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

Saumitra Yadav and Manish Shrivastava
Language Technologies Research Center, KCIS,
International Institute of Information Technology Hyderabad, India
saumitra.yadav@research.iiit.ac.in and m.shrivastava@iiit.ac.in

Abstract

This paper presents our approach to Grammatical Error Correction (GEC) for five low-resource Indic languages, a task severely limited by a scarcity of annotated data. Our core methodology involves two stages: synthetic data generation and model optimization. First, we leverage the provided training data to build a Statistical Machine Translation (SMT) system, which is then used to generate large-scale synthetic noisy-to-clean parallel data from available monolingual text. This artificially corrupted data significantly enhances model robustness. Second, we train Transformer-based sequence-to-sequence models using an asymmetric and symmetric Byte Pair Encoding (BPE) configuration, where the number of merge operations differs between the source (erroneous) and target (corrected) sides to better capture language-specific characteristics. For instance, source BPE sizes 4000, 8000 and 16000, with target sizes at 500, 1000, 2000, 3000 and 4000. Our experiments demonstrated competitive performance across all five languages, with the best results achieving a GLUE score of 94.16 for Malayalam (Rank 4th) followed by Bangla at 92.44 (ranked 5th), Tamil at 85.52 (ranked 5th), Telugu at 81.9 (7th), and Hindi at 79.45(10th) in the shared task. These findings substantiate the effectiveness of combining SMT-based synthetic data generation with asymmetric BPE configurations for low-resource GEC.

1 Introduction

Grammatical error correction (GEC) in Indian languages is a vital yet challenging research area due to the complex morphological nature, rich syntactic structures, and diverse scripts prevalent among these languages (Bhattacharyya and Bhattacharya, 2025; Sharma and Bhattacharyya, 2025b,a). The digital proliferation of Indic languages such as Hindi, Tamil, Telugu, Bangla, and Malayalam has highlighted the importance of automated GEC systems to assist language learners, enhance machine translation, and improve natural language understanding.

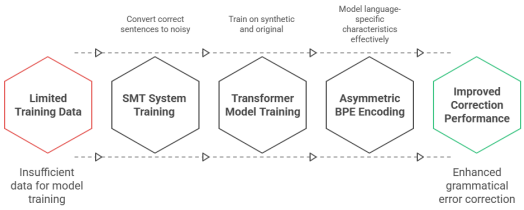

Figure 1: Our Pipeline for GEC for Indic Languages.

Unlike English and other widely studied languages, Indian languages exhibit a high degree of inflection and derivation, complicating both error detection and correction tasks. Moreover, the limited availability of large-scale, annotated parallel corpora with grammatical errors and corresponding corrections presents a significant barrier to training effective GEC models (Felice and Yuan, 2014). These challenges have motivated research into data augmentation, synthetic error generation(Wang et al., 2024; Deng et al., 2025) for GEC. To address this resource limitation, we follow the paradigm of viewing GEC as a monolingual machine translation task, enabling us to leverage existing monolingual corpora for synthetic data augmentation (Junczys-Dowmunt et al., 2018). As summarized in Figure 1, our pipeline begins with generating erroneous-to-correct sentence pairs using an SMT system trained on the parallel data provided.

Beyond data augmentation, the choice of subword tokenization is crucial for morphologically rich, low-resource languages (Ding et al., 2019; Abid, 2020). While BPE (Sennrich et al., 2016) hyperparameters have been explored, most research employs symmetrical BPE (same number of merge operations/NMO for source and target) (Huck et al.,

Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025), pages 147–154

2017; Ortega et al., 2020; Lankford et al., 2021; Domingo et al., 2019; Lee et al., 2024). Given that GEC involves translating noisy, often fragmented input to clean, correct output, we use flexibility offered by subword tokenization to increase performance of systems in this regard. Building on work that uses differing NMOs for word alignment (Ngo Ho and Yvon, 2021), we propose and systematically investigate the use of asymmetric BPE configurations for GEC to better model the distinct characteristics of the erroneous (source) and corrected (target) sides.

We generate synthetic noisy data by training a statistical machine translation (SMT) system on the small provided training data to convert correct sentences sampled from Doddapaneni et al. (2023) into noisy (erroneous) sentences. This synthetic data, paired with the original correct sentences, is then used to train transformer-based models for grammatical error correction across five Indic languages while employing asymmetric Byte Pair Encoding, using different number of merge operations for source and target tokenization model, to effectively model language-specific characteristics and improve correction performance.

