

MarathiGEC: A Transformer-Based Grammar Error Correction Framework for Marathi Using Fine-Tuned IndicBART

Ashwini Pijdurkar¹, Samarth Sabale², Krushna Yeole³, Imam Shaikh⁴,

^[1] Assistant Professor, ^[2-4] Students of Computer Engineering

^[1-4] Department of Computer Engineering, SPPU University/Zeal College of Engineering and Research

¹ ashwini.pijdurkar@zealeducation.com, ² samarthstable2907@gmail.com, ³ yeolekrushnar@gmail.com,
⁴ emamshaikh2003@gmail.com

Abstract – Grammar Error Correction (GEC) is an essential problem in the Natural Language Processing (NLP) domain where the objective is to detect and automatically correct the grammatical mistakes present in any piece of text. Though there have been many advanced grammar correction models developed for high resource languages like English, robust grammar correction tools are yet to be created for regional Indian languages. The Indo-Aryan language, Marathi, poses distinct challenges in grammar correction owing to its heavy morphology, gender based verb agreement, complex sentence structure, and dependencies on context. This work introduces MarathiGEC, a transformer based Marathi Grammar Error Correction tool using the IndicBART model. An annotated corpus of sentence pairs which contain grammatically wrong sentences and their corrections in Marathi language were created to train the IndicBART model. The presented GEC model uses a sequence to sequence text generation method and thus can correct grammatical errors in the context of their meaning. Experimentation proves the effectiveness of the model to correct errors in gender agreement, verb conjugation, tense consistency, misspellings, and sentence structure..

Keywords: Deep Learning, Grammar Error Correction, Marathi Grammar correction, MarathiGEC, Language Analyzer, Marathi NLP, Transformer Models, Low Resource Languages

I. INTRODUCTION

Among the domains of artificial intelligence, natural language processing has come to be regarded as one of the most significant fields, making it possible for machines to process natural languages. Deep learning, along with other recent innovations, has led to better performances of NLP tasks such as machine translation, text summarization, sentiment analysis, question answering, and grammar error correction. Of these, GEC (grammar error correction) is one of the most significant tasks which help improve written communication through automatic detection and correction of any grammatical, syntactical, and contextual errors in texts. Modern GEC systems have thus come to be essential in education and professional settings, as well as in digital communications.

Despite the phenomenal success in developing grammar checking tools for highly resourceful languages like English, French, and Chinese, there is little research being conducted on developing such tools for the regional languages of India. Marathi, which is one of the most popular languages of the Indo-Aryan family of languages, is used by millions of speakers of the language. However, there are some unique problems that can be seen in the Marathi language which make the process of grammar correction very difficult. There are complicated morphological structures, inflections, gender-based verbs, tense-based conjugations, flexible word order, and contextual grammar rules in the language. This results in a number of errors in the grammar of the language.

Marathi text. However, there is still room for improvement concerning their capacity to perform well in relation to grammar rules that are unique to Marathi, which include, but are not limited to, gender agreement, verb conjugations, number agreements, case marking, and correcting erroneous sentences depending on context. Multilingual models lack fine-tuning regarding Marathi grammar since they are primarily trained on multilingual datasets, thus performing grammatically or contextually incorrect corrections. The lack of annotated grammar correction benchmarks in the Marathi language makes it low-resource in terms of Grammar Error Correction.

The developments in transformer models that use the sequence-to-sequence paradigm have provided promising possibilities to address these challenges. For example, IndicBART is a multilingual transformer-based model that can be used for Indian languages. IndicBART has performed well on different text generation and language understanding tasks. Unlike rule-based methods, IndicBART uses context representation and attention mechanism to identify the patterns of language. As a result, the ability to identify the patterns of language directly from the data makes this model useful for grammar correction tasks.

For the current study, MarathiGEC framework, which is transformer-based Marathi grammar error correction framework using IndicBART architecture, is proposed. MarathiGEC learns context-dependent mappings from erroneous sentences to corrected sentences by treating grammar correction problem as a sequence-to-sequence text generation task. The model is trained on a specially curated parallel corpus of grammatically erroneous and corrected Marathi sentences. By capturing the context-dependent relationship between erroneous sentences and corrected sentences, the model can automatically correct different types of grammatical errors such as gender agreement errors, verb conjugation errors, tense errors, spelling errors, and structure errors.

This project's main contribution is twofold. First, a parallel dataset of Marathi Grammar Error Correction (incorrect and grammatically corrected sentence pairs) is created to enable supervised training of the model. Secondly, the IndicBART model is fine-tuned for Marathi grammatical error correction in order to generate context-aware corrections. Moreover, the efficiency of the proposed approach is estimated using the traditional NLP evaluation metrics. With this project, we will be able to contribute to Marathi Natural Language Processing studies and create a basis for developing intelligent writing assistants in Indian regional languages.

