach for low-resource GEC.

Pipeline for GEC for Indic Languages



1. Treat Grammatical Error Correction as monolingual Machine Translation.
2. **Pipeline stages:**
   a. Train **SMT model** on limited parallel GEC data
   b. Use SMT to generate **synthetic noisy sentences** from clean monolingual text
2. **Two types of sentence pairs:**
   a. Synthetic noisy-to-clean pairs
   b. Identity (clean-to-clean) pairs
3. Pair noisy sentences as **source** with original clean sentences as **target**

Dataset

Language	Data Made Available				Generated Data (approx.)		
	Train	Train (No Id)	Val	Test	Synthetic	Identical	Different
Bangla	659	418	103	331	446K	35K	411K
Hindi	600	541	108	237	461K	254K	207K
Malayalam	313	294	51	103	492K	251K	241K
Tamil	91	91	17	66	487K	270K	217K
Telugu	604	552	101	316	483K	251K	232K

Table 1. Data provided by the organizers and synthetic data generated for training GEC models.

Results

Source BPE	Target BPE	GLEU
8000	500	93.78
16000	500	93.92
4000	**4000**	**93.99**
4000	500	93.75
8000	4000	93.97
8000	2000	93.47
4000	2000	94.04
8000	1000	93.88
4000	1000	93.96
4000	**3000**	**94.16**

Table 2. Malayalam

Source BPE	Target BPE	GLEU
8000	500	84.44
16000	500	84.87
4000	**4000**	**85.05**
4000	500	84.86
8000	**4000**	**85.52**
8000	2000	85.26
4000	2000	85.50
8000	1000	84.42
4000	1000	85.25
4000	3000	84.74

Table 3. Tamil

Source BPE	Target BPE	GLEU
8000	500	91.71
16000	500	91.68
4000	**4000**	**92.45**
4000	500	91.65
8000	4000	92.35
8000	2000	92.35
4000	2000	92.19
8000	1000	91.44
4000	1000	92.14
4000	**3000**	**92.44**

Table 4. Bangla

Source BPE	Target BPE	GLEU
8000	500	79.94
16000	500	80.07
4000	**4000**	**81.90**
4000	500	81.18
8000	4000	80.78
8000	2000	80.72
4000	**2000**	**81.68**
8000	1000	80.39
4000	1000	80.68

Table 5. Telugu

Source BPE	Target BPE	GLEU
8000	500	79.27
16000	500	79.08
4000	**4000**	**79.45**
4000	500	79.27
8000	4000	79.27
8000	2000	79.39
4000	2000	78.70
8000	**1000**	**79.38**
4000	1000	78.93
4000	3000	79.29

Table 6. Hindi

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

AFNLP IJCNLP AACL 2025 December 20-24 Mumbai, India

More on Asymmetric BPE for MT:
To be presented in WAT2025