2 Related Work

2.1 Grammar Error Correction as MT

Since the early work of Brockett et al. (2006), Grammatical Error Correction (GEC) systems have often employed a monolingual machine translation approach, training models to map erroneous sentences directly to their corrected counterparts.

Statistical machine translation (SMT) played an important role in GEC research. Yuan and Felice (2013) and Wang et al. (2014) provide evidence regarding SMT's capabilities and limitations, particularly in addressing complex error types and local contexts. Felice and Yuan (2014) discuss the use of SMT systems trained on learner data to artificially generate noisy sentences from correct sentences, effectively enriching training data for translation-based correction models.

Xie et al. (2016) presented a neural network-based approach to GEC, employing a character-level encoder-decoder recurrent neural network with an attention mechanism.

Hoang et al. (2016) and Chollampatt et al. (2016) utilized machine translation systems enhanced with a feed-forward neural translation model and n-best list re-ranking methods to improve correction accuracy. Sequence-to-sequence methods were further explored by Yuan and Briscoe (2016), Chollampatt and Ng (2018), and Yuan et al. (2019). Junczys-Dowmunt et al. (2018) highlighted strategies leveraging larger annotated corpora and data augmentation to overcome resource limitations, drawing parallels between low-resource machine translation and grammatical error correction.

The generation of synthetic error-laden data has been systematically studied by Htut and Tetreault (2019), who compare rule-based and neural approaches for artificial error creation. Building on this, Stahlberg and Kumar (2021) introduce tagged corruption models to create large-scale synthetic datasets, such as C4_200M, which improve the performance of neural GEC systems. More recently, Wang et al. (2024) propose a contextual data augmentation technique that combines rule-based and model-based generation methods, followed by relabeling to reduce noise in synthetic data. Deng et al. (2025) focus on automatic synthetic data generation within an unsupervised GEC framework.

Transformer-based language models have also shown remarkable effectiveness in GEC. Alikaniotis and Raheja (2019) demonstrate that transformers outperform conventional recurrent models, providing a strong baseline for future research. Furthermore, Kubal and Nagvenkar (2025) explore multilingual transformer architectures for robust correction across diverse languages. Bhattacharyya and Bhattacharya (2025) used LLMs to improve GEC for Bangla. Together, these studies establish the translation paradigm as central to the development of powerful grammatical error correction systems. They also underscore the importance of synthetic data generation, model augmentation, and hybrid strategies in improving grammaticality—especially in low-resource scenarios.

2.2 Symmetric BPE Configuration

In many bilingual machine translation (MT) systems, especially in low-resource scenarios, it is a common practice to apply the same number of merge operations (NMO) for both source and target languages when using Byte Pair Encoding (BPE). Several studies have adhered to this symmetry: Ding et al. (2019) observed that smaller vocabulary sizes (0–4K NMO) can outperform the widely used 32K setting by up to 4 BLEU points in low-resource transformer setups. Similar trends have been reported in English–Egyptian and English–Levantine (Abid, 2020), as well as

English–Irish (Lankford et al., 2021) translation tasks.

Other research has adapted segmentation strategies to account for typological or morphological characteristics of languages. For instance, segmentation restrictions for polysynthetic languages were proposed by Ortega et al. (2020), while Lee et al. (2024) addressed over-segmentation issues in morphologically rich languages. Target-side tokenization variations have also been explored to better capture language-specific features (Domingo et al., 2019). Alternative approaches include cascading segmentations (Huck et al., 2017), vocabulary refinement through VOLT (Xu et al., 2021), and concatenation of corpora tokenized with multiple BPE settings (Poncelas et al., 2020). Ngo Ho and Yvon (2021) experimented with differing NMO settings on source and target sides to improve word alignment, though this did not extend to training MT systems with asymmetric BPE. Yadav and Shrivastava (2025) extensively experimented with asymmetric BPE and showed efficacy of using asymmetric BPE while training NMT models in low resource setting for multiple language pairs.

Our present work builds on these foundation by generating additional noisy-to-correct sentence pairs to expand parallel training data and reinforce the effectiveness of translation-based approaches for grammar correction.

3 Data and Synthetic Generation

The dataset statistics, shown in Table 1, include the initial data provided by the organizers and the sentences we generated. The raw *Training Data* was cleaned by excluding pairs where the source and target sentences were identical. The *Validation* and *Test* sets comprise the incorrect sentences utilized for system development and final performance assessment.