II. LITERATURE REVIEW

Grammar Error Correction (GEC) has drawn considerable interest in NLP in view of its practical use in intelligent writing assistance systems. The early work in the field focused on the use of rules-based and statistical methods for identifying and fixing grammatical errors. In this context, Arora et al. [1] made an extensive review of GEC techniques and discussed the shift from conventional linguistic approaches to machine learning and deep learning-based ones. Despite the recent progress achieved in the field through deep learning and statistical methods, the efficiency of these techniques depends on annotated large-scale data sets that are difficult to obtain for Marathi. Though they can process Mar

Low-resource Indian language research is becoming increasingly prevalent. For instance, Sonawane et al. [2] offered a Hindi grammar correction framework using inflectional error generation and demonstrated the relevance of synthetic error data sets. Likewise, Sharma and Bhattacharyya [3], [8] designed the Hi-GEC benchmark data set for Hindi grammatical error correction.

Several works have helped in developing Marathi NLP resources and tools. Firstly, Gawali et al. [5] created a morphology-based Marathi spell checker, and secondly, Khandale and Mahender [10] suggested a rule-based approach for discourse analysis in Marathi texts. In addition, Singh et al. [12] and Deshpande and Gore [13] designed statistical and hybrid POS tagging systems, which were crucial for the linguistic basis of Marathi NLP.

With recent developments in transformer-based language models, Marathi language processing research has progressed significantly. For instance, Joshi [6] created L3Cube-MahaCorpus and MahaBERT; both are Marathi language resources and pre-trained language models. Moreover, Joshi et al. [4] built MahaNLP – an open source framework that provides support for Marathi NLP tasks using transformer-based models.

The research carried out in other low-resource languages further confirms the success of transformer-based methods. The grammatical error correction system for Gujarati language was created by Raj [9], and the morphological analyzer for Sinhala language was introduced by Kumarasinghe et al. [7]. This shows the increasing application of deep learning models for resource-deficient languages.

However, very little has been done on Marathi Grammar Error Correction using transformer-based models. All Marathi Natural Language Processing studies have focused mainly on spell correction, POS tagging, and language modeling, while there are few datasets available for grammar correction. To fill this gap, the proposed study uses IndicBART, a sequence-to-sequence transformer model, which is fine-tuned using parallel data consisting of incorrect and correct sentences in Marathi language.

III. MATERIAL AND METHODOLOGIES:

The MarathiGEC system being considered employs the sequence-to-sequence transformer framework that incorporates the IndicBART for automating the process of identifying and correcting grammatical mistakes in the Marathi language. This framework is intended to enable the learning of contextual correlation between the wrong sentences and their corrections to generate fluent outputs. The entire process involves various steps such as data gathering, dataset preparation, data preprocessing, model fine-tuning, grammar correction, and evaluation.

The first step is the acquisition of Marathi texts from different sources like educational materials, books, newspapers, blogs on the internet, and literary texts. For the creation of a parallel dataset that can be used in supervised learning, wrong sentence variants were generated, which were then verified manually against the correctly formed sentences. This led to the generation of a parallel dataset, where each record comprised a pair of wrong Marathi sentences and their grammatically correct forms.

Post acquisition, the dataset was arranged in the form of a parallel corpus that could be used for sequence to sequence learning. The data set comprises of one wrong Marathi sentence as the source and another grammatically correct sentence of Marathi as the target. This corpus comprises of several categories of grammatical errors, including gender errors, verb conjugations, tense errors, spelling errors, number agreement errors, and sentence structure errors.

Before training the model, the dataset that is collected goes through many preprocessing steps in order to clean it and make it consistent. Preprocessing includes text cleaning, Unicode normalization, removing symbols and special characters, sentence segmentation, tokenization, and dataset validation. Finally, the cleaned text is converted into a format readable by IndicBART's tokenizer.

IndicBART is an Indic sequence-to-sequence transformer model trained specifically on Indian languages. IndicBART is used as the base model of our proposed approach. It is trained further using the parallel dataset of incorrect and correct Marathi sentences. While training, the encoder takes the incorrect sentence as input and generates contextual embeddings from it that represent both semantic and grammatical aspects. Then, these contextual embeddings are passed to the decoder, which learns to generate the correct version of the sentence.

In the process of grammar correction, there is a step where the input sentence goes through preprocessing and tokenization and then sent to the IndicBART encoder. Here, context-based representations of the input sentence are generated, which later become the basis of the decoder for generating a grammatically correct sentence. With attention mechanisms and contextual learning, the model is able to capture the long dependencies present in the sentences and perform corrections without losing their meanings.

with optimized hyperparameters for better accuracy and fluency. Finally, the system will be tested and evaluated using metrics like BLEU scores.