The synthetic dataset is generated via a Statistical Machine Translation (SMT) system (Koehn et al., 2003), which is highly effective for low-resource translation (Koehn and Knowles, 2017). We train the SMT to model the error-generating process, then apply it in reverse to 0.45 million clean monolingual sentences per language (Doddapaneni et al., 2023) to create pairs of "incorrect" (noisy) input and correct output. This method ensures the synthetic data reflects realistic error patterns. The SMT utilizes symmetric BPE with 500 merge operations. Table 1 provides a breakdown of this generated corpus, indicating the percentage of sentences that remained *Identical* or were generated *Different* then correct monolingual text.

The final training set for the Transformer models (Vaswani et al., 2017) was a combination of two types of sentence pairs: the SMT-generated noisy-to-clean synthetic pairs, and a crucial set of identity pairs (correct sentence to correct sentence). Including these identity pairs ensures the model learns not only how to correct errors but also the identity function—that is, how to preserve correct sentences when no error is detected, thereby preventing unnecessary over-correction. For validation set we use training data provided by the organizers. All the datasets are preprocessed using Indic NLP library (Kunchukuttan, 2020).

4 Experimental Setup and Results

Then we train a transformer model using Fairseq (Ott et al., 2019) with hyperparameters and gpu usage given in Appendix A. For subword tokenization we use both symmetric ($m = n$) and asymmetric ($m > n$) BPE for incorrect (source) and correct (target) respectively, where m and n are respective NMOs. For source we chose 16K, 8K, 4K and for target we chose 500, 1K, 2K, 3K and 4K. GLEU was used for calculating the performance of each system. For clarity we are showing only top performing models for each langauge and their respective ranks in leaderboard (Table 2). Performance on other *BPE configurations* are given in appendix B.

The best-performing configurations (Table 2) were overwhelmingly asymmetric, such as the (Source BPE 4K, Target BPE 3K) pairing for both Malayalam and Bangla, and (8K source BPE, 4K target BPE) for Tamil. This empirically confirms our hypothesis that distinct tokenization granularities are beneficial for modeling the noisy source and clean target spaces in GEC. From a learning standpoint, using a smaller decoder-side vocabulary encourages tighter coupling between source and target representations, which facilitates more reliable alignment and mapping of segments, echoing observations on subword choices and alignment behavior in prior work (Ngo Ho and Yvon, 2021). This is consistent with earlier evidence (Domingo et al., 2019) that target-side vocabulary design has a direct impact on NMT effectiveness.

	Data made available				Generated		
Language	**Training Data**	**After removing identical sentence**	**Validation**	**Test**	**Synthetic**	**Identical**	**Different**
Bangla	659	418	103	331	446,805	35,622	411,183
Hindi	600	541	108	237	461,862	254,039	207,823
Malayalam	313	294	51	103	492,248	251,325	240,923
Tamil	91	91	17	66	487,344	270,074	217,270
Telugu	604	552	101	316	483,696	251,634	232,062

Table 1: Data shared by organizers and generated by us for our models.

Languages	Source BPE	Target BPE	GLEU Score	Rank
Malayalam	4000	3000	94.16	4
Bangla	4000	3000	92.44	5
Telugu	4000	4000	81.9	5
Tamil	8000	4000	85.52	7
Hindi	4000	4000	79.45	10

Table 2: GLEU score of models with BPE configuration (source, target BPE) and respective Ranks in the leaderboard

5 Future Work and Conclusion

This research successfully validates the combined utility of SMT-based data augmentation and asymmetric Byte Pair Encoding (BPE) for Grammatical Error Correction (GEC) in low-resource settings. Building on these promising results, several key areas remain for future investigation:

5.1 Scaling Data Augmentation and Quality Control

While the current work demonstrated strong performance using approximately 0.45 million synthetic sentences per language, a critical next step is to evaluate the effects of massive-scale data augmentation. This involves utilizing the entirety of available monolingual corpora (such as the full Doddapaneni et al. (2023) dataset) to push the synthetic data volume into the millions.

5.2 Ablation Study on Training Data Composition

Our current model is trained on a mixture of synthetic noisy-to-clean pairs, and identity pairs (correct-to-correct sentences). An important ablation study would be to isolate the components of the training set. Specifically, we plan to train models exclusively on the synthetic noisy-to-clean pairs, removing the identity pairs. This experiment would conclusively determine the true generalization capability of the SMT-generated errors and quantify the necessity of training the model on the identity function to prevent over-correction.

5.3 Generalization to Other Low-Resource Languages

The demonstrated effectiveness of our approach for morphologically rich Indic languages suggests its broad applicability. We aim to expand this methodology to GEC tasks in other low-resource languages. The combination of leveraging readily available monolingual text for synthetic error generation and fine-tuning subword tokenization via asymmetric BPE gives possibility of threads of experiments for any language pair where parallel error-annotated data is scarce.

We presented our effective Grammatical Error Correction (GEC) systems for five Indic languages—Bangla, Hindi, Malayalam, Tamil, and Telugu—developed as a low-resource solution to the BASHA Task 1 shared challenge. We did this in two step approach. First, we leveraged a minimal training set to train a Statistical Machine Translation (SMT) system, which was then used to generate large-scale, contextually relevant synthetic noisy-to-clean sentence pairs from extensive monolingual text. Second, we demonstrated the critical importance of asymmetric Byte Pair Encoding (BPE) configurations. By systematically applying different numbers of merge operations for the source (erroneous) and target (corrected) vocabularies, we were able to tailor the subword segmentation to build good models.The results, which placed our systems competitively in the shared task (e.g., Rank 4th for Malayalam and 5th for Bangla), provide strong empirical evidence for the combined

benefits of synthetic data and optimized subword tokenization. This work validates a highly resource-efficient and generalizable methodology for advancing GEC capabilities in morphologically complex, low-resource language environments.

Limitation

The primary constraint of this work stems from the inherent computational expense associated with exhaustively training and evaluating diverse Byte Pair Encoding (BPE) configurations across all target languages. This practical limitation necessitated a focused selection of configurations. Additionally, applying the insights derived from this study to the context of decoder-only architectures is anticipated to introduce considerable technical challenges that warrant further investigation. Moving forward, the scope of this research could be significantly enhanced by utilizing larger training datasets and dedicating focused effort to investigating and improving the quality and efficacy of synthetic data generation.

References

Wael Abid. 2020. The SADID evaluation datasets for low-resource spoken language machine translation of Arabic dialects. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6030–6043, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dimitris Alikaniotis and Vipul Raheja. 2019. The unreasonable effectiveness of transformer language models in grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 127–133, Florence, Italy. Association for Computational Linguistics.

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.

Chris Brockett, William B. Dolan, and Michael Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 249–256, Sydney, Australia. Association for Computational Linguistics.

Shamil Chollampatt and Hwee Tou Ng. 2018. Neural quality estimation of grammatical error correction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2528–2539, Brussels, Belgium. Association for Computational Linguistics.

Shamil Chollampatt, Kaveh Taghipour, and Hwee Tou Ng. 2016. Neural network translation models for grammatical error correction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 2768–2774. AAAI Press.

Jiayi Deng, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2025. InstructGEC: Enhancing unsupervised grammatical error correction with instruction tuning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 110–122, Abu Dhabi, UAE. Association for Computational Linguistics.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

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.

Miguel Domingo, Mercedes García-Martínez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. 2019. How much does tokenization affect neural machine translation? In *Computational Linguistics and Intelligent Text Processing: 20th International Conference, CICLing 2019, La Rochelle, France, April 7–13, 2019, Revised Selected Papers, Part I*, page 545–554, Berlin, Heidelberg. Springer-Verlag.

Mariano Felice and Zheng Yuan. 2014. Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–126, Gothenburg, Sweden. Association for Computational Linguistics.

Duc Tam Hoang, Shamil Chollampatt, and Hwee Tou Ng. 2016. Exploiting n-best hypotheses to improve an smt approach to grammatical error correction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 2803–2809. AAAI Press.

Phu Mon Htut and Joel Tetreault. 2019. The unbearable weight of generating artificial errors for grammatical error correction. In *Proceedings of the Fourteenth*

Workshop on Innovative Use of NLP for Building Educational Applications, pages 478–483, Florence, Italy. Association for Computational Linguistics.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67, Copenhagen, Denmark. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, New Orleans, Louisiana. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Divesh Ramesh Kubal and Apurva Shrikant Nagvenkar. 2025. Leveraging multilingual models for robust grammatical error correction across low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 505–510, Abu Dhabi, UAE. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Seamus Lankford, Haithem Alfi, and Andy Way. 2021. Transformers for low-resource languages: Is féidir linn! In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 48–60, Virtual. Association for Machine Translation in the Americas.

Jungseob Lee, Hyeonseok Moon, Seungjun Lee, Chanjun Park, Sugyeong Eo, Hyunwoong Ko, Jaehyung Seo, Seungyoon Lee, and Heuiseok Lim. 2024. Length-aware byte pair encoding for mitigating oversegmentation in Korean machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2287–2303, Bangkok, Thailand. Association for Computational Linguistics.