IV. PROPOSED MODEL

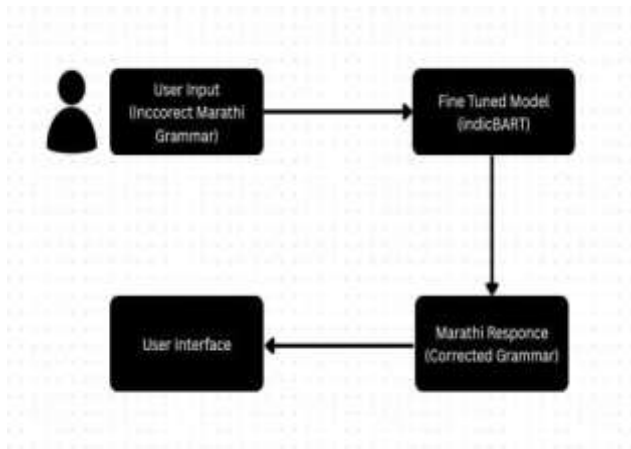


Figure 1. Proposed Model.

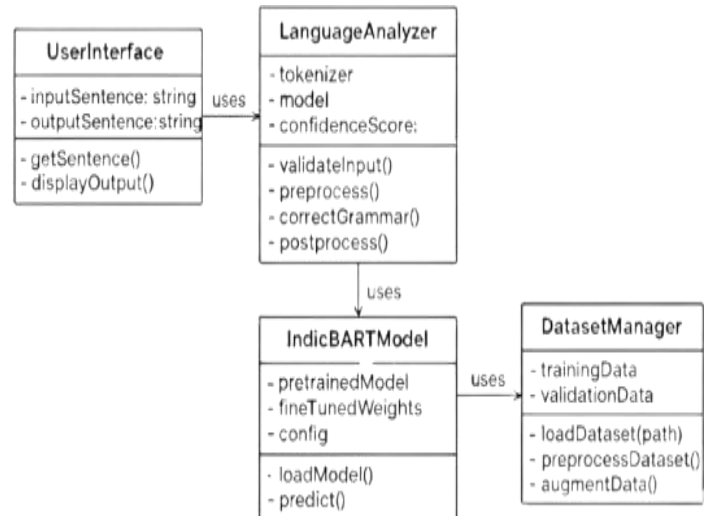


Figure 3. UML Diagram.

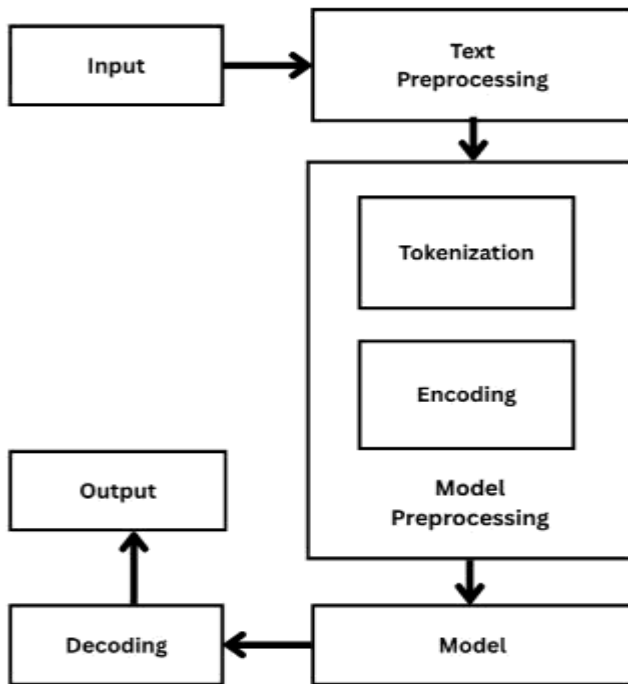


Figure 2. Proposed Architecture.

Table 1: Dataset

Dataset	BLEU ↑	Remarks
Corpus of Books and News (300k pairs)	75	stable training and evaluation outcomes; broad domain coverage; high-quality parallel data of incorrect–correct sentence pairs.

Table 2: Performance

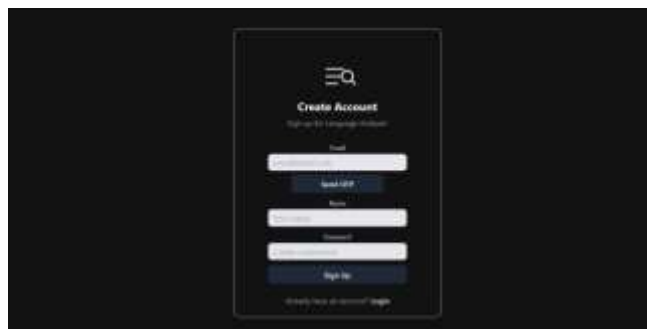
BLEU	75.0
GLEU	0.72
METEOR	87
Chrf++	88

Table 2: Data Split

Training Set	90%
Testing Set	5%
Validation Set	5%

V. Implementation.

1. signup



2. login



Observation and conclusion

MarathiGEC, a transformer-based Marathi Grammar Error Correction framework based on IndicBART, is introduced in this paper. MarathiGEC framework helps in addressing the increasing demand for automation of grammar correction systems for Marathi, an Indian language that is morphologically complex and low-resource. Parallel sentences in the form of grammatically wrong and correct Marathi sentences were collected and employed for the fine-tuning process of the pre-trained IndicBART model. Using grammar correction as a sequence-to-sequence text generation problem helps in learning contextual and grammatical connections between grammatically wrong and correct sentences.

MarathiGEC framework is capable of correcting several kinds of grammatical errors including gender agreement, verb conjugation, tense, spelling, number agreement, and sentence structure errors. Contextual learning ability of the IndicBART model results in generating fluent and grammatically correct sentences without altering the semantics of the sentence.

Besides, this study provides useful linguistic materials and sets a platform for further research in the area of Marathi Grammar Error Correction. Both the prepared data set and the trained model may be used in a variety of tasks, such as education websites, writing assistants, and content creation applications, as well as other related fields. The future work is planned to be directed at the data set expansion, taking into account the dialect differences, improvement of the model in regard to difficult sentence structures, and implementation of the framework as a web/mobile application. The proposed MarathiGEC framework is an important stage in development of Marathi NLP.

References

- [1] Arora, A., Kumar, V., and Ailani, N. (2019). An overview of correcting grammatical errors. 3536–3540 in International Journal of Scientific & Technology Research, 8(11).
- [2] Sonawane, S., Choudhury, M., & Mahata, D. (2020). Hindi grammar correction using inflectional error generation. 242–251 in Proceedings of the 17th International Conference on Natural Language Processing (ICON).
- [3] Bhattacharyya, P., and Sharma, V. (2025). Hi-GEC: A standard for correcting grammatical errors in Hindi. Proceedings of the Association for Computational Linguistics' (ACL) 2025 Annual Conference.
- [4] Joshi, A., Patil, M., and Magdum, P. (2023). MahaNLP is an open-source Marathi natural language processing library that uses models based on MahaBERT. 112–121 in Proceedings of the 2023 Conference on Computational Linguistics.
- [5] Gawali, M., Joshi, A., and Kulkarni, S. (2005). A Marathi spellchecker based on morphology. 67–74 in Journal of Indian Linguistics and Technology, 4(2).
- [6] R. Joshi (2022). L3Cube-MahaCorpus and MahaBERT: Resources, Marathi BERT Language Models, and Marathi Monolingual Corpus. arXiv preprint arXiv:2202.01159.

[7] K. Kumarasinghe, G. Dias, and I. Herath(2021) "SinMorphy: A Morphological Analyzer for the Sinhala Language," in 2021 Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, 2021, pp. 681-686, doi: 10.1109/MERCON52712.2021.9525636.

[8] Ujjwal Sharma and Pushpak Bhattacharyya. 2025. Hi-GEC: Hindi Grammar Error Correction in Low Resource Scenario. In Proceedings of the 31st International Conference on Computational Linguistics, pages 6063–6075.

[9] Raj, J. (2024). Development of a grammatical error correction tool for the low-resource Gujarati language. Bachelor's Thesis, Innopolis University. DOI: 10.13140/RG.2.2.21773.76003

[10] K. B. Khandale and C. N. Mahender(2020), "Natural Language Processing based Rule Based Discourse Analysis of Marathi Text," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, Jul. 2–4, pp. 356–362, doi:10.1109/ICESC48915.2020.9155653.

[11] M. C. Madhavi, S. Sharma, and H. A. Patil(2014), "Development of language resources for speech application in Gujarati and Marathi," in International Conference on Asian Language Processing (IALP), Kuching, Malaysia, Oct. 20–22, 2014, pp. 115–118, doi:10.1109/IALP.2014.6973517.

[12] J. Singh, N. Joshi, and I. Mathur(2013), "Development of Marathi part of speech tagger using statistical approach," in 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, India, Aug. 22–25, pp. 1554–1559, doi:10.1109/ICACCI.2013.6637411.

[13] M. M. Deshpande and S. D. Gore(2018), "A Hybrid Part-of-Speech Tagger for Marathi Sentences," in 2018 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India.

[14] L. R. Nair and S. D. Peter(2011), "Development of a Rule Based Learning System for Splitting Compound Words in Malayalam Language,"