Anh Khoa Ngo Ho and François Yvon. 2021. Optimizing word alignments with better subword tokenization. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 256–269, Virtual. Association for Machine Translation in the Americas.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Alberto Poncelas, Jan Buts, James Hadley, and Andy Way. 2020. Using multiple subwords to improve English-Esperanto automated literary translation quality. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 108–117, Suzhou, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ujjwal Sharma and Pushpak Bhattacharyya. 2025a. HiGEC: Hindi grammar error correction in low resource scenario. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6063–6075, Abu Dhabi, UAE. Association for Computational Linguistics.

Ujjwal Sharma and Pushpak Bhattacharyya. 2025b. IndiGEC: Multilingual grammar error correction for low-resource Indian languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22393–22407, Suzhou, China. Association for Computational Linguistics.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yiming Wang, Longyue Wang, Xiaodong Zeng, Derek F. Wong, Lidia S. Chao, and Yi Lu. 2014. Factored statistical machine translation for grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 83–90, Baltimore, Maryland. Association for Computational Linguistics.

Yixuan Wang, Baoxin Wang, Yijun Liu, Qingfu Zhu, Dayong Wu, and Wanxiang Che. 2024. Improving grammatical error correction via contextual data augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10898–10910, Bangkok, Thailand. Association for Computational Linguistics.

Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky, and Andrew Y. Ng. 2016. Neural language correction with character-based attention. *Preprint*, arXiv:1603.09727.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021. Vocabulary learning via optimal transport for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7361–7373, Online. Association for Computational Linguistics.

Saumitra Yadav and Manish Shrivastava. 2025. Segmentation beyond defaults: Asymmetrical byte pair encoding for optimal machine translation performance. *Preprint*, arXiv:2511.03383.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California. Association for Computational Linguistics.

Zheng Yuan and Mariano Felice. 2013. Constrained grammatical error correction using statistical machine translation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 52–61, Sofia, Bulgaria. Association for Computational Linguistics.

Zheng Yuan, Felix Stahlberg, Marek Rei, Bill Byrne, and Helen Yannakoudakis. 2019. Neural and FST-based approaches to grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 228–239, Florence, Italy. Association for Computational Linguistics.

A Training Hyperparameters

Table 3 gives hyperparameters we used for training GEC systems. And Table 4 shows the gpu hours used with respective GPUS to train these models.

B Performance for All BPE configurations

Table 5 shows performance of all BPE configurations for GEC for all the languages. Due to resource constraints we didnt explore all possibilities only some promising ones.

Parameter	Value
arch	transformer
optimizer	adam
adam-betas	(0.9, 0.98)
clip-norm	0.0
lr	5e-4
lr-scheduler	inverse_sqrt
warmup-updates	4000
warmup-init-lr	1e-07
dropout	0.3
attention-dropout	0.1
activation-dropout	0.1
weight-decay	0.0001
criterion	label_smoothed_cross_entropy
label-smoothing	0.1
max-tokens	30000
max-update	300000
patience	20
update-freq	10

Table 3: Training hyperparameters used across all experiments.

GPUs	GPU Hours
4090 RTX	356.86
2080 TI	100.64

Table 4: GPU usage for training the models.

Language	Source BPE	Target BPE	GLUE Score	Language	Source BPE	Target BPE	GLUE Score
Bangla	8000	500	91.71	Malayalam	8000	2000	93.47
	16000	500	91.68		4000	2000	94.04
	4000	4000	92.45		8000	1000	93.88
	4000	500	91.65		4000	1000	93.96
	8000	4000	92.35		**4000**	**3000**	**94.16**
	8000	2000	92.35	Tamil	8000	500	84.44
	4000	2000	92.19		16000	500	84.87
	8000	1000	91.44		4000	4000	85.05
	4000	1000	92.14		4000	500	84.86
	4000	**3000**	**92.44**		**8000**	**4000**	**85.52**
Hindi	8000	500	79.27		8000	2000	85.26
	16000	500	79.08		4000	2000	85.5
	4000	**4000**	**79.45**		8000	1000	84.42
	4000	500	79.27		4000	1000	85.25
	8000	4000	79.27		4000	3000	84.74
	8000	2000	79.39	Telugu	8000	500	79.94
	4000	2000	78.7		16000	500	80.07
	8000	1000	79.38		**4000**	**4000**	**81.9**
	4000	1000	78.93		4000	500	81.18
	4000	3000	79.29		8000	4000	80.78
Malayalam	8000	500	93.78		8000	2000	80.72
	16000	500	93.92		4000	2000	81.68
	4000	4000	93.99		8000	1000	80.39
	4000	500	93.75		4000	1000	80.68
	8000	4000	93.97				

Table 5: GLEU score of models with BPE configuration (m,n) with Bold marking the top performing from respective languages and BPE configurations